Smallest Causal Decision Tree Creation

Dr. D. Mabuni
Assistant Professor, Dept. of Computer Science and Technology, Dravidian University, Kuppam, Chittoor (District), Andhra Pradesh, India

***Abstract*—Causal inference concepts are inherently present in many real-world tasks. Cause-and-effect relationships are frequently occurring in many applications. In any system if causal inference knowledge is known then it will be easy to take effective decisions under risky conditions. Decision trees such as C4.5 and CART are popular classification algorithms. Causal decision trees can be created by combining with the best features of normal decision tree and popular statistical parameters. Splitting procedure of normal decision tree is different from the splitting procedure of causal decision tree. In causal decision tree splitting is done based on values of treatment effects. Causal decision trees can be created using different ideas, plans and procedures. Correlations between predictor attributes and target attribute play an important role. One way of creating causal decision tree is by using all the attributes in the training dataset. Alternate way for creating causal decision tree is use only highly correlated attributes and remove attributes whose correlation value is less than the specified correlation threshold. In this paper both above said causal decision trees are created. Correlation threshold is also variable according to desired requirements.**

***Index Terms*— machine learning, decision tree, causal decision tree, treatment effects, causal inference, causal relationships, data analytics, decision tree analysis, causal Bayesian networks (CBNs), correlation, smallest causal decision tree. Curse of dimensional, statistical correlation, adult dataset, UCI machine learning.**

I. INTRODUCTION

Machine learning technology has made our lives easily in executing many tasks. Machine learning is a branch of artificial intelligence and computer science and it interacts with many other branches such as statistics, artificial intelligence, data mining, social science, and medical science for finding solutions to many problems. Machine learning is used for effective management of data and algorithms. Machine learning is an important component of the data science. Machine learning algorithms are mainly used for classification and prediction. Machine learning models are divided into three types-supervised, unsupervised, semi-supervised. Popular machine learning algorithms are - decision trees, clustering, logistic regression, K-nearest neighbour classification, association, linear regression, random forests. Some important machine learning applications are-computer vision, market data analysis, speech recognition, customer service, medical data analysis, fraud detection, automated stock trading, retail data management, stream data management.

Analyse data clearly, correctly, completely, in a controlled way and build data analytics models which are used for predicting the future outcomes. Predictive data analytics models are designed to analyse, process and manage historical data, discover useful patterns, ideas, plans, observe trends and use that available information (knowledge) to predict future trends.

I.1 Decision Trees

Decision trees such as ID3, C4.5 and CART (classification and regression trees) are predominant algorithms used for classification and regression. C4.5 algorithm is continuing its top-1 position in the in the entire list of top-10 positions of data mining algorithms. Many improvements have been continuously applying for C4.5 algorithm. Interpretability and comprehensibility are the important features of the decision trees. Decision can process both numerical and categorical data.

I.2 Causal Decision Trees

Causal decision trees are modified versions of normal decision rules by combining with statistical parameters. Each branch of the causal decision tree leads to a relationship of causality instead of classification. Causal tree learning is a machine learning technique used for determining cause and

effect relationships. Causal inference and decision tree analysis are mainly based on statistical parameters.
The main goal of data analytics is to find true causal relationships from the data. Efficient, effective, scalable, automated, robust and fast excellent techniques are needed to find causal relationships from the data. Causal relationships provide clear understanding of the relationships between predictor attributes and target attribute. Causal relationships also provide useful knowledge that is used for correct decision making under high risk conditions. One way to find causal relationships is by using randomized controlled trials but these tests are costly and even sometimes impossible to conduct. Observational studies is the second method that is useful for extracting causal signals from the data but it requires the knowledge of domain experts and generally it is a time-consuming process.
Causal relationships extracting methods must be scalable, automated, ability to handle high dimensional datasets, and must have the capability to find causal relationships without any prior knowledge. In the case of weak-models' ensembles are created by combining weak models. Sufficient research has been done to find parameter values of the heterogeneous effects.
Main goal of machine learning is predicting unknown data from the known data using learned knowledge from the known data. Currently there is a lot of demand for scalable, efficient, effective and automated techniques for causal relationships exploration in the data without involving domain experts. Whenever a target variable is fixed, finding causal relationships from the data becomes a supervised learning technique. A perfect classification algorithm may not find causal relationships. Main goal of causal decision tree is to provide interpretable and actionable knowledge that is used strictly to reduce the effects of the undesirable effects and to increase the effects of the desirable outcomes.

I.3 Graphical Causal Models

In causal inference the main focus is estimating the causal effect of a treatment in the selected outcome. Causal relationships among the variables can be represented by graphs and these graphs are called causal Bayesian networks (CBNs). These graphs are divided into two types. The first type is called direct acyclic graph (DAG) and the second type of graph is called acyclic directed mixed graph.
As for as graphical causal models are concerned only causal Bayesian networks (CBNs) have achieved a greater success in establishing causal relationships from the data. Among the graphical causal data models only causal Bayesian networks have been the most developed one and predominantly used one. The main disadvantage of CBNs is that they are not efficient or not up to the mark as expected. The time complexity of learning a complete CBN is NP-Hard.
Different types of well-established techniques are available for causal relationships exploration in data. Some of these methods are potential outcome models, causal Bayesian network models and structural equation models. Potential outcome model is not suitable for large scale causal relationships exploration. Causal decision tree method uses a statistical test to find the cause-and-effect relationship between a predictor variable and the target variable. If the cause-and-effect score is high then it is an indication that there exists a causal relationship between a predictor variable and the target variable. Decision trees are simple to understand and interpret.

II. LITERATURE SURVEY

Decision trees are mainly used in decision analysis. Decision tree can be converted into a set of interpretable rules. These set of rules are called knowledge base of the decision tree and it is useful for effective decision making. Causal decision trees are created by combining the best features of the classification tree and the best features of the statistical parameter values. J. Li et al. [1] have developed a causal decision tree algorithm for extracting causal relationships from the data. The proposed algorithm establishes a causal relationship between a predictor variable and the target/output variable by using a standard and well-established statistical framework. Proposed method is a hybrid data model consisting of a decision tree model and a standard statistical test. The proposed hybrid data model is a graphical data model that is completely represented by the causal decision tree structure for finding the causal relationships. The proposed technique is able for finding real causal relationships from the large datasets. Experimental results have shown that proposed method is scalable and practical for identifying causal relationships from the large datasets. The proposed causal decision tree technique is an excellent technique that is not only useful for representing causal relationships in the data but also useful for identifying interpretable true causal relationships in the data.

[2] This paper summarizes some of the important and advanced causal relationships. Causality details based on the structural causal model are described in detail. This paper also explains mathematical tools used for inferring causal queries. Normally causal relationships cannot be represented in probabilistic notation. Causal effects can be identified when the model is Markovian (acyclic graph). J. Pearl [3] said that more importantly humans implicitly employ Bayesian networks to infer cause and effect relationships. Pearl wanted to extend Bayesian causality to statistics and economics related principles. E. Eells [4] said that in probabilistic causation, probabilistic tools are used for inferring causal relationships. Probability causation says that causes change the probability of their effects. An effect may occur without a cause and an effect may not occur with a cause. Causal data modelling methodology is generally used principles from other disciplines such as statistics, artificial intelligence, philosophy, econometrics, epidemiology, and many other inter disciplinary domains.

D. B. Rubin [5] said that causal effects are defined as the difference between potential outcomes when treatments are applied and the potential outcomes when treatments are not applied. Potential outcomes are useful methods for better understanding of the causal relationships. Imbens and Rubin [6] said that in social and biomedical sciences majority of the questions are causal. The authors have given clear explanations about how random experiments are allowed to assess the causal effects. S. Greenland and B. Brumback [7] in their paper provides a brief overview of important types of causal data models such as graphical models, potential outcome models, structural equation models and sufficient-component causal models. Main important concept of this paper is to provide logical connections among the different types of causal data models.

J. Li et al. [8] said that causality inferences are primarily used for prediction, explanation, and decision making in applications such as medicine, engineering, biology, economics, epidemiology, research, physics, business, and social sciences. Four types of causal inference methods are innovative, efficient, effective, and robust in identifying true causal relationships with respect to a given target variable. P. Spirtes et al. [9] said that causation means the relationship between particular events. Each cause is treated as a particular event and each effect is also treated as a particular event. A causal graph is a directed acyclic graph that represents the relationships among the variables in the form of causal structure. Z. Jin et al. [10] identifying causal relationships from a large dataset of observational data is a tedious task. Because Bayesain learning is NP-complete many researchers are trying to develop local and constraint based causal relationships. Authors proposed an algorithm for mining causal rules in large databases of binary variables. Causality finding procedures are extended to identify causal relationships with multiple cause variables.

J. Li et al. [11] said that a very urgent requirement is needed to design and develop an efficient method for extracting causal relationships directly from the data. Causal Bayesian network is one such method for obtaining direct causal relationships from the data but its time complexity is NP-Complete. So, it is not suitable and useful as expected. Another alternative is association rule mining for extracting causal relationships directly from the data. Authors [12] said that causal Bayesian networks (CBNs) can be used in order to extract required explanations about the observed subset of causal variables. Special type causal trees called explanation causal trees are designed and developed for explicit causal inference.

Authors [13] proposed techniques for estimation of heterogeneous causal relationships by constructing regression tree methods. Authors [14] have used matching techniques for causality determination and also, they have clearly explained that matching techniques can be used effectively to increase the strength of prosecution of causal questions in sociology. R. P. Rosenbaum and B. D. Rubin [15] said that propensity score is the conditional probability assigned to a particular treatment from a given set of covariates. Propensity scores are sufficient to remove bias due to all observed covariates. D. B. Rubin [16] the aim of analysing large databases is to extract causal relationships about the effects of actions or treatments. Authors said that propensity score-based methods are more reliable and standard tools for causal relationships extraction.

E. A. Stuart [17] said that estimation of causal effects using observational data requires replication of randomized experiments very closely for the distributions. Matching is a method that equates the distribution of covariates in the treated and control groups. Well matched methods are convenient for extracting causal inference details and these methods are already popular in sociology, medicine, research, business, epidemiology, science, physics, economics, political science, and artificial intelligence related fields. This paper provides basics for understanding matching methods. P. C. Austin [18] said that propensity score is the probability of treatment assignment for finding the causal inference. Propensity score is used in randomized experiments and in observational studies. The propensity score is generally estimated using logistic regression. Propensity score is computed using various statistical packages.

Authors [19] said that propensity score is a conditional probability assigned to a particular treatment from the given set of covariates. Propensity score method removes bias and maintains balance properties in the case of incomplete data. Logistic regression [20] is predominantly used in machine learning, statistics, and in many data analysis tasks. It is a probabilistic model useful for better explanation of data. For some applications logistic regression outperforms the existing advanced methods. B. K. Lee et al. [21] said that popular decision tree classifiers such as C4.5 and CART are alternative techniques to logistic regression for finding propensity scores. Authors have examined various CART-based propensity score data models and experiments are conducted through simulation. Various models of CART are - CART, pruned CART, ensemble CART, and logistic regression CART. Through experimental study they observed that ensemble method particularly boosted CART is superior in performance when compared with other methods.

N. Younas et al. [22] said that ensemble methods can be used for finding causal relationships for effective decision making under high-risk conditions. This paper explains the importance of causal decision tree forest and a new ensemble of optimal causal decision tree forest is created for accurately estimating heterogeneous effects. Initially a large number of causal decision trees are created and then a subset of important and accurate causal decision trees is selected and then ensemble forest is created. Performance of the proposed method is evaluated by using many metrics such as mean square error, root mean square error, mean absolute deviation, Pearson correlation and ordinary least squares. Finding causal relationships in the data is an important task for accurate decision making in risk-based situations.

S. Betnard et al. [23] created an ensemble of causal decision trees from the given very large forest of causal decision trees. Authors have used idquoclassicalrdquo random forest technique for ensemble creation but the used technique is suffering with very low interpretability and in the ensemble number of trees must be specified at the beginning itself. From the given set of initial causal decision trees actually ‘n’ number of ensembles are possible but the proposed method is capable in creating better subsets of causal decision trees.

T. M. Oshiro et al. [24] said that random forest is an efficient technique for ensemble creation and it has been used in many real-world applications. Here the main problem is that how to decide ‘n’ where n is the number of trees in the random forest. There is no direct method for finding exact value of n. One of the heuristic approaches is that at the beginning include some optimal number of trees and then add one by one and immediately compute accuracy and see that, is there any better accuracy improvement than the previous accuracy then include the selected tree; otherwise discard the selected tree. The process continuous until a certain point where there is no improvement in the accuracy even a new tree is added. Stop the process at that point.

P. Latinne et al. [25] said that the main goal of the present paper is to determine minimum number of decision trees that must be included in the ensemble. A new procedure is proposed for selecting and then deciding the value of n at the beginning itself. The proposed technique is based on McNemar non-parameter test of significance. S. Athey et al. [26] proposed generalized non-parametric random forest based on statistical estimations. The proposed method uses weighted set of nearby training tuples. So, authors have proposed efficient, effective, flexible and scalable generalized random forests. T. G. Dietterich [27] has created ensembles of decision trees using bagging and boosting techniques but bagging and boosting are suffering with the instability of the base learner. A new randomization technique is proposed for generating ensemble.

P. Geurts et al. [28] proposed a new randomized method for ensemble creation. The structures of the trees are independent of the output values. The strength of the randomization is increased with parameter estimation and tuning. G. Biau [29] said that random forest is an ensemble created with a set of decision trees. Random forest creation details are discussed in detail. Z. Hu et al. [30] proposed a description length guided Granger causality analysis for sequential medical imaging. Using minimum description length parameter, the performance of the proposed technique is evaluated and found that the performance of the proposed technique is superior than many other methods.

III. PROBLEM DEFINIION

It is quite natural that many real-world datasets contain hundreds or even thousands of attributes. Curse of dimensional is a major problem in many machine learning applications. The main problem in many real time applications of data science is how to remove unnecessary or unimportant attributes before using of the selected training dataset. To alleviate curse of dimensional problems there is a need to reduce the number of attributes in the training datasets before actually using them in the real-time tasks. Different types of methods are in use for removing unwanted or unnecessary attributes from the very large datasets. Always smallest models are preferable. To make the machine learning model simple and effective, explicitly remove unnecessary attributes in the training dataset. Here, the main problem is how to remove unnecessary attributes from the training datasets.

IV. PROPOSED METHOD FOR FINDING SMALLEST CAUSAL DECISION TREE

Machine learning models are very useful and convenient models for effective decision making. According to Occam razor simple model is the best model in many real time situations because they are very easy to understand and interpret. Main goal of machine learning is to create a simple, optimal but effective model for correct decision making. Causal decision trees are convenient models to represent cause-and-effective relationships in the data and the knowledge of these relationships is very useful to take correct decisions under high risky situations. To create such knowledgeable models many methods are available. Correlations based knowledge between predictor attributes and output attribute is useful for creating smallest causal decision trees. Correlations represent important relationships between predictor attributes and target attributes. In general, high correlation between predictor and target attributes indicates high causality. Causality implies correlation but the converse may not be true in all the cases. Smallest causal decision tree can be constructed by discarding or removing attributes whose correlation value is less than the specified correlation threshold value. Attributes whose correlation value is less than the specified correlation threshold value are called weak attributes and attributes whose correlation value is greater than or equal to the specified correlation threshold value are called strong attributes. The causal decision tree that is constructed by using only strong attributes is called smallest causal decision tree or strong causal decision tree (SCDT). In this paper both weak and strong causal decision trees are created actually by taking real-time UCI machine learning dataset called adult dataset.

IV.1 Causal Decision Tree Construction using Full Training Dataset

Table-1 Attributes of the UCI Machine Learning Adult Training Dataset

S.No	Attributes
1	Age < 30
2	Age > 60
3	Private
4	Self-emp
5	Gov
6	Education-num > 12
7	Education-num < 9
8	Prof
9	White
10	Male
11	Hours > 50
12	Hours < 30
13	US
14	> 50k

In general, training datasets contain thousands of attributes. Whenever a causal decision tree is constructed for the complete training dataset containing high correlated and low correlated predictor attributes with target attribute, the time complexity is very high because the number of input attributes is very high. In the present paper, first a causal decision tree is constructed for the UCI machine learning adult training dataset by taking all the predictor attributes including highly correlated and weakly correlated attributes with the target attribute. Secondly, causal decision tree is created after removal of unnecessary attributes and for some selected attributes from the training dataset.

UCI machine learning adult dataset is taken as a training dataset. Adult dataset contains total 13 predictor attributes (1 to 13) and one target attribute (14) as shown in the Table-1. Causal decision treeis constructed by taking the complete set of all the given attributes in the adult training dataset. Dimensionality reduction technique plays a major role in many real applications. In machine learning, different types of dimensionality reduction technique are already available and still new dimensionality reduction technique are always welcoming to develop.

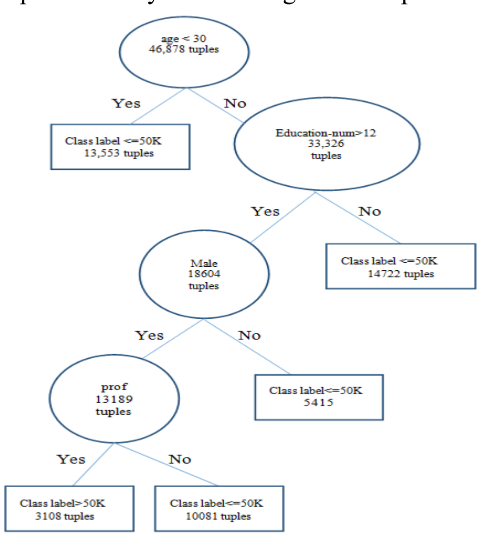

Figure-1 Causal decision Tree for All Attributes of the Adult Dataset

IV.2 Causal Decision Tree Construction using only with the Selected Highly Correlated Predictor Attributes

Before creating causal decision tree, correlation values between predictor attributes and the target attribute are computed and all the attributes whose correlation value is less than the specified correlation are removed from the list. Causal decision tree is constructed by taking only highly correlated or strong predictor attributes. Time complexity is reduced approximately up to 50%. Attributes with low correlated values are removed from consideration and used only strongly correlated attributes for causal decision tree creation. Causal decision tree constructed by using only strong predictor attributes is called smallest or strong causal decision tree. Strong causal decision trees are always preferred ones because their usage is highly desirable. Smallest causal decision trees are always preferable to design, develop and then use as much as possible in day-to-day applications for achieving real benefits.

Same causal decision tree is created for the smart training dataset. Smart training dataset is a training dataset with low correlated attributes are removed from the training dataset. That is causal decision tree model is generated only for the selected and highly correlated attributes. But the wonder is that both causal decision trees are exactly the same. This procedure is called one type of attribute pruning from the initial training dataset. That is, it is an example for simple dimensionality reduction technique. Correlation is an important statistical technique or property used for dimensionality reduction. Note that there may exist other statistical methods and formulas for reducing the dimensions form the original training dataset.

Table-2 Included training dataset attribute numbers are – 1, 3, 6, 8, 10, 11, 12. Selected correlation threshold value is 0.1

S.No	Attributes	Correlations against target attribute	Selected attributes
1	Age < 30	0.2846	Age < 30
2	Age > 60	0.0018	Removed
3	Private	0.1178	0.1178
4	Self-emp	0.0974	Removed
5	Gov	0.0563	Removed
6	Education-num > 12	0.2381	Education-num > 12
7	Education-num < 9	0.0775	Removed

8	Prof	0.1834	Prof
9	White	0.0811	removed
10	Male	0.2141	Male
11	Hours > 50	0.1386	Hours > 50
12	Hours < 30	0.1481	Hours < 30
13	US	0.0382	removed
14	> 50k	Class attribute	Class attribute

In this case selected correlation threshold value is 0.1. Here, one can specify correlation threshold as 0.2 or 0.3 or 0.4. Correlation threshold specification value depends on the particular application and its selection is not fixed and it differs from one application to another application. Based on the experience one must correctly select and then fix correct correlation threshold value.

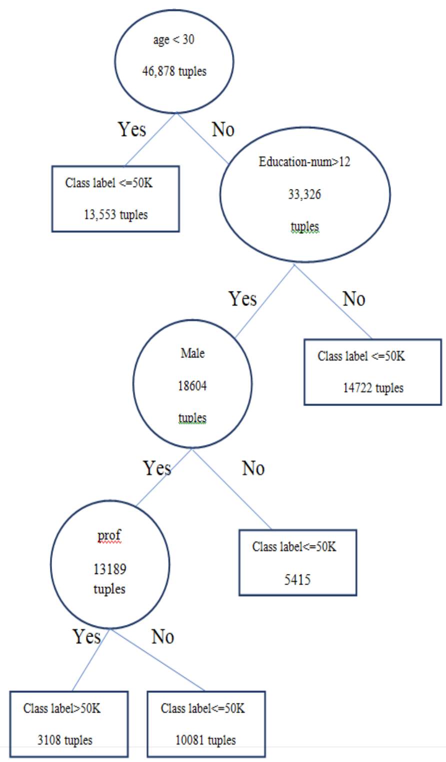

Figure-2 Causal Decision Tree after removal of unnecessary attributes of the Adult Dataset

IV.3 Algorithm for Strong Causal Decision Tree
Algorithm Strong_Causal_Decision_Tree (Training Dataset T)
Input: Training Dataset, T
Output: Strong Causal Decision Tree (SCDT)
1.for each predictor attribute, Ai in the training dataset do
2. find the correlation, corr(Ai, T), between Ai and the target attribute,
3. if corr (Ai, T) >= corr_threshold Then
4. S = S U Ai
5. endif
6.endfor
7. for each attribute Xi in S do
8. compute causality value causality (Xi) using statistical formula
9. end for
10. find highest causality score attribute
11. split data in the current node based on the highest causality score attribute
12. repeat steps 10 and 11 for each subbranch of the current node with highest causality score
13. if the causality score of all the attributes is less than the threshold causality then
14. create a leaf node with appropriate class label
15. endif

V. CONCLUSIONS

The knowledge of causal relationships present in the data is useful for effective decision making under risk-based conditions. In this paper, the causality knowledge is represented in the form of causal decision trees models. Thus, causal decision tree models are important in data science in multitude ways. The models must be as simple as possible but should contain more than 90% of correct and accurate details. In this paper, a new algorithm is proposed for constructing smallest causal decision tree that includes as simple as possible smallest subset of highly correlated attributes of the given training dataset. Many of the unimportant attributes are removed from the dataset based on correlated values and then causal decision tree is created by selecting only strong predictor attributes. This causal decision tree is called smallest or strong causal decision tree.

In this paper, two types of causal decision trees are created without removing attributes and then with removing unnecessary attributes. In the future, different and efficient machine learning and statistical methods and formulas will be studied and then investigated thoroughly for finding better dimensionality reduction techniques.

REFERENCES

[1] J. Li, S. Ma, T. Le, L. Liu, and J. Liu, "Causal decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 257–269, Feb. 2017.

[2] P. Spirtes, "Introduction to causal inference," *Journal of Machine Learning Research*, vol. 11, pp. 1643–1662, 2010.

[3] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.

[4] E. Eells, *Probabilistic Causality*. New York, NY, USA: Cambridge University Press, 1991.

[5] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[6] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge, U.K.: Cambridge University Press, 2015.

[7] S. Greenland and B. Brumback, "An overview of relations among causal modelling methods," *International Journal of Epidemiology*, vol. 31, no. 5, pp. 1030–1037, 2002.

[8] J. Li, L. Liu, and T. D. Le, *Practical Approaches to Causal Relationship Exploration*. Berlin, Germany: Springer, 2015.

[9] P. Spirtes, C. C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[10] Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang, "Discovery of causal rules using partial association," in *Proc. IEEE 12th Int. Conf. Data Mining (ICDM)*, Brussels, Belgium, Dec. 2012, pp. 309–318.

[11] J. Li, T. Le, L. Liu, J. Liu, Z. Jin, and B. Sun, "Mining causal association rules," in *Proc. IEEE 13th Int. Conf. Data Mining Workshops (ICDMW)*, Dallas, TX, USA, Dec. 2013, pp. 114–123.

[12] U. H. Nielsen, J. Philippe Pellet, and A. Elisseeff, "Explanation trees for causal Bayesian networks," in *Proc. 24th Conf. Uncertainty in Artificial Intelligence (UAI)*, Helsinki, Finland, 2008, pp. 427–434.

[13] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.

[14] S. L. Morgan and D. J. Harding, "Matching estimators of causal effects: Prospects and pitfalls in theory and practice," *Sociological Methods & Research*, vol. 35, no. 1, pp. 3–60, 2006.

[15] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[16] D. B. Rubin, "Estimating causal effects from large data sets using propensity scores," *Annals of Internal Medicine*, vol. 127, no. 8, pp. 757–763, 1997.

[17] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statistical Science*, vol. 25, no. 1, pp. 1–21, 2010.

[18] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011.

[19] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[20] P. Komarek, "Logistic regression for data mining and high-dimensional classification," Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 2004.