International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Inferring the Causal Relationships in Student Placements' Performance using Causal Machine Learning

[1]D. Naga Jyothi*, Dr. Uma N. Dulhare[2]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: This research proposes using causality models to analyse and infer student placement data. It demonstrates the distinctions between applications of Causal Machine Learning and Machine Learning for resolving different education-related processes. *Association does not equal Causation*. In traditional machine learning, the focus is often on predicting outcomes or patterns based on input data. However, causal machine learning goes beyond prediction by aiming to uncover cause-and-effect relationships between variables. The review of causal inference in the presence of massive data sets is a rich and expanding field of contemporary research. The goal of causal inference is to understand how changes in one variable affect another, and to identify the underlying mechanisms that lead to certain outcomes. The causal Inferencing which is the key concept for causal machine learning can be implemented using the DAG (Directed Acyclic graph). Through this paper we aim to provide some useful insights using 3 causal discovery tools (PC, GES, LiNGAM) to produce the DAGs. We proposed a novel 3D framework (Data correlation, Discovery tool using Causal ML, Domain knowledge) which combines the merits of both manual and causal discovery tools. The causal graph obtained is checked for falsification i.e. the correctness of the graph. The obtained graph needs to be informative and significance level (p-value < 0.05) so that the DAG would be accepted. Thus, a final Causal Model is formed that represents relationships between the variables to understand and predict the effects of interventions or changes in the system.

Keywords: *Causal relationships, Causal discovery techniques, Directed Acyclic Graph (DAG), 3D Framework, Treatments, Confounders, Falsification, Causal Modelling.*

1. Introduction

Despite all the hype surrounding AI, the majority of ML initiatives prioritise outcome prediction over causality analysis. Indeed, after several AI projects, It is realized that ML is great at finding correlations in data, but not causation. This problem severely restricts our ability to use Machine Learning for Decision Making.

Machine learning is a powerful tool to find patterns and to examine associations and correlations, particularly in large data sets [1]. Although the use of machine learning has led to the emergence of numerous productive sectors for research in social science, public health, economics, education and medicine, these disciplines still need approaches that can address causal issues rather than just correlational analysis. Various tools can be used manage the student projects, placements[14]. ML algorithms in their current state can be biased, suffer from a relative lack of explainability, and are limited in their ability to generalize the patterns they find in a training dataset for multiple applications. Exploration is done on how to combine the different parameters that affect the accuracy of the machine-learning algorithms with respect to different products[15]. It is reasonable to assume that considering causality in a world model will be a critical component of intelligent systems in future. In traditional machine learning, the focus is often on predicting outcomes or patterns based on input data. However, causal machine learning goes beyond prediction by aiming to uncover cause-and-effect relationships between variables.

The commonly held belief that "correlation does not imply causation" refers to the fact that a causal relationship between the variables cannot be inferred just from correlation. It is important to remember that correlations do not always imply the presence of potential causal relationships. On the other hand, strong correlations also overlap with identity relations (tautologies), where no causal process is present, and the factors underlying the correlation, if any, may be indirect and unknown. As a result, establishing a causal relationship (in either direction) requires more than just a correlation between two variables.

The review of causal inference in the presence of massive data sets is a rich and expanding field of contemporary research. Understanding how changes in one variable impact another and figuring out the underlying mechanisms that produce particular results are the two main objectives of causal inference. This is important in various fields such as healthcare, economics, social sciences, and more, where understanding causality can be crucial for making informed decisions and interventions.

[1]*Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*
ORCID ID : 0000-0001-9498-2882
[2]*Muffakam Jah College of Engg. And Tech, Hyderabad ,Telangana, India*
ORCID ID : 0000-0002-4736-4472
** Corresponding Author Email: dnagajyothi_cseaiml@cbit.ac.in*

The difference between the prediction and causal inference can be described as follows.

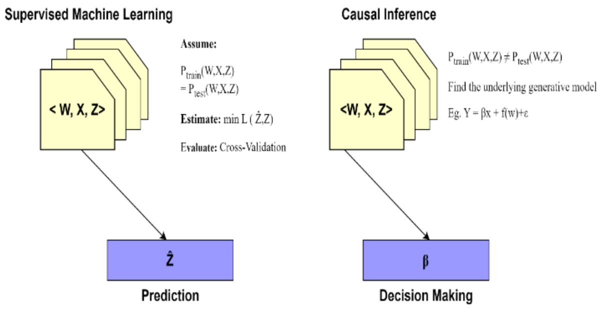

Supervised Machine Learning
< W, X, Z>
Assume:
$P_{train}(W,X,Z)$
$= P_{test}(W,X,Z)$
Estimate: $\min L(\hat{Z},Z)$
Evaluate: Cross-Validation
$\hat{Z}$
Prediction
Causal Inference
<W, X, Z>
$P_{train}(W,X,Z) \neq P_{test}(W,X,Z)$
Find the underlying generative model
Eg. $Y = \beta x + f(w) + \varepsilon$
β
Decision Making

Fig .1. Difference between the prediction and causal inference

Making well-informed decisions requires the use of causal inference, which goes beyond the simple connections found in predictive models to reveal the actual processes that generate data. Even in the lack of interventional data, it allows us to estimate the impacts of treatments and counterfactual outcomes. To truly comprehend linkages in the actual world and generalise knowledge, it is imperative to go beyond correlation-based research. For example, it is crucial to accurately forecast and comprehend the causal impacts of computing systems that actively intervene in societally significant domains like healthcare, education, and governance.

1.1 Machine Learning Applications

- ***Personalized Learning:***

Adaptive Learning Systems use different Platforms using ML to adapt educational content to individual student needs based on their learning pace and style.

Intelligent Tutoring Systems (ITS) like Carnegie Learning provide personalized tutoring by analyzing student performance data to adapt instructional strategies in real-time.

- ***Predictive Analytics:***

Student Performance Prediction is carried out using ML models by predicting student outcomes, such as grades and graduation rates, enabling early intervention for at-risk students[17].

Dropout Prevention can be done by analyzing historical data, ML models identify students at risk of dropping out and suggest interventions[18].

- ***Natural Language Processing (NLP):***

Automated Essay Scoring can be done using tools like Grammarly to evaluate and score essays, providing immediate feedback on writing quality.

Chatbots and Virtual Assistants are used to answer student queries, offer homework help, and provide administrative support.

- ***Content Recommendation:***

Recommender systems suggest the next best course or module based on a student's performance and interests for Learning Path Optimization. ML helps in recommending textbooks, research papers, and supplementary materials tailored to individual learning needs.

- ***Classroom Analytics:***

Attendance Monitoring can be done using computer vision and ML by tracking student attendance and engagement during classes. Behavioral Analysis can be done using ML models by analyzing classroom behavior to understand engagement levels and identify disruptive patterns.

1.2 Causal Machine Learning Applications

- ***Impact Evaluation of Educational Interventions:***

Causal ML models, such as causal forests or double machine learning, evaluate the effectiveness of educational programs and policies by isolating causal effects from confounding variables. These models help in understanding the impact of policy changes, such as changes in curriculum or teaching methods, on student outcomes.

- ***Personalized Education Plans:***

Causal ML can identify which educational interventions work best for which students, allowing for more precise and effective personalized education plans. By understanding how different students respond to different interventions, educators can design strategies that maximize overall educational outcomes.

- ***Resource Allocation:***

Causal ML helps in determining the most effective allocation of educational resources, such as funding, teachers, and technology, to maximize student success. These models evaluate how different student demographics are affected by resource distribution, ensuring more equitable educational opportunities. By leveraging ML and Causal ML, the education sector can significantly enhance its ability to deliver personalized, effective, and equitable learning experiences.

1.3 Causal ML Models

Our data contains patterns, which machine learning lets us identify and use to inform decisions. Our decision-making process is being revolutionised by machine learning. Its foundational premise is that the data it sees during training is indicative of the data it sees in production and the data we used for testing. Errors occur when this assumption is violated. Thus, once the model has been implemented, it is crucial to monitor its performance as well as the distribution of the variables. We must retrain the model and teach it the correlation patterns found in the new data set whenever there is a noticeable drift in any of the important variables. Thus, this basic premise and this assumption is what machine learning is based on.

Causal machine learning refers to the application of machine learning techniques to infer and understand causal relationships between variables. Causal machine learning faces unique challenges because establishing causation requires more than just observing correlations. It often involves dealing with confounding variables, selection bias, and other factors that can influence both the treatment and outcome. Researchers and practitioners use various methods, such as randomized controlled trials, observational studies, and causal modeling techniques, to provoke causal relationships from observational data. So, for good decision making we just not only need predicting the target variable but also to find the variables that caused the outcome. We also need to estimate how the outcome would change if we changed these variables. This is called **causal inference.**

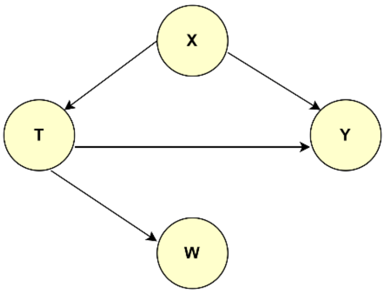


Fig .2. Causal Graph where T is the Treatment, Y is the outcome, X is the Confounder

1.4 Causal Effect

Causality can be defined as follows. We state that a treatment T (which may also refer to a decision or an action) results in an outcome Y if and only if altering T modifies Y while maintaining all other parameters constant.

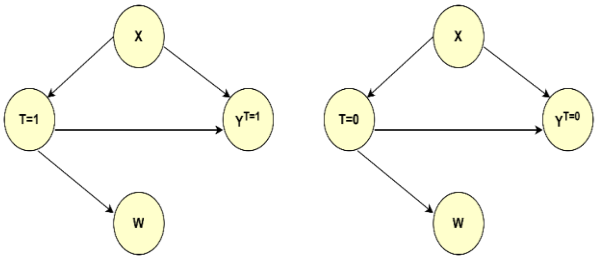


Fig .3. Causal graph representing the real world(left) & counterfactual world(right)

There are two main obstacles to causal inference: 1. We never see the counterfactual reality; We can't compute the causal effect directly; We have to estimate the counterfactuals; and There are validation issues. 2. A single data distribution can be fitted with many causal mechanisms.

Causal inference requires assumptions and domain expertise; data alone is insufficient.

2. Literature Review

One of the study tells about experimental designs for educational systems are introduced with DAGs and graphical models. Sequential interventions and confounding model control are especially emphasized. Comparing the effectiveness of g-formula and (Inverse Probability of Treatment Weighting) IPTW for obtaining unbiased causal estimates in educational contexts, it shows how important it is to account for confounders in order to provide causal estimates that are reliable. A popular experimental design in education is the Randomised Controlled Trial (RCT), which assigns participants at random to treatment and control groups with the goal of removing confounding variables. Sequentially Multiple Assignment Randomized Trial (SMART) is another design that uses randomization at every stage to enable several interventions and assessments over time. This suggests more clever experimental designs by gathering more comprehensive student data, which includes possible confounders in diagnostic exams, and modelling the educational system with time-varying interventions and feedback from confounders. There is a need for more sophisticated methods because traditional assessments of educational systems do not account for confounding variables, provide feedback to students, or account for real-world study variances. This study addresses the sequential character of learning and the significance of several interventions in educational systems. It also emphasises how large amounts of educational data are available for the development of intelligent modelling and inference algorithms. Time-varying confounders may not be captured by several frameworks used in education, such as Intelligent Tutoring Systems (ITS), Hidden Markov Models (HMM), Dynamic Bayesian Networks (DBN), and others. The implementation and assessment of these suggested models and methods in actual educational environments may be the main focus of future study in order to gauge their efficacy in raising student performance and enhancing educational outcomes. The paper emphasizes the importance of modeling educational systems using graphical causal models and directed acyclic graphs (DAGs) to quantify interventions and confounders accurately, enabling the derivation of unbiased causal estimates of joint interventions on outcomes. It discusses the limitations of traditional cross-sectional studies in educational research and highlights the need for experimental and quasi-experimental designs that consider confounding variables, feedback mechanisms, and real-world deviations from ideal conditions [1]. One research work describes about the Mathematical foundations of machine learning with various examples. An empirical study of supervised learning algorithms like Naïve Bayes, KNN and semi-supervised learning algorithms viz. S3VM, Graph-Based, Multiview are also explained in the view of Machine learning algorithms[16].

A research work explains to uncover and analyse the correlations between the important student traits associated with low performance, the research applies machine learning and causal discovery algorithms. It uses a variety of machine learning and causal discovery techniques to predict and explain correlations between children who score poorly in reading, emphasising patterns and data insights already in place. The introduction places a strong emphasis on the value of education in society and the necessity of comprehending the connection between academic achievement and student characteristics. Gradient Boosting, K-nearest neighbours, SVM, Random Forest, and Decision Tree are the machine learning algorithms used in this work to analyse student data and performance. The impact of variables on low-performing students was inferred and causal linkages were found using four different causal discovery algorithms: PC, GES, LinGAM, and GOLEM. The study report made use of information from the PISA 2018 database[2]. Causal discovery algorithms were evaluated for their effectiveness in characterising the relationships among at-risk students using evaluation criteria such as FDR, Recall, Precision, F1 score, and SHD score. To study the correlations between student performance and attributes in greater detail, future research may require investigating novel causality models. Researchers can gain more understanding of the variables impacting students' academic performance by utilising various causal discovery algorithms or combinations of algorithms.

3. Methodology

In this section the proposed work is implemented using the various causal discovery algorithms using the Placement Dataset to draw a causal graph which talks about the treatments, outcomes, confounders and instrument variables which are discussed in the above sections. A novel framework called 3D Framework is proposed and utilised to consolidate the aspects of the different causal discovery algorithms with the help of Domain knowledge which is necessary for Causal Modelling.

3.1 Python Libraries

DoWhy is a Python library designed to encourage causal analysis and thought, similar to what machine learning libraries have done for prediction. DoWhy offers an extensive range of algorithms for root cause analysis, interventions, effect estimation, prediction, quantification of causal influences, learning causal structures, diagnostics of causal structures, and counterfactuals. DoWhy's response API, which can verify causal assumptions for any estimation method, is a crucial component that improves inference's robustness and makes it more understandable for non-experts.

Modelling causal relations as a causal graph is the first step in carrying out a causal job in DoWhy. The "cause-effect-relationships," found in a system domain are modelled by a causal graph. This helps to clarify each causal presumption. We need the causal graph to be a directed acyclic graph (DAG), with an edge X→Y signifying that X is the cause of Y. A causal graph represents the conditional independence relationships between variables statistically.

Causal Inference Libraries:

DoWhy: A library that provides a unified interface for causal inference methods. It is built on top of popular libraries like Pandas, NumPy, and scikit-learn.

CausalML: A library that offers a suite of methods for causal inference and machine learning, including methods for estimating treatment effects and dealing with confounding.

EconML (Econometric Machine Learning): Developed by Microsoft Research, this library focuses on providing tools for estimating treatment effects using machine learning methods.

3.2 Data Preprocessing

The dataset is loaded, and different preprocessing techniques are performed to remove the unwanted data and clean the data by encoding, removing the missing values. Causal modelling, also known as causal discovery, is the initial stage. This involves encoding our assumptions in a causal graph. Our subject expertise will be added to our observational dataset. This step focuses on discovering the network of influences that exist between the features. The directed acyclic graphs (DAGs) are used to quantify the experimental design for the educational system that is proposed. Moreover, we propose to represent the educational system as a mixture of confounders, time-varying treatments, and feedback between the former two. Modelling and assessment take up most of the time in the early stages, while feature engineering can take much longer as the system ages [1].

3.3 Causal Discovery Algorithms

A widely used causality framework is the graphical model that use Directed Acyclic Graphs (DAGs).

- Directed edges are used in the graphical model to indicate the cause and effect.
- Simple to show the outcome, treatment/intervention, and confounding.
- It depicts the procedure used to generate data.

- It is simple to show d-separation and data independence.

3.3.1 Peter – Clark (PC) algorithm

PC stands for the Peter and Clark algorithm, used for learning the structure of Bayesian networks from data. It is a constraint based causal discovery method.

Algorithm steps:

1. Graph Initialization: Form a complete undirected graph on the vertex set V.
2. Edge Deletion: Iteratively select pairs of variables X and Y that are adjacent in the graph, test for d-separation, and delete edges based on conditional independence tests.
3. Graph Refinement: Update the graph structure by recording subsets of adjacent variables that lead to d-separation, ensuring accurate representation of conditional independence relationships.
4. Efficiency and Reliability: The PC algorithm is computationally efficient and asymptotically reliable but may take unnecessary risks on sample data, impacting edge elimination decisions.

3.3.2 Greedy Equivalence Search (GES) algorithm

The Greedy Equivalence Search (GES) algorithm with the Bayesian Information Criterion (BIC) score is a method used for score-based causal discovery in graphical models. Here's a step-by-step explanation of how this algorithm works:

Algorithm steps:

1.Graph Initialization: Begin with an empty graph where no edges are present.

2. Initialize Possible Edges: Specify the possible directions for edges between variables (e.g., A → B, A ← B, A ↔ B).

3. Iterative Edge Addition and Deletion: For each pair of variables X and Y, consider adding an edge X→Y, X←Y, or X↔Y (bidirectional). Also, consider deleting existing edges to explore different graph structures.

4. Score Computation: For each proposed graph (with added or deleted edges), compute the BIC score. The BIC score balances model fit and complexity. The BIC score for a graph G given data D is calculated as:

BIC(G)=log(P(D∣G))−2dlog(n) ,where P(D∣G) is the likelihood of the data given the graph G where d is the number of parameters in the model and n is the sample size.

5. Model Selection: Choose the graph structure that maximizes the BIC score among all considered structures.

6. Repeat and Refine: Continue the process of adding or removing edges, re-computing the BIC scores, and selecting the best-scoring model until convergence or a predefined stopping criterion (e.g., maximum iterations) is met.

7. Final Model: The resulting graph with the highest BIC score represents the inferred causal relationships among variables based on the given data.

3.3.3 Linear, Non- Gaussian Acyclic Model (LiNGAM)

Constrained functional causal discovery using LiNGAM (Linear Non-Gaussian Acyclic Model) involves a methodology for inferring causal relationships among variables from observational data while incorporating specific constraints or assumptions about the underlying causal structure.

LiNGAM is a model used for causal discovery that assumes a linear causal relationship between variables but allows for non-Gaussian (non-normally distributed) noise. The fundamental assumption of LiNGAM is that the causal relationships among variables can be represented by a directed acyclic graph (DAG), where nodes represent variables and edges represent causal relationships.

Algorithm steps:

1. Model Specification: Define the LiNGAM model and its assumptions, including linearity of causal relationships and non-Gaussian noise.
2. Data Preprocessing: Prepare the data, ensuring that it meets the assumptions of the LiNGAM model (e.g., linearity, non-Gaussian noise).
3. Causal Discovery: Use the LiNGAM algorithm to infer the causal structure from the data. This involves estimating the parameters of the model that best explain the observed data.
4. Incorporating Constraints: Apply constraints to the causal discovery process. These constraints might include:

- Variable Subset Constraints: Limit the

  search for causal relationships to a specific

  subset of variables.
- Structural Constraints: Enforce specific structures in the causal graph (e.g., acyclicity, sparsity).
- Interventional Constraints: Incorporate known interventions or causal relationships based on prior knowledge.

5. Evaluation and Validation: Assess the quality and validity of the inferred causal model. This may involve cross-validation, testing against independent datasets, or comparison with ground truth if available.

In summary, constrained functional causal discovery using LiNGAM-based algorithms provides a principled approach to uncovering causal relationships while incorporating domain knowledge and respecting specific constraints on the causal structure. This methodology is particularly useful in fields such as economics, genetics, and neuroscience, where understanding causal relationships is critical for making informed decisions and drawing meaningful conclusions from data.

3.4 Manually determining the DAG based on the Domain Knowledge (using MCMD Generalization)

The predominant outcomes in the educational field are imperatively & evidently comparative in real time execution.

- For instance, selection of a student in placements is a comparative selection, based on an entry criterion common for all students and a selection criterion where their performance is compared with other students.
- Another instance, the institute ranking is a comparative scale among other institutes offering same academic functions.
- Similarly, student performance even though is absolute in nature, but its applied usage is done comparatively.

Considering this very nature of Education field, its important and will be effective if the decisions are taken based on the attributes/factors which serve as differentiators. For example, the number of Internships, communication skills, might have greater influence on the selection of a student in placements, however only after meeting the minimum criteria of entry eligibility.

Minimum Criteria and Maximum Differentiators (MCMD), a technique that can be used when using the Domain Knowledge in determining the Causal Graph. Proposing a generalization for identifying the treatments in a DAG model of a given use case. Using the technique of Minimum Criteria and Maximum Differentiators (MCMD**)**, decisions in Education domain can be taken effectively, based on two factors.

1. The features that must meet a minimum criterion in-order to be eligible for pursuing an aspired outcome – **Treatments**.
2. The features that influence to maximize the chances of outcome by aiding to generate differentiating attributes that favors the interest of an outcome – **Instrument Variables**.

This theorem can be used in deciding the datasets, determining to model the causal graphs, identifying the treatments that can used in decision models.

Framework for deriving the Causal Graph - 3D Framework (Data Correlation, Discovery by Causal ML, Domain Knowledge)

This name is picked from the perspective that the analysis is done from 3 dimensions to determine a graph which can be approximated to the best possibility of the interested outcome. It illustrates a hybrid Approach combining the merits of both manual and causal discovery tools.

1. Based on the **Data Correlation**, by running the correlation on the given dataset, from the Correlation Matrix, pick the most related (+ve and -ve) features onto the interested outcome. This will give the relationship between the features, offcourse mathematically (statistical association). Significantly correlated/related (both +ve / -ve) features are noted as the ***Treatments/Interventions/Causes***, and their corresponding related features are considered as the ***Treatment Variables***. Also, the moderately correlated features are also considered as ***Instrument Variables***.

From the correlation matrix:

a. **Treatments/Interventions/Causes**: those with strong correlation coefficient to the interested outcome (i.e more than 0.5), or the very -vely correlated ones.
b. **Instrument variables:** those with moderate correlation coefficient to the interested outcome (i.e more than 0.2).
c. **Confounders:** those with correlation to both Treatment and Outcome features.

2. By using the **Causal Discovery** algorithms (PC, GES, LiNGAM), we will use the DoWhy package's CausalLearn (CDT-Causal Discovery Toolset) library. By this step, we will have causal model identified CPDAG (completed partial DAG), from which, the nodes which are directly associated to the outcome node will be picked as the ***Treatments*** and the indirectly related nodes are noted as the ***Instrument Variables*** and the ones which are directly related to both the outcome and treatment variables as ***Confounders***.
3. By using the **Domain Knowledge**, we use the MCMD approach (Minimal Criteria and Maximum Differentiators). Based on this, the features which are factored for meeting a minimal required criteria are noted as ***Treatments***, and the features that can enhance / maximize the chances of the outcome in a desired/targeted way are noted as ***Instrument***

Variables. Typically used technique for gathering the inputs from multiple people is called *Event-Storming*.

Combining the above 3 steps, the framework is used to determine the Causal Graph. This results in a causal model which can be effective for decision making towards the interested target outcome factoring the ML techniques validated by the domain expertise.

The table that captures the Treatments, Instrument Variables, Confounders from the above 3 approaches is called the 3D decision table. It's a composite and tabular layout to pick up the features which are identified by the 3 approaches commonly.

Table 1. 3D Table

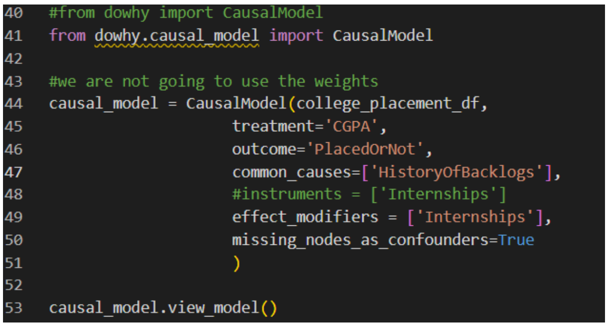
Outcome: <<>>	D#1 Data Correlation	Causal Discovery Toolset			D#3 Domain Knowledge
		PC	GES	LiNGAM	
Treatments					
Instrument Variables					
Confounders					

3.5 Falsifying or validating the identified Causal graph for the correctness

An informative causal graph is one that captures the true causal structure to some extent. It aligns with the actual dependencies and causal mechanisms present in the observed data. In other words, the graph is not purely random or arbitrary. It reflects meaningful relationships. During falsification process, we evaluate whether the given causal graph (DAG) is consistent with the observed data. We test whether the graph adheres to certain statistical properties (such as conditional independence relationships) that are expected based on causal assumptions. The concept of Markov equivalence class is crucial here. Two DAGs are in the same Markov equivalence class if they imply the same set of conditional independence relations. If a given DAG is informative, it means that it lies within the same Markov equivalence class as the true causal graph. When the falsification tests are performed, we compare the given DAG against permuted (randomized) versions of the graph. If the given DAG is significantly better (in terms of adhering to LMCs) than most permuted DAGs, we do not reject it. The p-value associated with the Markov equivalence class informs us about the informativeness of the graph. In summary, an **informative causal graph** provides meaningful insights into the causal relationships, and during falsification, we assess whether it aligns well with the observed data.

3.6 Creating the Causal Model

A causal model is a conceptual or mathematical framework that represents relationships between variables to understand and predict the effects of interventions or changes in the system. It describes how variables are interconnected and how one variable can influence another, often through a series of direct and indirect paths. Causal models help in identifying the cause-and-effect relationships rather than just associations or correlations.



```
#from dowhy import CausalModel
from dowhy.causal_model import CausalModel

#we are not going to use the weights
causal_model = CausalModel(college_placement_df,
                    treatment='CGPA',
                    outcome='PlacedOrNot',
                    common_causes=['HistoryOfBacklogs'],
                    #instruments = ['Internships']
                    effect_modifiers = ['Internships'],
                    missing_nodes_as_confounders=True
                    )

causal_model.view_model()
```

Fig .4. Code Snippet to create Causal Model

Causal models provide the structure and framework for causal inference, enabling researchers to uncover and quantify causal relationships from data.

4. Results and Discussion

The Proposed Model is explored with various stages like the Preprocessing, applying the Correlation model, Causal Discovery techniques and 3D Framework which uses the principle of Minimum Criteria and Maximum Differentiators (MCMD) on the student placement data to form the Directed Acyclic Graph (DAG). The robustness of the graph is checked by falsification process which assesses and tells whether the graph is informative and aligns well with the observed data.

4.1 Dataset/Usecase/ Tools

The dataset used in the research work is Placement of the student by the end of their graduation which is extracted from Kaggle[7]. The dimensions of the dataset are 2967x8 meaning there are 2967 records of students and 8 features in total. The features include age, gender, stream, internships, cgpa, hostel facility,HistoryofBacklogs, PlacedorNot. All the data is numerical and not categorical. In the fields where the values are just 0's and 1's, 0 stands for no, and 1 is for yes.

4.2 Identifying the correlated features using the Correlation matrix

Here, as per the correlation matrix , the highly positively correlated features with the outcome variable PlacedOrNot are CGPA with 0.59 and Internships with 0.18.

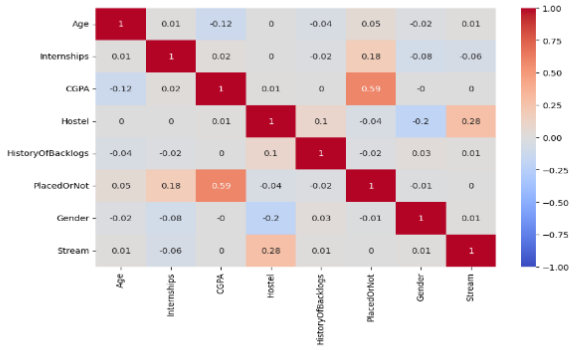


Fig .5. Correlation Matrix showing the Positively and negatively correlated features

4.3 Identifying the treatments and producing the DAG using PC (Peter Clark Algorithm), GES (Greedy Equivalence Search Algorithm), LiNGAM (Non - gaussian Linear causal model Algorithm) for causal discovery

The directed acyclic graphs (DAGs) are used to quantify the experimental design for the educational system that is proposed. Causality attempts to describe the relationship that can exist between two variables [2]. Causality can be usually divided into two main subjects:

- **Causal Inference** which designates a branch of knowledge that examines the presumptions, research plans, and estimating techniques that enable researchers to infer causal relationships from data.
- **Causal Discovery** which is related to the process of discovering causal relationships by analysing the statistical properties of observational data.

The different Causal discovery methods are

- constraint based causal discovery – It aims to infer causal structure from data by leveraging independence structure between the variables. The known algorithms is **PC**
- Score based causal discovery – It generates candidate graphs iteratively, then chooses the best one after assessing how well each one describes the data. **GES** belong to this family.
- constrained functional causal discovery – Based on structural equations that define the causal relationships . **LiNGAM- based** models

Considering the results of PC, GES, LiNGAM methods for the causal discovery in the proposed work.

Peter – Clark(PC) Algorithm

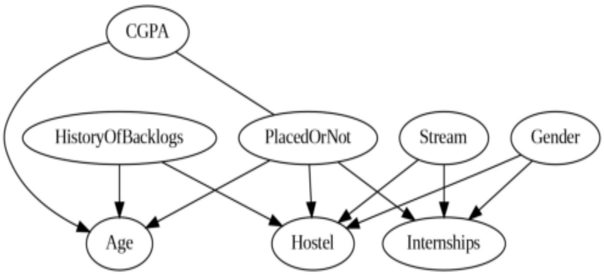


Fig .6. Snapshot of Final Causal graph using PC Algorithm

The above graph gives a CPDAG (Completed Partially Directed Acyclic Graph). From the features of the placement dataset, the PlacedOrNot (outcome) is caused by CGPA (Treatment). Instrument variables & confounders are not present.

Greedy Equivalence Search (GES) algorithm

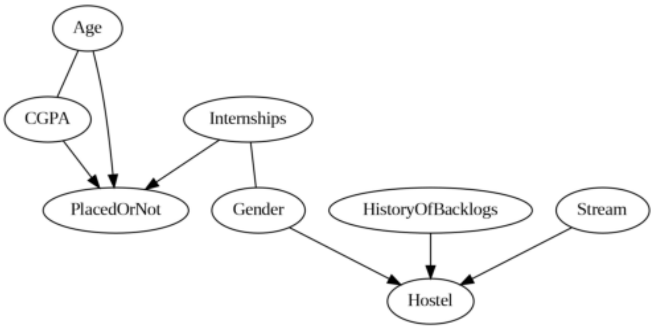


Fig .7. Snapshot of Final Causal graph using GES Algorithm

From the Graph, Treatments – CGPA,

Internships Outcome – PlacedOrNot,

Confounder – Age,

Instrument Variable – Gender

Linear, Non- Gaussian Acyclic Model (LiNGAM)

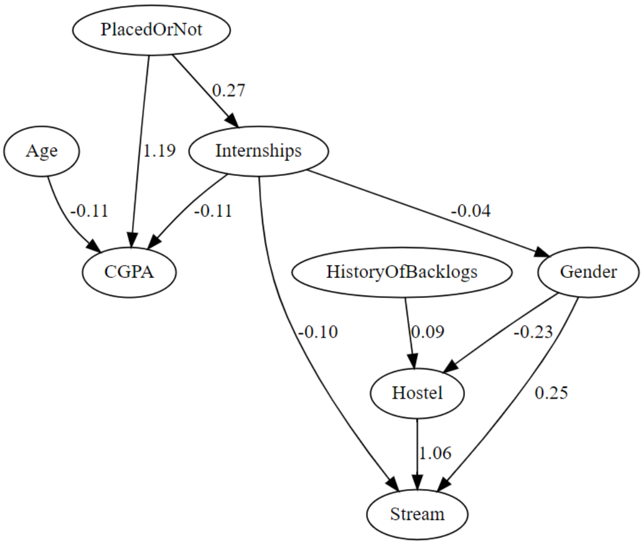


Fig .8. Snapshot of Final Causal graph using LiNGAM Algorithm

From the Graph,

Treatments – CGPA, Internships,

Outcome – PlacedOrNot

Confounder – Nil,

Instrument Variable – Gender, Age

4.4 3D Framework – Hybrid Approach

Determining the DAG involves 3 ways to execute. It uses a hybrid approach, a composite render of ML correlation-based relations + Causal ML based causal discovery + application of Domain knowledge to determine the DAG.

Table 2. Example 3D from the implementation

Outcome: PlacedOrNot	D#1 Data Correlation	Causal Discovery Toolset			D#3 Domain Knowledge
		PC	GES	LiNGAM	
Treatments	CGPA	CGPA	CGPA, Internship	Internships, CGPA,	CGPA (MC)
Instrument Variables	Internships	-	Gender	Age, Gender	Internships (MD) Gender (MD)
Confounders	HistoryOfBacklogs, Gender	-	Age	-	HistoryOfBacklogs

The method for discussing and determining the variables by the domain experts is called as **Event-Storming**.

From the above table:

Treatments: CGPA

Instrument Variables: Internships, Gender

Confounders: HistoryOfBacklogs

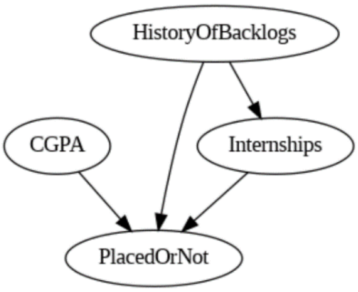
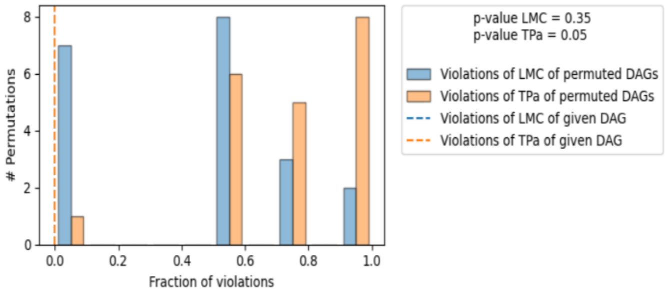
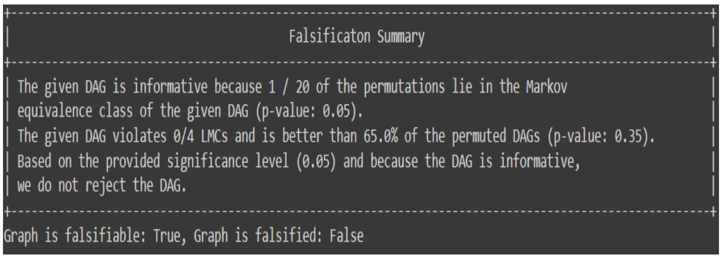


Fig. 9. Snapshot of Final Causal graph using the Hybrid(3D) Approach

4.5 The **falsify_graph** function is part of the **DoWhy** library, which provides tools for causal inference. Specifically, it focuses on falsifying causal graphs (Directed Acyclic Graphs or DAGs) using observational data.





```
+------------------------------------------------------------------------------------------+
|                                  Falsificaton Summary                                    |
+------------------------------------------------------------------------------------------+
| The given DAG is informative because 1 / 20 of the permutations lie in the Markov       |
| equivalence class of the given DAG (p-value: 0.05).                                      |
| The given DAG violates 0/4 LMCs and is better than 65.0% of the permuted DAGs (p-value: 0.35). |
| Based on the provided significance level (0.05) and because the DAG is informative,     |
| we do not reject the DAG.                                                                |
+------------------------------------------------------------------------------------------+
Graph is falsifiable: True, Graph is falsified: False
```

Fig. 10. Output after the Falsifying or validating the identified Causal graph for the correctness

Here considering the placements in an educational institution after analysing the correlations and establishing the causation produces the following Causal Model. The **Causal Model or framework** reveals that the PlacedOrNot feature can be influenced by the changes made to the CGPA, Internships features directly which are called as **Treatments** and can be influenced by the HistoryOfBacklogs indirectly which are called as **Confounders**.

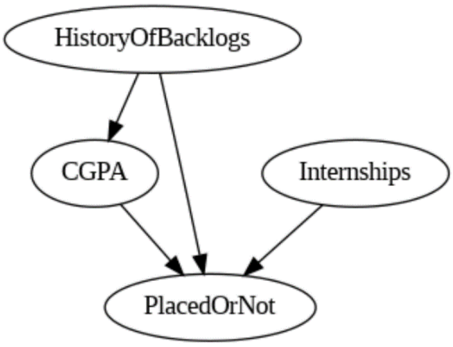


Fig .11. Causal Model

5. Conclusion & Future Work

In this work, the various Causal Discovery Algorithms are implemented, and the resultant causal graphs are analysed with the Minimum Criteria and Maximum Differentiator (MCMD) concept of Domain Knowledge to form the final validated Causal Graph (DAG). The Treatments, outcome, Confounders are Identified, and the robust Causal Model is built. The Future work would involve taking the causal model as input for the estimation of effect of Treatments on the outcomes for this placement use case. It is based on the mean value and the p-value (significance Level) calculations, So that we can quantify that the estimated treatment effect on the outcome. The refutation of the effect estimate is also done for robust results. Ultimately, a strong decision-making model is formed through the communication of the treatments as decisions are chosen for the desired outcomes.

6. Conflicts of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agreed with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

7. Author contributions

Conceptualization, methodology, software, validation, formal analysis, writing-original paper draft, *Naga Jyothi*; writing-review and editing, *Dr. Uma Dulhare*; visualization, *Naga Jyothi*; supervision, *Dr. Uma Dulhare.*

References

[1] Tadayon, Manie, and Greg Pottie. "**Causal inference in educational systems: A graphical modeling approach**." *arXiv preprint arXiv:2108.00654* (2021).

[2] Ouaadi, I., & Ibourk, A. (2023, July). **Causal Discovery and Features Importance Analysis: What Can Be Inferred About At-Risk Students?.** In *International Conference on Business Intelligence* (pp. 134-145). Cham: Springer Nature Switzerland.

[3] Sharma, Amit, and Emre Kiciman. "DoWhy: An end-to-end library for causal inference." *arXiv preprint arXiv:2011.04216* (2020).

[4] Dominik Janzing, Lenon Minorics, Patrick Blöbaum. **Feature relevance quantification in explainable AI: A causal problem** International Conference on Artificial Intelligence and Statistics, 2907-2916, 2021.

[5] Weidlich, Joshua, Ben Hicks, and Hendrik Drachsler. "**Causal reasoning with causal graphs in educational technology research**." *Educational technology research and development* (2023): 1-19.

[6] Burnett, J. Wesley, and Calvin Blackwell. "Graphical causal modelling: an application to identify and estimate cause-and-effect relationships." *Applied Economics* (2023): 1-15.

[7] Forney, Andrew, and Scott Mueller. "**Causal inference in AI education: A primer.**" *Journal of Causal Inference* 10, no. 1 (2022): 141-173.

[8] Kaur, Prableen, Agoritsa Polyzou, and George Karypis. "**Causal inference in higher education: Building better curriculums.**" In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pp. 1-4. 2019.

[9] de Carvalho, Walisson Ferreira, and Luis Enrique Zarate. "**Causality relationship among attributes applied in an educational data set**." In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pp. 1271-1277. 2019.

[10] Ellison, George TH. "Introducing causal inference to the medical curriculum using temporal logic to draw directed acyclic graphs." *medRxiv* (2020): 2020-08.

[11] Kaddour, Jean, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. "**Causal machine learning: A survey and open problems.**" *arXiv preprint arXiv:2206.15475* (2022).

[12] https://www.pywhy.org/dowhy/main/example_notebooks/tutorial-causalinference-machinelearning-using-dowhy-econml.html

[13] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf. **Quantifying causal influences** The Annals of Statistics, Vol. 41, No. 5, 2324-2358, 2013.

[14] Kudikyala, Udai Kumar, and Uma N. Dulhare. "Using Scrum and Wikis to manage student major projects." In *2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pp. 15-20. IEEE, 2015.

[15] Arif, Fayeza, and Uma N. Dulhare. "A machine learning based approach for opinion mining on social network data." In *Computer Communication, Networking and Internet Security: Proceedings of IC3T 2016*, pp. 135-147. Springer Singapore, 2017.

[16] Dulhare, Uma N., Khaleel Ahmad, and Khairol Amali Bin Ahmad, eds. *Machine learning and big data: concepts, algorithms, tools and applications*. John Wiley & sons, 2020.

[17] U.N. Dulhare, D.N. Jyothi, B. Balimidi, and R. R. Kesaraju, “Classification Models in Education Domain Using PSO, ABC, and A2BC Metaheuristic Algorithm-Based Feature Selection and Optimization”, In Machine Learning and Metaheuristics: Methods and Analysis, pp. 255-270, Singapore: Springer Nature Singapore, 2023.

[18] Jyothi, D. Naga, and Uma N. Dulhare. "Student Learning Based Data Science Assisted Recommendation System to Enhance Educational Institution Performance." *International Journal of Intelligent Engineering & Systems* 17, no. 4 (2024).