International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 850-872
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Understanding and Analysing Causal Relations through Modelling using Causal Machine Learning

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

[1] Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India
* **Corresponding Author Email:** dnagajyothi_cseaiml@cbit.ac.in - **ORCID:** 0000-0001-9498-2882

[2] Muffakam Jah College of Engg. And Tech, Hyderabad, Telangana, India
Email: uma.dulhare@mjcollege.ac.in - **ORCID:** 0000-0002-4736-4472

Article Info:

DOI: 10.22399/ijcesen.1018
Received : 03 December 2024
Accepted : 11 February 2025

Keywords :

Causal Inference,
Machine Learning,
Directed Acyclic Graphs,
DoWhy Library,
Causal Discovery Algorithms,
Student Placement.

Abstract:

The study of causal inference has gained significant attention in artificial intelligence (AI) and machine learning (ML), particularly in areas such as explainability, automated diagnostics, reinforcement learning, and transfer learning.. This research applies causal inference techniques to analyze student placement data, aiming to establish cause-and-effect relationships rather than mere correlations. Using the DoWhy Python library, the study follows a structured four-step approach—Modeling, Identification, Estimation, and Refutation—and introduces a novel 3D framework (Data Correlation, Causal Discovery, and Domain Knowledge) to enhance causal modeling reliability. Causal discovery algorithms, including Peter Clark (PC), Greedy Equivalence Search (GES), and Linear Non-Gaussian Acyclic Model (LiNGAM), are applied to construct and validate a robust causal model. Results indicate that internships (0.155) and academic branch selection (0.148) are the most influential factors in student placements, while CGPA (0.042), projects (0.035), and employability skills (0.016) have moderate effects, and extracurricular activities (0.004) and MOOCs courses (0.012) exhibit minimal impact. This research underscores the significance of causal reasoning in higher education analytics and highlights the effectiveness of causal ML techniques in real-world decision-making. Future work may explore larger datasets, integrate additional educational variables, and extend this approach to other academic disciplines for broader applicability.

1. Introduction

Causal inference has emerged as a critical area of research in machine learning, offering a robust framework for understanding cause-and-effect relationships beyond traditional correlation-based models. While machine learning techniques have been widely employed in various domains, including education, their predominant focus on predictive accuracy often overlooks the underlying causal mechanisms influencing outcomes. This limitation is particularly significant in the context of student placement prediction, where identifying the key determinants of placement success is essential for informed decision-making by educational institutions [1]. By integrating causal inference methodologies, this study aims to bridge this gap by establishing a systematic approach to uncovering the causal relationships between academic and non-academic factors that impact student employability. Through the application of causal discovery tools and statistical modelling, this research provides a comprehensive analysis that enhances the interpretability and reliability of predictive models in the education domain [2].

1.1 Background and Motivation

The integration of machine learning (ML) techniques in the education sector has significantly improved predictive analytics, enabling institutions to forecast student performance, recommend personalized learning paths, and assess employability potential [3]. These advancements have facilitated data-driven decision-making, enhancing the overall efficiency of academic and career guidance systems. However, traditional ML models primarily rely on correlation-based methods,

which, while effective for pattern recognition, fail to capture the underlying causal mechanisms that drive outcomes. In the context of student placements[4], existing models often identify associations between variables—such as CGPA, internships, and extracurricular activities—and employment success but do not establish whether these factors directly influence placement outcomes or are confounded by unobserved variables[5]. This limitation raises critical concerns regarding the interpretability and generalizability of ML-driven placement prediction models. Correlation-based approaches may lead to spurious relationships, misinforming institutional policies and intervention strategies. For instance, a model may indicate a strong correlation between high academic performance and placement success without considering whether other latent factors, such as industry exposure or employability skills, mediate this relationship. To address this gap, causal inference techniques provide a structured methodology to distinguish true causal effects from mere associations, thereby enabling more reliable decision-making[6]. By leveraging causal inference, this study aims to construct a data-driven framework that identifies and quantifies the causal impact of key academic and non-academic factors on student placement outcomes. This approach not only enhances the robustness of predictive models but also equips educational institutions with actionable insights to refine their placement strategies, optimize training programs, and support students more effectively in their transition to the job market[7].

1.2 Problem Statement

Accurately predicting student placement outcomes remains a significant challenge for educational institutions, as existing predictive models predominantly rely on correlation-based techniques rather than causal analysis. Traditional machine learning approaches identify statistical associations between factors such as academic performance, internships, and extracurricular activities with placement success. However, these models do not establish whether these factors directly influence employment outcomes or if their effects are confounded by other latent variables. This reliance on correlation-based predictions limits the interpretability and effectiveness of decision-making processes in academic institutions. Without a clear understanding of causal relationships, institutions may implement policies or training programs based on misleading insights, resulting in suboptimal resource allocation and ineffective interventions. For instance, a model might suggest that students with higher CGPA[8] are more likely to secure job placements, but without causal inference, it remains unclear whether CGPA itself is the primary determinant or if other factors—such as employability skills, industry exposure, or structured placement training—mediate this relationship. Consequently, interventions based solely on correlation-based predictions risk being ineffective or even counterproductive[9]. To address this issue, there is a critical need for causal inference methodologies that can differentiate between correlation and causation, ensuring that the factors influencing student placement outcomes are correctly identified. By leveraging causal discovery techniques and structured statistical modeling, this study seeks to bridge this gap, providing a more reliable and interpretable framework for understanding student employability and optimizing placement strategies.

1.3 Objectives of the Study

The primary objective of this study is to investigate the causal impact of various academic and non-academic factors on student placement outcomes using causal machine learning techniques. Unlike traditional predictive models that establish statistical associations, this research aims to develop a structured causal framework to determine whether specific factors directly influence placement success. By leveraging causal discovery methods, the study seeks to identify key determinants—such as CGPA, internships, employability skills, and extracurricular activities—and assess their causal relationships with employment outcomes. Additionally, the research intends to validate the robustness of these causal estimations through systematic statistical techniques, ensuring the reliability and generalizability of the findings. Ultimately, the insights derived from this study will contribute to more effective decision-making in educational institutions, enabling data-driven interventions to enhance student employability.

1.4 Significance of Causal Inference in Educational Analytics

Causal inference plays a crucial role in optimizing student placement strategies by providing a systematic approach to distinguish between mere associations and genuine cause-effect relationships. Traditional machine learning models, while proficient in identifying correlations, do not offer actionable insights into how modifying specific variables—such as improving soft skills training or increasing internship opportunities—can impact placement rates. The application of causal reasoning in educational analytics allows institutions to move beyond predictive accuracy and focus on designing targeted interventions that yield measurable

improvements in student employability. By integrating causal inference, educational policymakers can make evidence-based decisions regarding curriculum development, placement training programs, and resource allocation. For example, if causal analysis reveals that internships significantly influence placement outcomes, institutions can prioritize industry collaborations and internship opportunities as a strategic intervention.. This study highlights the transformative potential of causal ML techniques in the education sector, advocating for their integration into institutional decision-making processes to improve student success rates.

1.5 Overview of the Proposed Approach

To achieve the study's objectives, a structured methodological framework is developed, integrating causal discovery algorithms and domain knowledge to construct a robust causal model. The research introduces a novel 3D Framework, which comprises three critical components:

1. Data Correlation – Analyzing statistical associations between student attributes and placement outcomes.
2. Causal Discovery – Employing causal inference techniques such as DoWhy and causal discovery algorithms (Peter Clark (PC), Greedy Equivalence Search (GES), and Linear Non-Gaussian Acyclic Model (LiNGAM)) to infer causal structures.
3. Domain Knowledge – Incorporating expert knowledge to validate and refine causal relationships, ensuring the model's interpretability and real-world applicability.

The proposed framework systematically models, identifies, estimates, and refutes causal effects, ensuring a rigorous approach to causal inference in student placement prediction. By leveraging DoWhy, an open-source Python library for causal analysis, this study aims to establish a validated causal graph (DAG) that accurately represents the key drivers of placement success. The findings of this research are expected to contribute to the development of more effective educational policies, placement strategies, and personalized career guidance systems, ultimately enhancing the employability prospects of students.

2. Related Work

This section provides a comprehensive review of existing literature on causal inference in machine learning, causal discovery techniques, and their applications in educational analytics and student placement prediction. While traditional machine learning models have been extensively used in educational research, their reliance on correlation-based techniques limits their ability to infer causality. Recent advancements in causal machine learning have introduced more sophisticated frameworks that enable a deeper understanding of the cause-and-effect relationships in student performance and employability outcomes. This section examines the evolution of machine learning applications in education, the role of Directed Acyclic Graphs (DAGs) in causal discovery, and the significance of causal inference in student placement prediction.

2.1 Machine Learning Applications in Education

Machine learning has been widely utilized in education for various applications, including student performance prediction, dropout prevention, automated grading systems, and personalized learning environments [10]. Predictive models have been employed to forecast students' academic success by analyzing features such as attendance, assessment scores, and engagement in learning activities [11]. Additionally, deep learning techniques have been integrated into adaptive learning platforms to recommend personalized content based on students' learning styles and cognitive abilities [12].Despite these advancements, a critical limitation of traditional machine learning models is their reliance on correlation-based techniques, which do not establish causal relationships between educational factors and student outcomes [13]. For instance, a model may predict that students with higher CGPA are more likely to secure job placements, but it does not determine whether CGPA directly influences placement success or if other underlying variables, such as employability skills and internship experiences, mediate this relationship [14]. This shortcoming has raised concerns regarding the interpretability and applicability of these models in educational decision-making [15]. To address this limitation, researchers have explored causal inference techniques that enable a more rigorous analysis of the factors influencing student outcomes [16]. Causal models, particularly those employing Directed Acyclic Graphs (DAGs), have been introduced to identify confounders, mediators, and instrument variables, thereby providing a more reliable framework for educational interventions [17]. Studies have demonstrated that incorporating causal inference into student performance and placement prediction models enhances their robustness and ensures that institutional policies are informed by actionable insights rather than spurious correlations [18].By leveraging causal discovery

techniques, this study aims to bridge the existing gap in student placement prediction by distinguishing between correlation and causation, enabling educational institutions to design targeted interventions that improve student employability outcomes.

2.2 Causal Inference in Artificial Intelligence and Machine Learning

The significance of causal inference in artificial intelligence (AI) and machine learning (ML) has been increasingly recognized, particularly in decision-making systems that require interpretability and robustness. Traditional ML models predominantly rely on correlation-based learning, wherein statistical dependencies between variables are leveraged for predictive tasks. However, these models fail to differentiate between causation and correlation, leading to challenges in generalizability and explainability in real-world applications [19]. For instance, predictive models in healthcare and finance often suggest associations between variables without verifying whether changes in one factor directly influence another [20]. Causal machine learning offers an alternative approach that incorporates structured causal modeling techniques to identify, estimate, and validate cause-effect relationships. Studies have demonstrated that integrating causal inference methods improves decision-making in domains such as healthcare, economics, and education by enabling more reliable interventions [21]. Methods such as structural causal models (SCMs) and counterfactual analysis have been applied to mitigate selection bias and confounding factors in ML predictions, leading to more interpretable and actionable insights [22]. The growing interest in causal ML has led to the development of specialized frameworks, such as DoWhy, which facilitate causal reasoning in AI applications [23]. Given the increasing complexity of data-driven decision-making, the adoption of causal inference methodologies in AI is essential for ensuring reliable, explanation-aware models that support decision-makers across various domains.

2.3 Directed Acyclic Graphs (DAGs) and Causal Discovery Techniques

Directed Acyclic Graphs (DAGs) serve as a fundamental tool in causal inference, enabling the visualization and quantification of causal relationships among variables. DAGs offer a structured approach to representing causal dependencies by ensuring acyclic relationships between nodes, thereby preventing feedback loops and reinforcing a clear causal hierarchy [24]. Their application spans multiple disciplines, including economics, epidemiology, and education, where understanding the causal impact of interventions is critical [25]. Several causal discovery techniques have been proposed to infer DAG structures from observational data, each employing distinct methodologies to establish causality. The Peter Clark (PC) algorithm, a constraint-based method, determines causal structures by iteratively testing conditional independence among variables [26]. The Greedy Equivalence Search (GES) algorithm, a score-based approach, evaluates candidate DAGs based on goodness-of-fit criteria such as the Bayesian Information Criterion (BIC), iteratively refining causal structures to optimize model accuracy [27]. Meanwhile, the Linear Non-Gaussian Acyclic Model (LiNGAM) assumes that causal relationships follow a linear structure with non-Gaussian noise, making it particularly suitable for domains where traditional Gaussian assumptions do not hold [28].Although these algorithms provide valuable frameworks for causal discovery, challenges remain in handling high-dimensional datasets, addressing latent confounders, and integrating domain expertise into automated discovery processes. The effectiveness of DAG-based causal modeling depends on the robustness of assumptions and the quality of observational data, necessitating hybrid approaches that combine automated causal discovery with expert-driven validation.

2.4 Causal Inference in Educational Analytics

The application of causal inference in educational analytics has gained traction as researchers seek to improve student learning outcomes, academic performance, and employability prospects [29]. While traditional predictive models have been employed to assess student success, they often fail to identify causal mechanisms underlying academic achievement and workforce readiness. Causal modeling in education enables policymakers and institutions to design evidence-based interventions, such as targeted skill development programs and personalized learning pathways [30].Studies have applied causal inference to various aspects of education, including the impact of learning strategies on student performance, the effectiveness of online learning platforms, and the role of socioeconomic factors in academic success [31]. For instance, research leveraging instrumental variable (IV) techniques has shown that structured mentorship programs have a significant causal effect on student retention rates [32]. Similarly, propensity score matching (PSM) has been used to evaluate the

effectiveness of employability training initiatives, ensuring that observed differences in placement outcomes are attributable to the intervention rather than selection bias [33].

Despite these advancements, limited work has been done in applying causal ML techniques to student placement prediction, highlighting an area that requires further exploration. This study seeks to address this gap by developing a causally informed model for understanding student employability factors, integrating DAG-based causal discovery and domain knowledge validation.

2.5 Causal Machine Learning for Student Placement Prediction

Accurate student placement prediction is a crucial challenge for educational institutions, as traditional models primarily rely on correlation-based learning, which does not account for causal dependencies between student attributes and employability outcomes [34]. Previous studies have identified various predictors, such as CGPA, internships, certifications, and soft skills, but have not established whether these factors directly impact placement success or if their effects are mediated by unobserved variables [35].Early attempts at causal inference in student placement analysis have employed propensity score matching (PSM) and instrumental variable (IV) approaches to reduce confounding bias [36]. However, these methods often require strong assumptions about treatment assignment, limiting their applicability in complex, multi-factorial placement scenarios. More recent work has explored DAG-based modeling to infer causal pathways between student attributes and job placement probabilities [37]. This study extends prior research by introducing a hybrid causal framework that integrates data correlation, causal discovery algorithms, and domain knowledge validation, ensuring greater interpretability and reliability in placement predictions. The proposed methodology offers a robust decision-making tool for academic institutions seeking to refine placement training programs and optimize student employability.

2.6 Limitations and Research Gaps in Existing Studies

While causal inference has shown promise in educational analytics, several limitations persist in existing studies. One of the most pressing issues is the reliance on observational data without causal validation, leading to potential biases in interpreting student success factors [38]. Many studies use association-based metrics, such as linear regression coefficients and correlation matrices, without establishing true causal effects [39]. This limitation restricts the applicability of findings, as institutions may implement ineffective policies based on spurious relationships.Another gap in current research is the lack of hybrid approaches integrating causal discovery with domain expertise. Automated causal inference techniques, while powerful, often require human validation to ensure that inferred causal relationships align with theoretical and contextual knowledge [40]. Additionally, high-dimensional datasets in education pose challenges for existing causal discovery algorithms, as they require large sample sizes and computationally intensive procedures to yield reliable causal estimates [41]. Furthermore, causal ML models in student placement prediction have not been extensively evaluated across diverse educational settings. Most studies focus on a single dataset or institution, limiting the generalizability of causal inferences [42]. To address these gaps, this study proposes a three-dimensional (3D) framework that integrates statistical correlation analysis, causal discovery algorithms, and expert-driven validation, ensuring a more interpretable and actionable causal model for student placement prediction.

3. Theoretical Foundations of Causal Machine Learning

3.1 The Evolution of Machine Learning in Educational Analytics

Machine learning (ML) has been widely adopted in educational analytics to enhance student performance prediction, provide personalized learning experiences, and improve institutional decision-making [43]. These advancements have enabled institutions to develop automated grading systems, early warning systems for at-risk students, and student placement prediction models. However, despite their predictive power, traditional ML models primarily focus on correlation-based methods, which fail to establish causal relationships between variables. For instance, a predictive model may identify a strong correlation between a student's CGPA and placement success, but this does not imply that CGPA directly influences placement outcomes. Instead, other latent factors, such as internship experience, employability skills, or mentorship programs, may mediate this relationship. This limitation significantly affects decision-making in educational institutions, as interventions based solely on correlations may lead to misguided

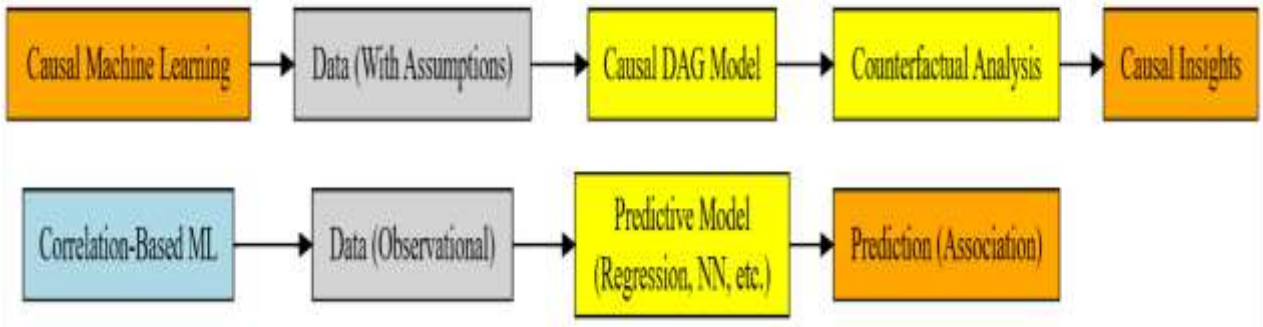


Figure 1. *Comparison of Correlation-Based ML vs. Causal ML*

policies. To address these challenges, researchers have emphasized the integration of causal inference techniques in ML-based educational analytics. By adopting causal machine learning (CML), institutions can identify key determinants of student employability and implement evidence-based policies that enhance learning and placement outcomes. Figure 1 is comparison of correlation-based ML vs. causal ML. Figure 2 is counter factual scenario in causal ML.

3.2 The Need for Causal Machine Learning

Traditional machine learning models are designed to optimize prediction accuracy based on **observational data**. However, these models struggle with **explainability**, **bias**, and **generalization to unseen scenarios**. In high-stakes domains such as **education and workforce analytics**, mere predictions are insufficient; institutions require **actionable insights** based on cause-and-effect relationships [44].

Causal ML provides a framework for answering **counterfactual questions**, such as:

- Would a student's placement outcome have changed if they had completed an internship?
- How much does employability training causally impact job placement success?
- If students were randomly assigned to different academic branches, would placement rates remain the same?

By addressing these causal questions, CML enables data-driven decision-making and ensures that educational interventions are targeted and effective.

3.3 Defining Causal Inference and Its Role in Machine Learning

Causal inference aims to determine whether a change in one variable (treatment) leads to a change in another variable (outcome), while controlling for

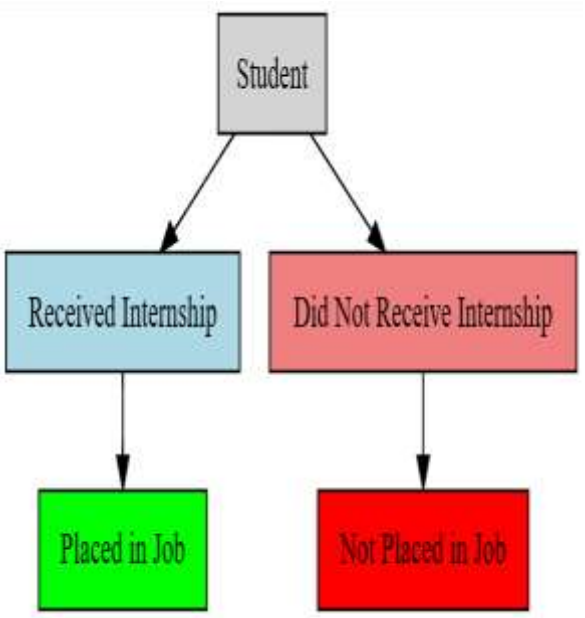


Figure 2. *Counter factual Scenario in Causal ML*

confounding factors. Unlike conventional ML models, which infer associations from data, causal ML establishes direct cause-and-effect relationships through structural causal models (SCMs) and Directed Acyclic Graphs (DAGs). A DAG is a graphical representation of causal relationships where nodes represent variables, and directed edges indicate causal influence. DAGs help in identifying confounders, mediators, and instrument variables, which are essential for unbiased causal estimation [45]. For instance, consider a student placement prediction model where Internship Experience (T) is hypothesized to influence Placement Success (Y). However, this relationship may be confounded by CGPA (X), as higher CGPA students are more likely to receive internships. In this case, failing to account for CGPA as a confounder would lead to biased causal estimates. Figure 3 is DAG showing causal relationships (Internship → Placement, Confounded by CGPA).

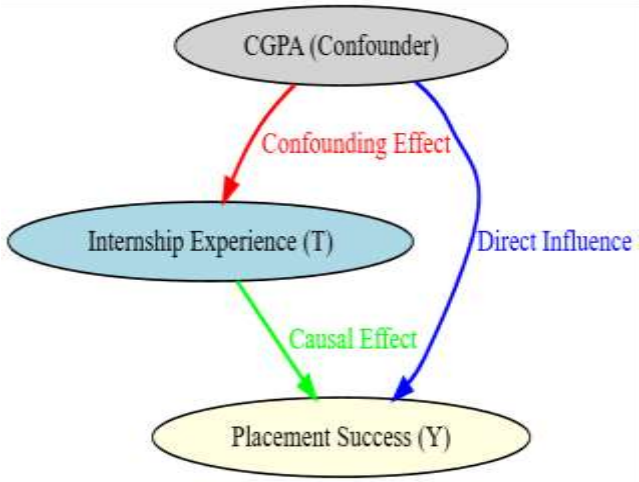


Figure 3. *DAG Showing Causal Relationships (Internship → Placement, Confounded by CGPA)*

3.4 Methodological Framework for Causal Analysis

The process of causal analysis in machine learning follows four key steps: Modeling, Identification, Estimation, and Refutation. Each step ensures that the causal effect of an intervention or treatment is correctly estimated and validated. Modeling involves defining a **conceptual framework** to represent causal relationships between variables. The key steps include:

- Defining the Research Question – Clearly stating the causal question being investigated.
- Developing a Theoretical Model – Constructing a DAG to outline causal pathways.
- Specifying Variables – Identifying treatment (T), outcome (Y), and confounders (X).

The identification step ensures that the causal effect can be **isolated from confounding factors**. This process involves:

- **Stating Assumptions** – Ensuring there are no unmeasured confounders and selecting appropriate instruments.
- **Choosing an Identification Strategy**, such as:
- **Randomized Controlled Trials (RCTs)** – Gold standard for causal inference [46].
- **Natural Experiments** – Leveraging external events that approximate randomization.
- **Instrumental Variables (IV)** – Using instruments that affect treatment but not the outcome directly.
- **Difference-in-Differences (DiD)** – Comparing changes in treated vs. control groups over time.

This step quantifies the causal effect using **statistical methods**:

- **Linear/Non-linear Regression** – Adjusting for confounders to estimate causal effects.
- **Matching Techniques** – Pairing treated and control units with similar characteristics.
- **Instrumental Variable Estimation** – Addressing endogeneity in causal inference.
- **Propensity Score Matching (PSM)** – Matching units based on the probability of treatment assignment.

Refutation techniques **validate the robustness of causal findings**:

- **Placebo Tests** – Ensuring no causal effect is observed where it should not exist.
- **Subgroup Analysis** – Checking if the causal effect holds across different subgroups.
- **Falsification Tests** – Testing whether causal assumptions hold in independent settings.

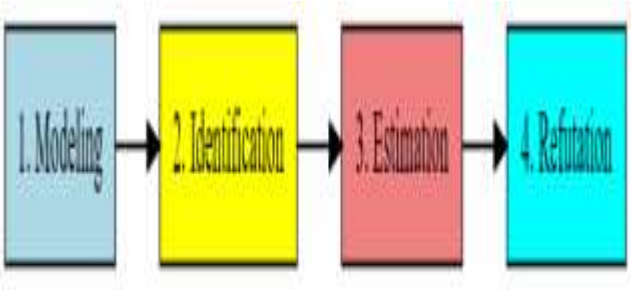


Figure 4. *Workflow diagram showing Modeling → Identification → Estimation → Refutation in causal inference.*

3.5 Challenges in Causal Machine Learning

Despite its advantages, causal ML presents several challenges:

- Observational data often suffer from selection bias, where certain groups may be over- or under-represented in the dataset.
- Example: If only students with strong academic records receive internships, the estimated effect of internships on placement may be overstated.
- Confounders introduce bias if not accounted for properly.
- Example: If students with high employability skills tend to enroll in internship programs, then the true causal effect of internships on placement must adjust for employability skills.
- Many causal ML models are developed using data from a single institution, limiting their

generalizability across different educational settings.

- Causal discovery algorithms, such as PC, GES, and LiNGAM, require large sample sizes and significant computational resources, making them challenging to scale.[47]

4. Methodology Overview

This section explains the implementation of the suggested work, which involves creating a causal graph that addresses variables named treatments, outcomes, confounders, and instrument variables using the Placement Dataset and additional causal discovering techniques. To combine the elements of the various causal discovery algorithms with the domain knowledge required for causal modelling, a novel structure known as the 3D Framework is put forth and use. Figure 4 is workflow diagram.
DoWhy is a Python module that aims to promote causal analysis and reasoning in a manner similar to the predictive machine learning frameworks. For root cause analysis, interventions, effect estimation, prediction, quantification of causal influences, learning causal structures, and creating counterfactuals, DoWhy provides a wide range of algorithms. It can validate causal assumptions for any estimating method. DoWhy's causal job execution process begins with modelling causal interactions as a causal graph. Causal graphs are used to model "cause-effect-relationships" that exist inside a system domain. A Directed Acyclic Graph (DAG) with an edge X→Y designating that X is the cause of Y is what we require for the causal graph. The statistical representation of the conditional independence relationships between variables is a causal graph.

4.1 Directed Acyclic Graphs (DAGs)

They play a crucial role in causal discovery. They are used to visually and formally represent the causal relationships between variables in a system. It is made up of Nodes and Edges where each node represents a variable, and each directed edge (arrow) represents a causal influence from one variable to another. The graph is acyclic, meaning it does not contain any loops, which ensures that it represents a consistent causal ordering. Causal discovery algorithms are computational methods used to uncover causal relationships from observational data. These algorithms aim to determine the underlying causal structure among a set of variables without the need for experimental intervention. Some common types of causal discovery algorithms are Constraint based methods (PC - Peter Clark Algorithm), Score based methods (GES - Greedy Equivalence Search Algorithm), Functional Causal methods (LiNGAM - Linear Non Gaussian Acyclic Model), hybrid methods.

4.2 Causal Discovery Algorithms

The process of deriving causal correlations between variables using observational data is known as causal discovery. Finding a set of causal relationships that make sense for the given is the aim of causal discovery. A directed acyclic graph (DAG) with the variables as nodes and the causal relationships as edges is a representation of the causal relationships. The two types of algorithms are constraint-based and score-based. A scoring function is used by score-based algorithms to assess a causal graph's quality. Constraint-based algorithms assess a causal graph's quality using a set of constraints. The more adaptable score-based methods can be used to identify causal graphs that deviate from the data. Only causal networks that are compatible with the data can be found by the more limited constraint-based methods. Compared to score-based algorithms, constraint-based algorithms are more efficient. Both constraint- and score-based approaches' benefits are utilised by the causal discovery module. You can define the search space's bounds by applying your domain expertise. We use a score-based technique to identify a causal graph that satisfies the given requirements with the available data.

4.3 Peter Clark Algorithm

This algorithm is based on constraints. To locate a causal graph that is consistent with the given constraints and the data, the PC technique is employed. The PC algorithm is a greedy search method that adds edges to a graph iteratively after beginning with an empty network. If the conditional independence test between two variables fails given the other variables in the graph, the algorithm adds an edge between the two variables. When the graph can no longer contain any more edges, the algorithm comes to an end. It is guaranteed by the PC algorithm to find a causal graph consistent with the given constraints and data. Every common cause for every pair of variables is included in the collection of observed variables. It adheres to the common markov condition, which states that, every variable in the causal graph is independent of its non-descendants given its direct causes (parents). The variables are placed in a causal order such that, for any pair of variables, the variable with a direct causal relationship between them comes first in the ordering. The correctness of these presumptions affects how well the algorithm performs, and failures can result in inaccurate or lacking causal structures.

4.4 Greedy Equivalence Search Algorithm

A popular technique for causal discovery is the GES (Greedy Equivalence Search) algorithm, which takes observational data and uses it to determine the causal structure, which is usually represented as a directed acyclic graph, or DAG. It operates by searching through the space of DAGs to find one that best represents the underlying causal relationships. The GES algorithm is divided into two main phases. The forward phase is starting from an empty graph, the algorithm iteratively adds edges that most increase a scoring criterion (usually based on a score like the Bayesian Information Criterion or BIC) until no more additions can recover the score. The Backward phase starts from the graph obtained and the algorithm iteratively removes edges that most increase the score until no more removals can improve the score.

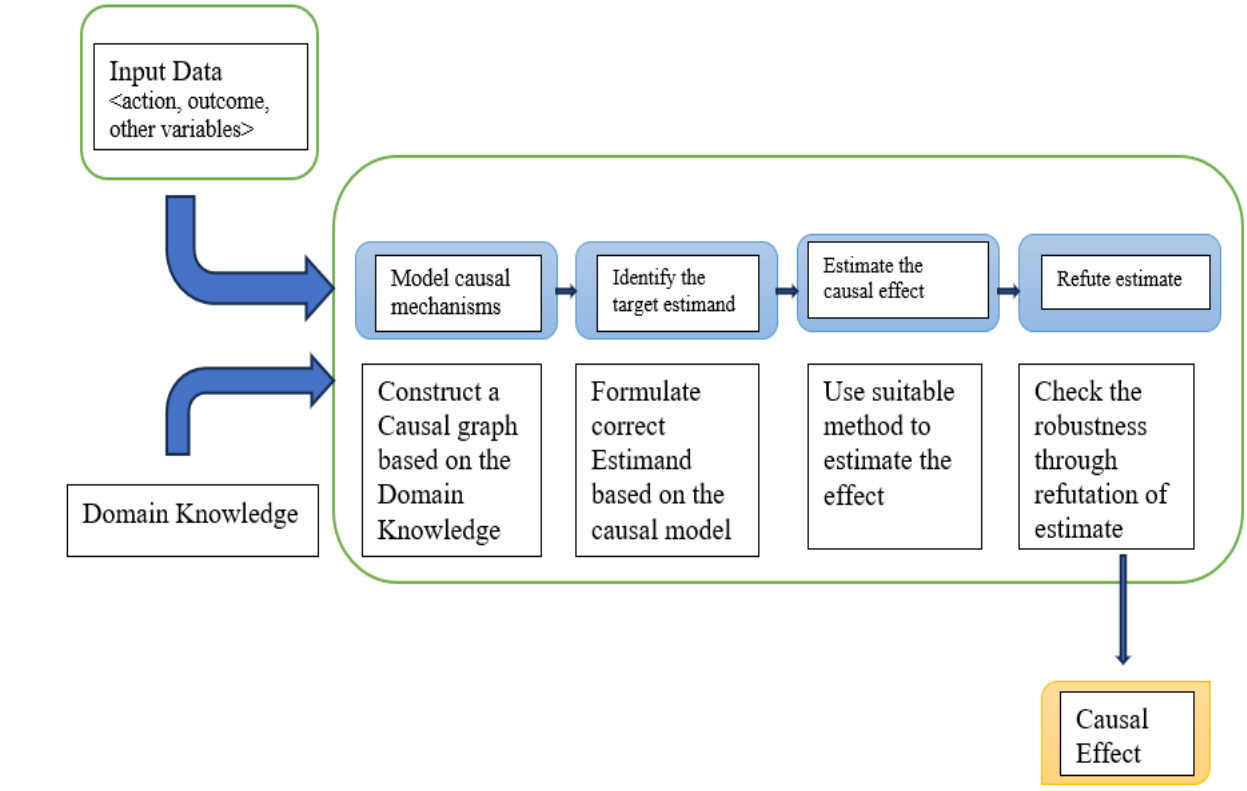
4.5 Linear Non-Gaussian Acyclic Model Algorithm

The causal relationships between the observable variables can be modelled using a linear structural equation model (SEM). In other words, every variable in the causal tree can be understood as a linear function of its independent noise factor plus its parents, or direct causes. It is assumed that the noise terms in the linear SEM are non-Gaussian. The non-Gaussianity is exploited by the DirectLiNGAM algorithm to discriminate between direct and indirect causal relationships. It is assumed that the causal graph is acyclic, which means that it lacks directed cycles and feedback loops.

4.6 3-Dimensional Framework

One method that can be applied when utilising the Domain Knowledge to ascertain the Causal Graph is Minimum Criteria and Maximum Differentiators (MCMD). It offers a generalisation for identifying the treatments of a specific use case in a DAG model. It illustrates a hybrid Approach combining the capabilities of both manual and causal discovery tools. The framework combines the 3 approaches of Data Correlation, Causal Discovery Tools and Domain Knowledge to represent the Causal Graph. This results in a causal model that considers the ML techniques verified by the domain expertise which might be useful for decision making towards the desired target outcome.

- Performing the causal tasks follows modelling causal relationships. DoWhy can assist with the following tasks. They're Effect estimation: How much would Y change if we adjust X?
- Attribution: What caused the event to occur? How do you interpret a result? What was the anomaly produced by my variables?



.

Figure 5. *Process of Estimating the Causal Effects*

- Counterfactual estimation: What would have happened if X had been assigned a value other than the one that was observed? What would be the values of the other variables?
- Prediction: What will the value of Y be given as input with new values for various input features?

4.7 Estimating the Causal Effects

The causal effect of a variable A on y is defined as the predicted change in y because of a change in A. Sometimes we are just interested in the effect on a specific subpopulation, or we want to examine the causal influence across subpopulations. Figure 5. Is process of estimating the causal effects.

Four phases to estimate the causal effect:
1. Make assumptions to model a causal inference problem.
2. Determine the "causal estimand," or expression, for the causal effect under these suppositions.
3. Calculate the expression with statistical techniques like instrumental variables or matching variables.
4. Finally, use a range of robustness checks to confirm the estimate's correctness.
There are four main verbs in DoWhy that describe this workflow:

- model (CausalModel or graph)
- identify (identify_effect)
- estimate (estimate_effect)
- refute (refute_estimate)

DoWhy employs a causal effect estimation API that supports several techniques by using these verbs. The model employs a formal causal structure to capture previous information, implements graph-based techniques to determine the causal effect, applies statistical methods to estimate the discovered estimand, and, lastly, attempts to challenge the generated estimate by examining its adaptability to presumptions. In another way, it's essential to decide on an identification technique before thinking about an estimating procedure. The identification algorithms supported by DoWhy are Backdoor, Frontdoor, Instrumental variable, ID algorithm.Once a causal effect is identified, we can choose an estimation method compatible with the identification strategy.

Identifying the Causal Effect
The second step of causal analysis is identification, which comes after we have modelled our causal assumptions. Identification of causal effect is the process of finding out whether the effect can be approximated using the data from the available variables given a causal graph and the set of observed variables. Formally speaking, identification transforms the intended causal effect expression—for example, $[Y|do(A)]$—into a form that may be estimated without the do-operator by using the observed data distribution.

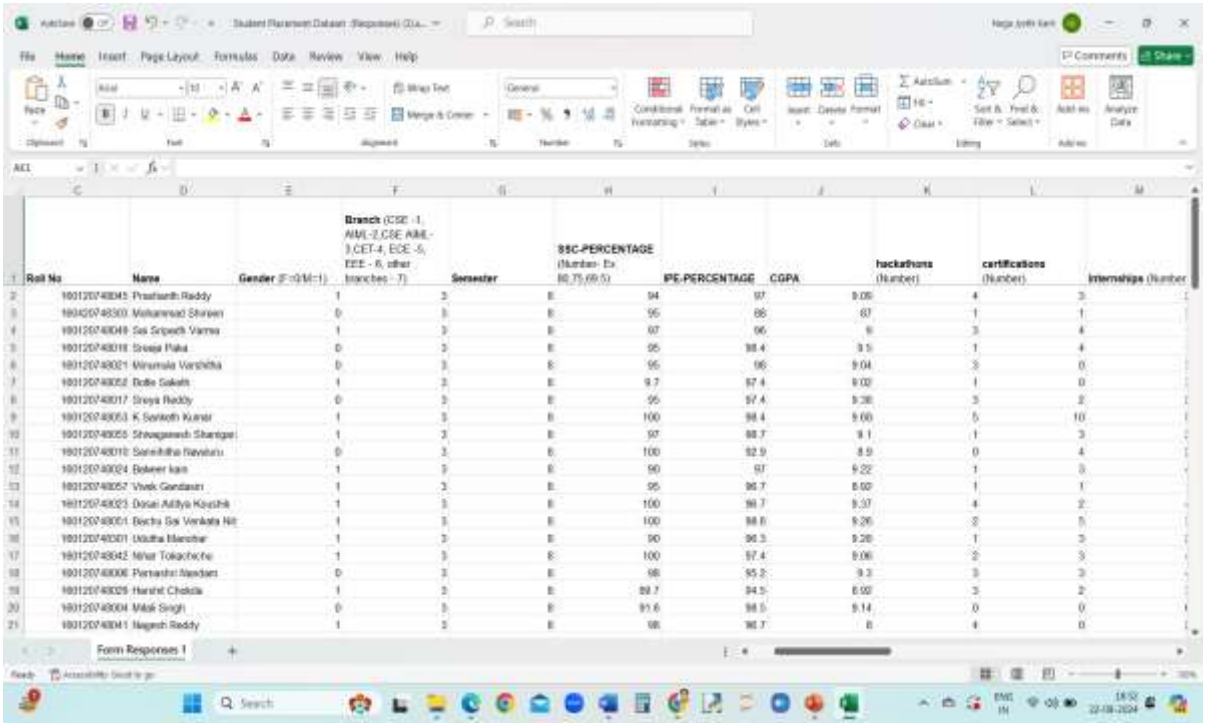
Backdoor Criterion
To identify causal effect using the backdoor criterion, any of the four basic kinds of adjustments can be used based on the requirements. Each of these are designed to return a valid backdoor set, but they vary in how they select the set of variables to return. They are maximal-adjustment, minimal-adjustment, exhaustive-search, default. A conditional probability distribution is estimated when effect estimation with backdoor is used. It is determined that the causal impact is $\sum wE[Y|A, W = w]P(W = w)$ given an action A, a result Y, and a collection of backdoor variables W. Any estimator that generates conditional expectation is suitable for our purposes. There are various types of average causal effect estimators that DoWhy supports. One widely used technique for estimating causal influence is linear regression. When there is a linear function approximation for the data-generating process leading to an output Y, it is helpful.

Quantify Causal Influence: In addition to estimating the average *total* causal effect, DoWhy can also be used for estimating the direct arrow strength between two nodes in a causal graph.

Direct Arrow Strength: This technique measures the change in distribution that occurs when an edge in a graph is eliminated to measure the causal influence of one variable on the other. When eliminating an edge, it employs a specific measure to estimate the change, such as the relative entropy or the variance difference. This yields a single, well-defined value that signifies the strength of a particular causal relationship in nonlinear interactions and explains how the removal of a particular causal link influences the target variable.

5. Results & Discussion

The present study includes the dataset used is the real time data taken from the undergraduate Engineering college students. The student details of third and final year students of Chaitanya Bharathi Institute of Technology, Hyderabad, India are included to check for placement of the student by the end of their graduation. The dataset measures 90 x 33, which indicates that there are a total of 90 student records and 33 features. The features are Roll No, Name, Gender, Branch, Semester, SSC-percentage, IPE- percentage, CGPA, Hackathons,



Roll No	Name	Gender (F=0/M=1)	Branch (CSE -1, AIML-2,CSE AIML-3,CET-4, ECE -5, EEE - 6, other branches - 7)	Semester	SSC-PERCENTAGE (Number- Ex 80.75 (69.5)	IPE-PERCENTAGE	CGPA	hackathons (Number)	certifications (Number)	internships (Number
160120748045	Prashanth Reddy	1	3	II	94	97	9.06	4	3	
160120748303	Mohammad Shireen	0	3	II	95	86	87	1	1	
160120748049	Sai Sripath Varma	1	3	II	97	96	9	3	4	
160120748018	Sowjanya Paka	0	3	II	95	98.4	9.5	1	4	
160120748021	Mirumula Vanshitha	0	3	II	95	96	9.04	3	0	
160120748058	Bolla Sakeeth	1	3	II	97	97.4	9.02	1	0	
160120748017	Sreya Reddy	0	3	II	95	97.4	9.38	3	2	
160120748053	K Sankeeth Kumar	1	3	II	100	98.4	9.66	5	10	
160120748055	Shivagowesh Shangar	1	3	II	97	88.7	9.1	1	3	
160120748010	Sanvitha Revuluru	0	3	II	100	92.9	8.9	0	4	
160120748024	Bakeer kan	1	3	II	90	97	9.22	1	3	
160120748057	Vivek Gandaveni	1	3	II	95	96.7	8.62	1	1	
160120748023	Dosa Aditya Kaushik	1	3	II	100	98.7	9.37	4	2	
160120748051	Bachu Sai Venkata Nit	1	3	II	100	98.8	9.26	2	5	
160120748301	Uddha Manohar	1	3	II	90	96.3	9.28	1	3	
160120748042	Nihar Tokacherla	1	3	II	100	97.4	9.06	2	3	
160120748006	Parmanbir Nandani	0	3	II	98	95.2	9.3	3	3	
160120748025	Harshit Chokda	1	3	II	88.7	94.5	8.62	3	2	
160120748004	Mitali Singh	0	3	II	91.6	98.5	9.14	0	0	
160120748041	Nagesh Reddy	1	3	II	98	96.7	8	4	0	

Figure 6. *Sample Data set*

Certifications, Internships, Projects, Employability skills marks, soft skills marks, self-learning capability, Job/ Higher studies, interested in games, placement training , worked in teams or not, moocs courses, mother's education, father's education, no. of siblings, accommodation type, parental status, mother's occupation, father's occupation, lab facility, class room facility, library, mentor allotted, frequency of meeting the mentor, mentoring attendance. The information is all numerical and not classified. In fields with only 0s and 1s as values, 0 denotes no and 1 denotes yes. Figure 6 is sample data set.A Python library called the DoWhy package is intended for causal modelling and inference. It offers a consistent interface for modelling causal relationships and testing causal hypotheses, integrating methods from econometrics, statistics, and machine learning. The key focus of DoWhy is to make causal inference easier and more accessible by providing tools for both the identification of causal relationships and the estimation of causal effects. The key features of DoWhy are Modeling Causal Relationships by defining a causal model using directed acyclic graphs (DAGs) or structural causal models (SCMs). DoWhy is a powerful tool for anyone interested in rigorous causal analysis, whether in academia, industry, or policy- making. The following packages need to be installed for modelling the causal relationships.

5.1 Install the dowhy packages

```
!pip install dowhy econml
!apt install libgraphviz-dev
!pip install pygraphviz
```

5.2 Data Load and Data Preparation

The outputs after the data loading and preparation are shown in figure 7 and 8.

5.3 Build the Causal Graphic Model (3D Framework)

The correlation heatmap analysis highlights key factors influencing student placement outcomes. CGPA positively correlates with placement success (0.47), but employability (0.60) and soft skills (0.54) play an even stronger role. Practical experiences like internships (0.41) and projects (0.47) significantly enhance employability, while teamwork (0.24) and extracurricular activities (0.24) also contribute. Academic resource utilization (library and lab grades) improves CGPA but has a weaker direct impact on placement. Field of study (0.52) is a major determinant, while gender has negligible influence (0.06). Accommodation type negatively affects placement (-0.26), possibly due to environmental constraints. Overall, a balanced approach integrating academics, soft skills, and experiential learning is crucial for maximizing employability. Figure 9 shows data correlation. The correlation heatmap (Figure 10) presents the relationship between the PlacedOrNot variable and other academic, extracurricular, and skill-related factors, providing insights into the key determinants

\nInternships (Number)	\nProjects (Number)	\n% in Employability Skills course (Number)	...	\ninterested in games/extracurricular activities (yes-1, No -0)	\nplacement training taken(y=1/n=0)	\nworked in teams ever? (Yes-1, No-0)	\nMoocs courses done? (Number)	\nAccommodation type (Hostel -1, Home -2)	\nLab Facility (good - 1/moderate - 0/ not good - 2)	\nclass room facility (good - 1/moderate - 0/ not good - 2)	Library/ E-Resources Availability (good - 1/moderate - 0/ not good - 2)	\nmentor allotted (1:y/0: n)	Placement Status(Placed-1/Not Placed -0) - final years fill it up, others need not fill this option
2	3	80	...	1	1	1	2	2	1	0	1	1	1
3	2	0	...	1	1	1	0	1	1	1	1	1	1
1	4	90	...	1	0	1	2	2	0	0	1	1	1
1	3	95	...	1	1	1	1	1	0	0	1	1	1
3	5	0	...	1	1	1	0	2	2	0	0	1	1

Figure 7. *After data preparation*

```
Index(['Gender', 'Branch', 'SSC_Grade', 'IPE_Grade', 'CGPA', 'Hackathons',
       'Certifications', 'Internships', 'Projects', 'Employability_Skills',
       'Soft_Skills', 'Extracurricular', 'Placement_Trainings',
       'Worked_In_Teams', 'Moocs_Courses', 'Accommodation_Type', 'Lab_Grade',
       'Class_Room_Grade', 'Library_Grade', 'Mentor_Alloted', 'PlacedOrNot'],
      dtype='object')
```

	Gender	Branch	SSC_Grade	IPE_Grade	CGPA	Hackathons	Certifications	Internships	Projects	Employability_Skills	...	Extracurricular	Placement_Trainings	Worked_In_
0	1	3	94.0	97.0	9.09	4	3	2	3	90	...	1	1	
1	0	3	95.0	88.0	8.70	1	1	3	2	0	...	1	1	
2	1	3	97.0	96.0	9.00	3	4	1	4	90	...	1	0	
3	0	3	95.0	98.4	9.50	1	4	1	3	95	...	1	1	
4	0	3	95.0	98.0	9.04	3	0	3	5	0	...	1	1	

5 rows × 21 columns

Figure 8. *After data preparation*

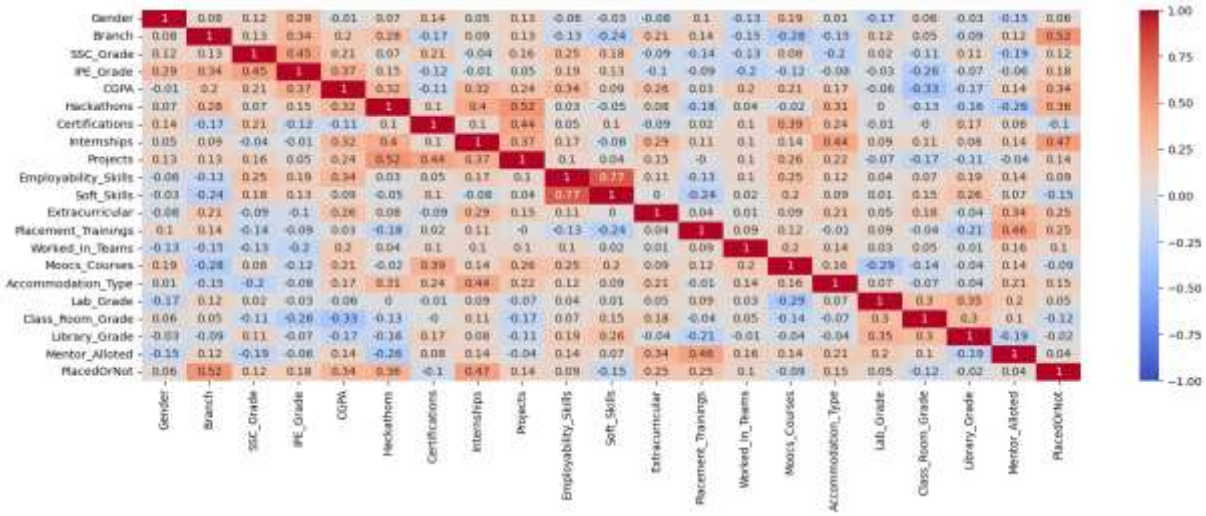


Figure 9. *Data Correlation*

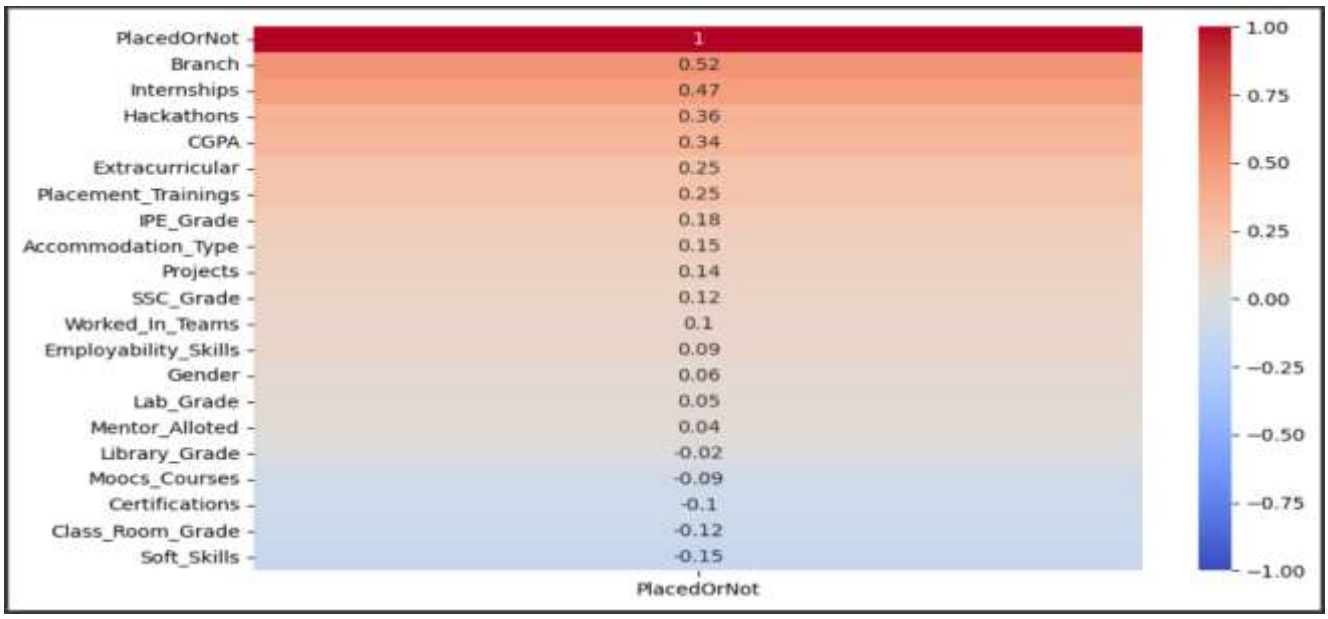


Figure 10. *Correlation between the placed Or Not variable and the other variables*

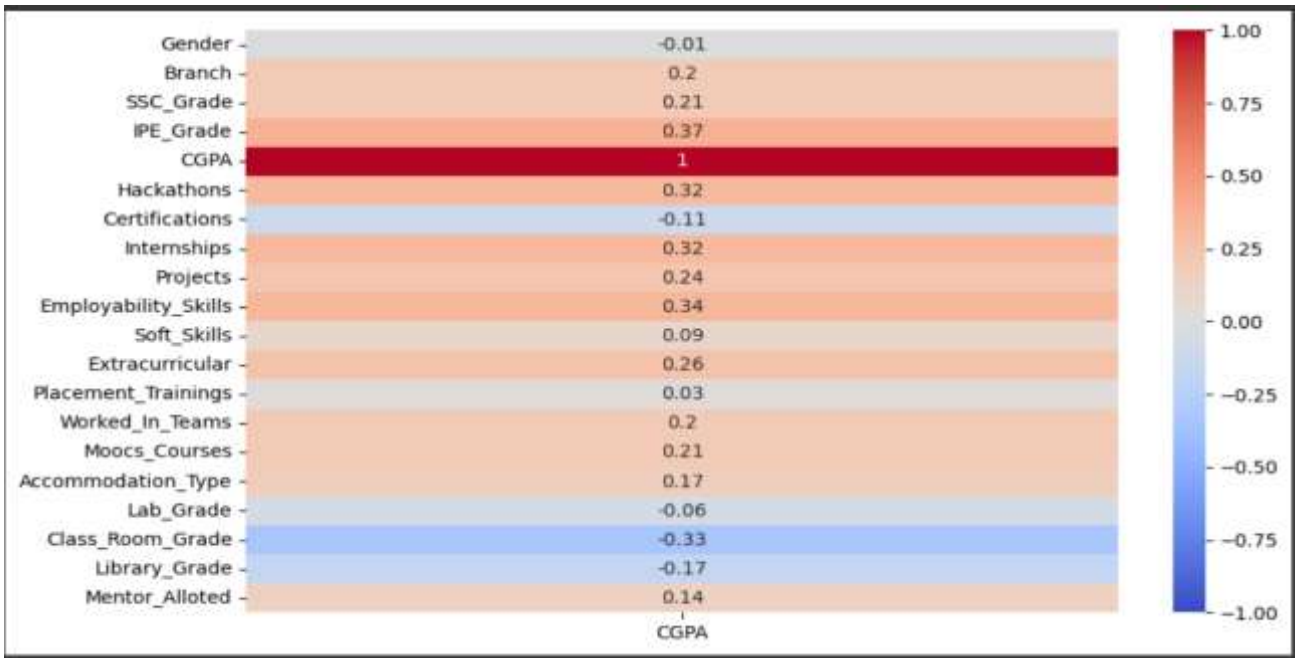


Figure 11. *Correlation between the CGPA variable and the other variables*

of student placement success. Branch of study exhibits the highest positive correlation (0.52), indicating that field of study significantly influences placement chances, likely due to industry demand variations. Internships (0.47) and Hackathons (0.36) show strong positive associations, emphasizing the importance of hands-on experience and industry exposure in securing employment. CGPA (0.34) also positively impacts placement, reinforcing the relevance of academic performance, though it is comparatively less influential than experiential learning. Participation in extracurricular activities (0.25) and placement training programs (0.25) further contributes positively. Moderate correlations are observed for IPE Grade (0.18) and Accommodation Type (0.15), suggesting a minor influence of past academic performance and living environment on placement success. Surprisingly, Soft Skills (-0.15) and Class Room Grade (-0.12) exhibit negative correlations, implying that traditional academic excellence and self-reported soft skills alone may not guarantee placement. Certifications (-0.10) and MOOCs (-0.09) also display weak negative correlations, indicating that these credentials alone might not significantly impact placement outcomes. Other variables such as Employability Skills (0.09) and Worked in Teams (0.12) show only weak positive correlations, suggesting that while these skills are essential, they may not directly influence hiring decisions as

strongly as hands-on project experience. Overall, the analysis underscores the importance of a balanced skillset combining academic performance, practical exposure, and industry-relevant experience in enhancing placement opportunities. Figure 11 illustrates the correlation between CGPA and various academic, extracurricular, and professional development factors, providing insights into the determinants of academic performance. IPE Grade (0.37) and SSC Grade (0.21) exhibit positive correlations with CGPA, indicating that students with strong pre-university academic backgrounds tend to maintain high performance in higher education. Employability Skills (0.34), Internships (0.32), and Hackathons (0.32) also show a positive correlation, suggesting that students who engage in skill-building activities and industry exposure tend to perform well academically. Projects (0.24) and Extracurricular Activities (0.26) further contribute positively, reinforcing the role of practical experience and holistic development in academic success. Worked in Teams (0.20) and Moocs Courses (0.21) demonstrate a mild positive correlation, highlighting the benefit of collaborative and self-paced learning. Conversely, Class Room Grade (-0.33) exhibits the strongest negative correlation, implying that students with high CGPA may not necessarily perform well in classroom assessments, possibly due to differences in assessment methods. Library Grade (-0.17) and Certifications (-0.11) also show weak negative correlations, suggesting that traditional academic resource utilization and certifications alone may not directly translate to higher CGPA. Lab Grade (-0.06) has a negligible correlation, indicating that laboratory performance does not significantly impact overall academic scores. Accommodation Type (0.17) and Mentor Allotted (0.14) have a small positive correlation, suggesting a minor influence of living conditions and mentorship on academic success. Overall, the findings suggest that while prior academic performance, employability skills, and industry engagement positively contribute to CGPA, traditional academic indicators such as classroom grades and library usage show a weaker or even negative correlation. This highlights the importance of a well-rounded educational approach integrating practical exposure and skill development alongside theoretical learning for academic excellence. Figure 12 presents the correlation between Hackathons and various academic, extracurricular, and professional factors. Projects (0.52) and Branch (0.28) show strong positive correlations, suggesting that students involved in hackathons often engage in project-based learning and belong to technical fields. A mild positive correlation with IPE Grade (0.15) and Certifications (0.10) indicates that prior academic performance and additional credentials contribute to hackathon participation. Accommodation Type (0.31) suggests that living conditions might influence participation rates. However, Placement Trainings (-0.18) and Mentor Allotted (-0.26) show negative correlations, implying that students relying on structured training or mentorship may engage less in hackathons. Weak or negative correlations with Employability Skills (0.03) and Soft Skills (-0.05) suggest that hackathons may primarily enhance technical rather

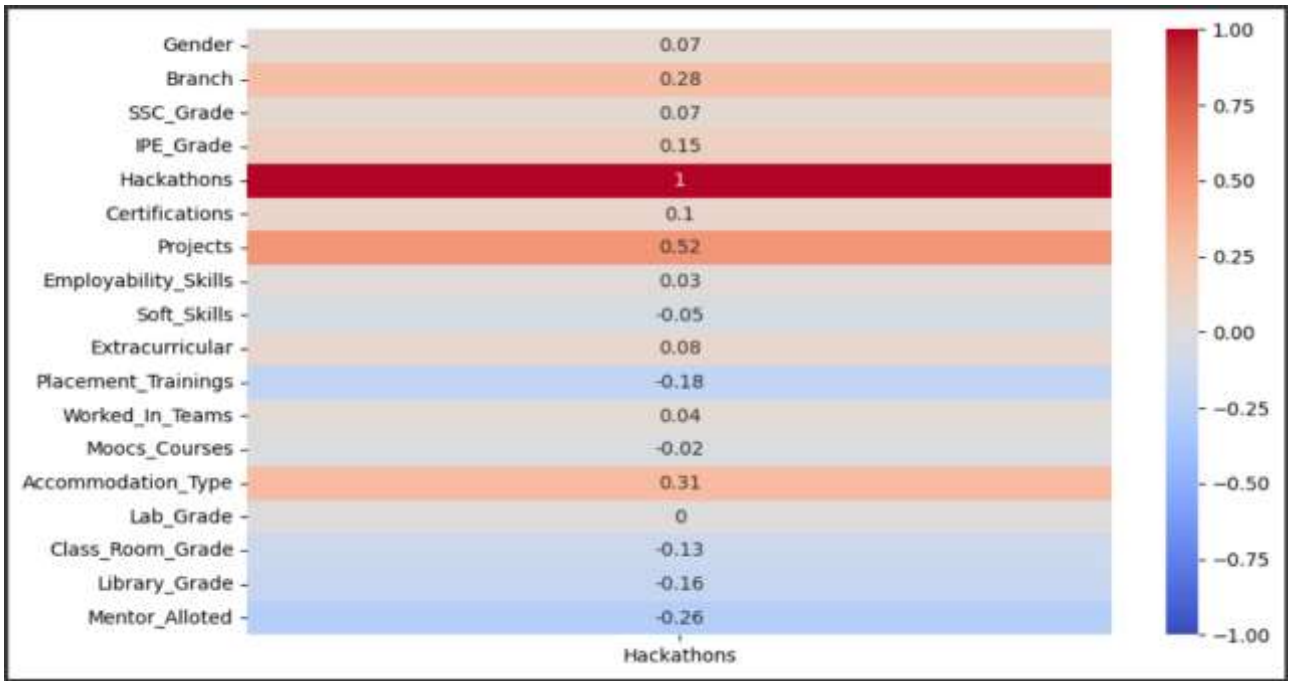


Figure 12. *Correlation between the Hackathons variable and the other variables*

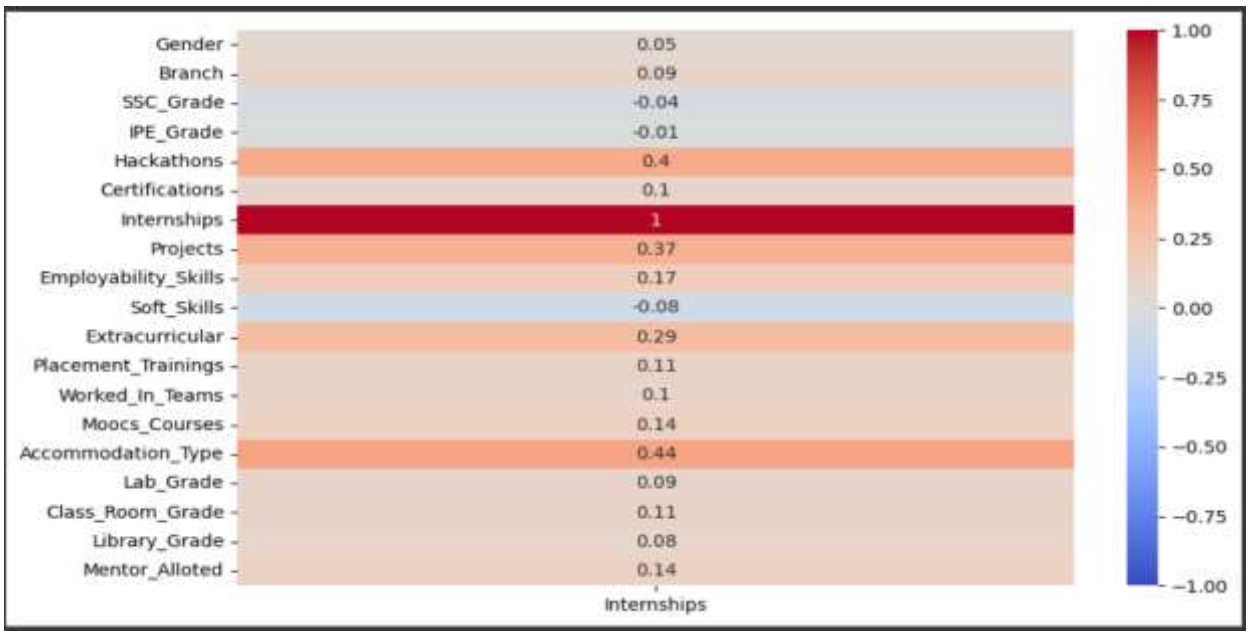


Figure 13. *Correlation between the Internships variable and the other variables*

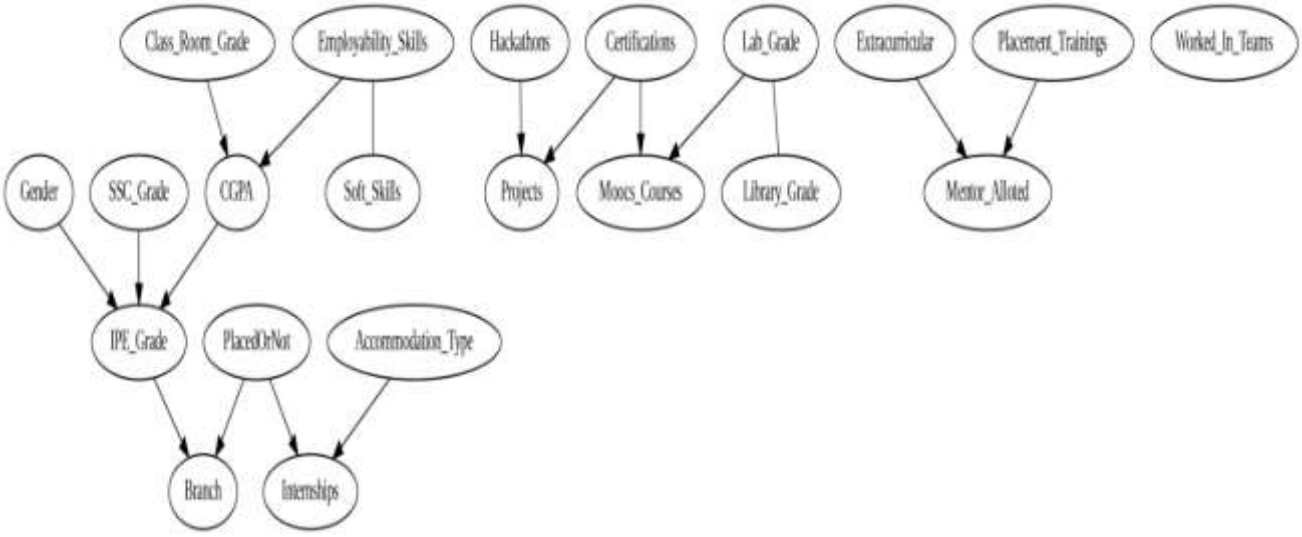


Figure 14. *CP-DAG using the PC Technique*

than interpersonal skills. Overall, hackathon participation is strongly associated with hands-on project experience and technical discipline but less linked to structured guidance or traditional employability training Figure 13 highlights the correlation between Internships and various academic, extracurricular, and employability factors. Strong positive correlations with Projects (0.37), Hackathons (0.40), and Extracurricular Activities (0.29) suggest that students engaged in internships often participate in hands-on learning and non-academic activities. Employability Skills (0.17) and Placement Training (0.11) show weaker correlations, indicating that internships contribute to skill development but are not the sole factor. Accommodation Type (0.44) suggests accessibility influences internship opportunities. Minor positive associations with Certifications (0.10) and Moocs Courses (0.14) indicate additional learning aids internships. A slight negative correlation with Soft Skills (-0.08) implies internships focus more on technical abilities. Overall, internships are strongly linked to practical experience, technical exposure, and accessibility rather than early academic performance.

5.4 Causal Discovery Algorithms

PC – Peter Clark Algorithm

Figure 14 presents the Causal Partially Directed Acyclic Graph (CP-DAG) generated using the PC (Peter-Clark) technique, illustrating causal relationships among academic, extracurricular, and

employability-related factors. The directed edges in the graph indicate inferred causal influences between variables. Academic performance is influenced by SSC Grade and IPE Grade, which directly impact CGPA, and in turn, CGPA affects Employability Skills and Soft Skills. PlacedOrNot is primarily determined by Branch and Internships, both of which are causally linked to IPE Grade, indicating that prior academic performance influences field selection and internship opportunities. Accommodation Type also impacts Internships, suggesting that living conditions play a role in practical exposure. Hands-on learning experiences, such as Projects, Hackathons, and.,

GES – Greedy Equivalence Search

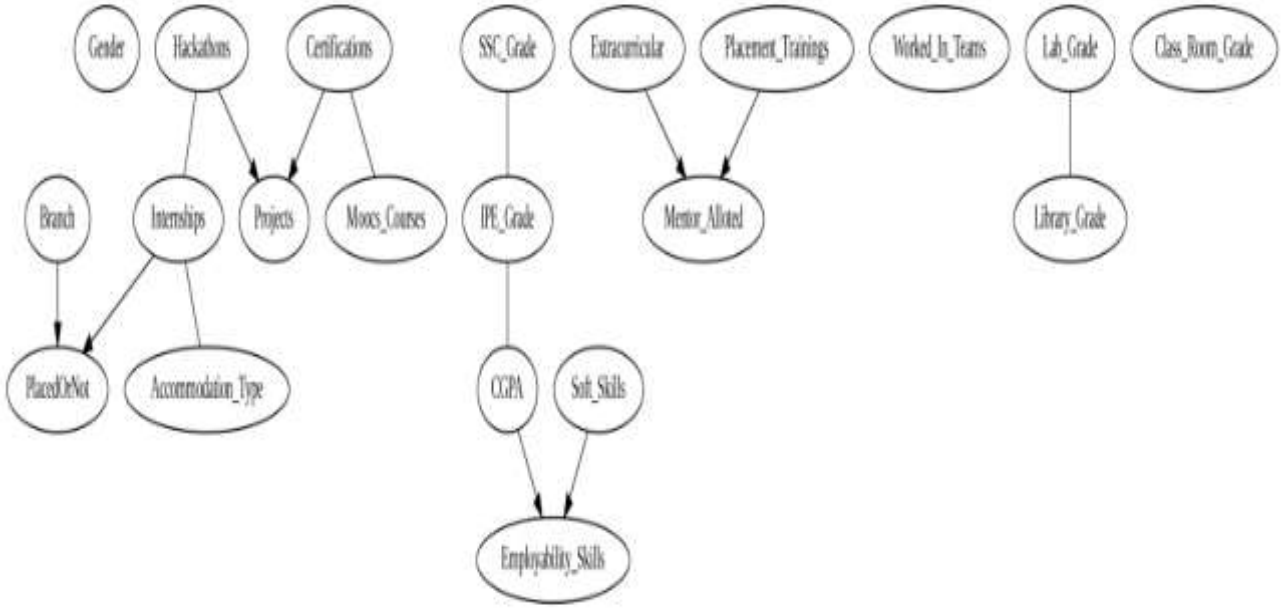


Figure 15. *CP-DAG using the GES Technique*

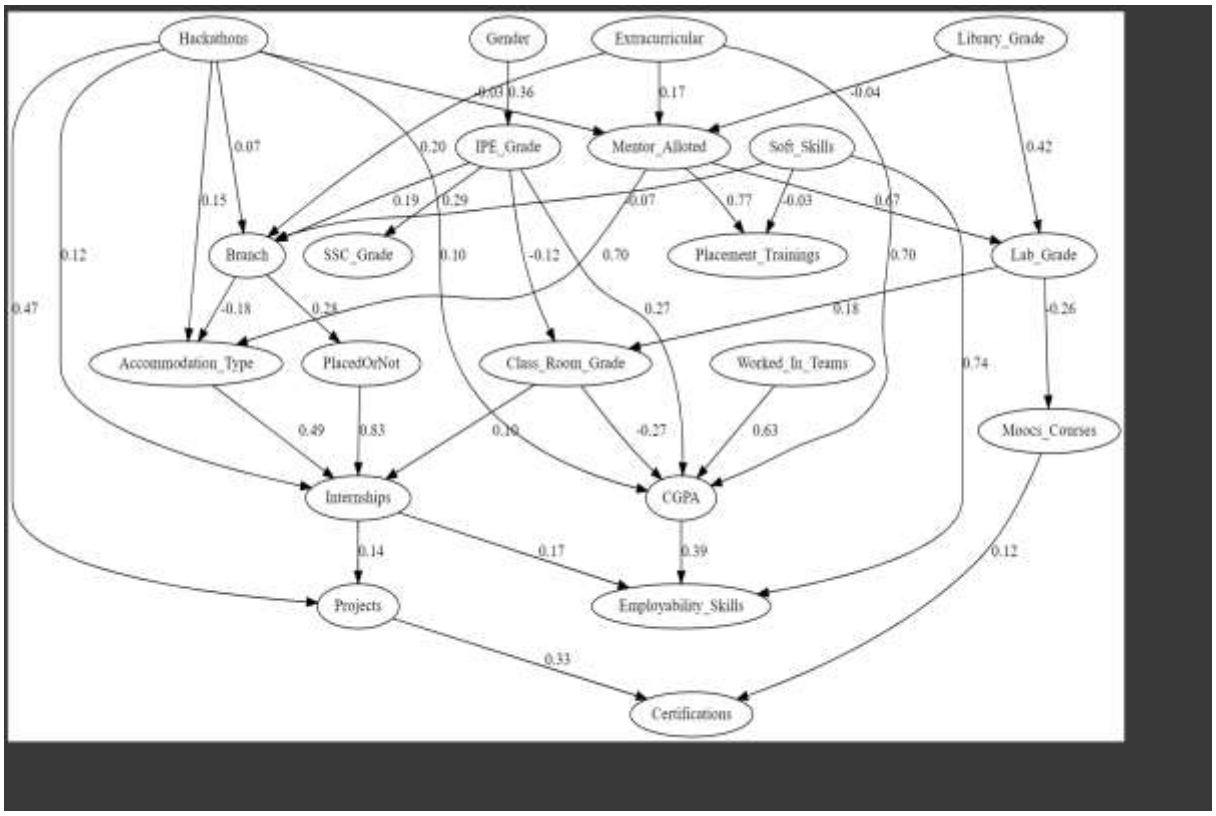


Figure 16. *DAG using the LiNGAM Technique*

Certifications, are interconnected, with Hackathons driving project engagement and Certifications leading to MOOCs participation. Library Grade is influenced by Lab Grade, indicating academic resource utilization patterns. Extracurricular activities drive Mentor Allotment and Placement Training, showing the importance of non-academic engagement in structured career support Interestingly, Worked in Teams appears isolated, implying it may not be directly influenced by other variables but could still play an indirect role in employability. Overall, this CP-DAG highlights that internships, field of study, and prior academic performance are key determinants of placement success, while skill development and practical exposure serve as crucial mediators.The model reinforces the need for a balanced academic and experiential learning approach to enhance employability outcomes. Figure 15 presents the Causal Partially Directed Acyclic Graph (CP-DAG) constructed using the Greedy Equivalence Search (GES) technique, which identifies causal structures among academic extracurricular, and employability-related factors. The directed edges indicate causal influences between variables, providing insights into the key determinants of placement success. Branch of study is a primary determinant of Placement Outcome (PlacedOrNot), reinforcing the importance of field selection in employability. Internships and Accommodation Type directly influence placement, indicating that both practical experience and environmental factors impact job opportunities. Hackathons contribute to Internships, while Certifications facilitate Project involvement, suggesting that technical skill-building fosters experiential learning. Moocs Courses emerge as a result of Certifications, highlighting self-driven learning pathways. Academic performance is shaped by SSC Grade and IPE Grade, both feeding into CGPA, which in turn drives Employability Skills and Soft Skills. This suggests that strong academic foundations support professional readiness. Extracurricular Activities impact Mentor Allotment and Placement Training, indicating that students engaged in non-academic pursuits receive more structured career guidance. Worked in Teams appears linked to Placement Training, emphasizing the role of collaboration in professional development. Interestingly, Library Grade is influenced by Lab Grade, indicating a relationship between practical coursework and academic resource utilization. However, Classroom Grade remains isolated, implying that traditional academic performance may not be a significant causal factor in employability. Overall, this CP-DAG highlights that field selection, internships, and experiential learning are crucial for placement success, while structured training, mentorship, and employability skills serve as key mediators. The model underscores the necessity of integrating academic excellence with hands-on learning and industry engagement to enhance career prospects.

LiNGAM – Linear Non-Gaussian Acyclic Model

Figure 16 presents a Directed Acyclic Graph (DAG) generated using the LiNGAM (Linear Non-Gaussian Acyclic Model) technique, illustrating direct causal influences among academic, extracurricular, and employability-related variables. PlacedOrNot is heavily influenced by Internships (0.83) and Branch (0.28), highlighting the significance of practical experience and field of study in securing placements. Internships are strongly linked to Projects (0.14) and influenced by Accommodation Type (0.49), indicating that living conditions impact internship opportunities. CGPA is a key determinant of Employability Skills (0.39), which is further enhanced by Certifications (0.33). IPE Grade (0.20) and SSC Grade (0.19) contribute to CGPA, reinforcing the role of academic history in performance. Soft Skills (0.03) and Worked in Teams (0.18) positively impact CGPA, while Placement Training (0.70) and Mentor Allotment (0.70) play a significant role in student development. Lab Grade (0.70) influences Moocs Courses (0.74) and Library Grade (0.42), suggesting that academic resources contribute to additional learning. Hackathons (0.12) influence Branch and Projects, indicating the importance of competitive learning. Overall, the DAG highlights that internships, academic performance, and practical experiences serve as key mediators for employability, emphasizing a holistic approach to career readiness.

5.4 Domain Knowledge

In the context of analyzing employability factors, a structured approach is adopted to identify key causal relationships using data correlation, causal discovery tools, and a domain knowledge-driven framework. The analysis incorporates treatments, instrument variables, and confounders to ensure a robust causal inference model.

D#1: Data Correlation Approach

The initial phase involves identifying treatments, instrument variables, and confounders based on data correlation patterns. Here, Branch is considered the primary treatment variable, given its substantial impact on placement outcomes. Instrument variables include Internships, Hackathons, and CGPA, which influence the treatment but are assumed to be exogenous in the causal model. The confounders,

which may introduce bias in estimating causal effects, include Projects, Accommodation Type, Extracurricular Activities, Placement Trainings, and IPE Grade. These factors impact both the treatment and the outcome, necessitating their inclusion for unbiased estimation.

D#2: Causal Discovery Toolset

To further refine causal relationships, three causal discovery algorithms—PC (Peter-Clark), GES (Greedy Equivalence Search), and LiNGAM (Linear Non-Gaussian Acyclic Model)—are employed.

- PC Algorithm: The identified treatments are Branch and Internships, while instrument variables include Accommodation Type, IPE Grade, CGPA, SSC Grade, Gender, Class Room Grade, and Employability Skills. Notably, no confounders are identified in this model.
- GES Algorithm: Similar to the PC approach, the treatments remain Branch and Internships, whereas instrument variables are Accommodation Type, Hackathons, Projects, Certifications, and MOOCs Courses, with no identified confounders.
- LiNGAM Algorithm: This approach also considers Branch and Internships as treatment variables, while instrument variables expand to Hackathons, IPE Grade, Accommodation Type, Projects, Employability Skills, CGPA, Soft Skills, and MOOCs Courses, again with no identified confounders.

D#3: Domain Knowledge-Based Approach (MCMD - Minimum Criteria and Maximum Differentiators)

A domain knowledge-driven approach, MCMD (Minimum Criteria and Maximum Differentiators), is employed to refine the causal model by ensuring the most explanatory yet distinct features are selected. In this approach:

- Treatments: Branch and CGPA are identified as the primary variables affecting employability.
- Instrument Variables: Class Room Grade and Certifications are included, as they contribute to skill-building but remain exogenous to the placement decision-making process.
- Confounders: Internships, MOOCs Courses, Extracurricular Activities, Employability Skills, Projects, and Hackathons are identified as they impact both the treatment variables and the final employment outcomes, necessitating their control in causal estimation.

Final Identification Using the 3D Framework

The final step involves integrating insights from the above methods within a 3D Framework, which systematically identifies causal structures based on data-driven and domain-specific criteria. The final classification yields:

- Treatments: Branch
- Instrument Variables: CGPA
- Confounders: Internships, Hackathons, Projects, Certifications, Extracurricular Activities, and MOOCs Courses

.

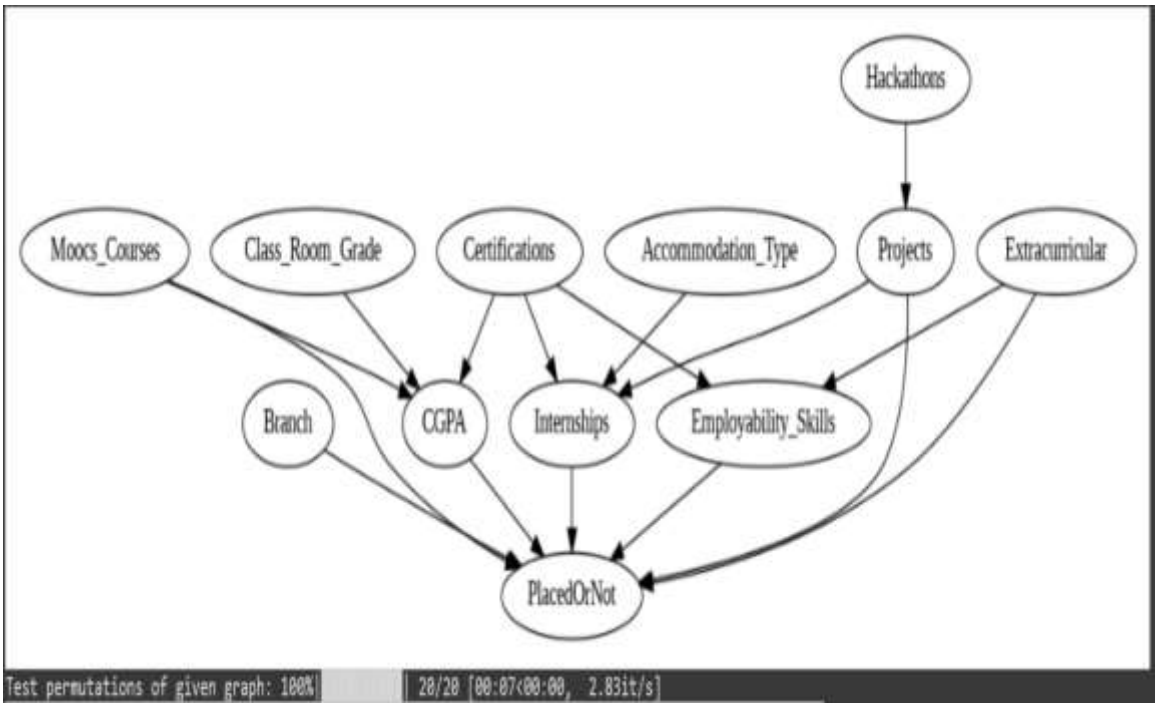


Figure 17. *Classifying the identified causal graph to ensure its validity before final confirmation*

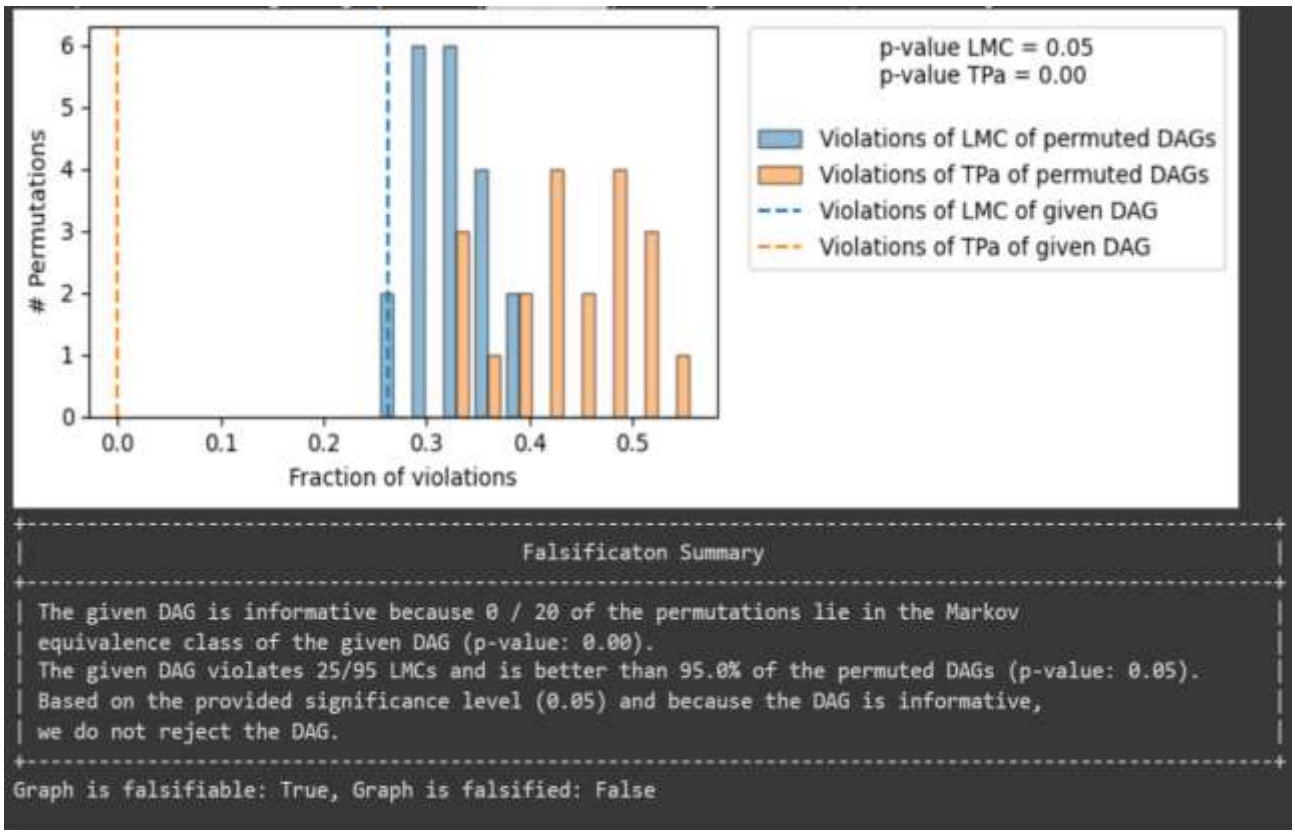


***Figure 18**. Graph Falsification*

This structured causal analysis framework ensures a robust and unbiased understanding of the factors influencing employability outcomes. The integration of data correlation, causal discovery algorithms, and domain knowledge-based refinement provides a comprehensive and empirically validated model for assessing employability determinants

5.5 Falsifying the identified graph before finally confirmation

Figure 17 presents the process of falsifying the identified causal graph to ensure its validity before final confirmation. The model undergoes permutation testing (as indicated by the 100% completion status), verifying whether the derived causal relationships hold under multiple iterations. Key determinants of Placement Outcome (PlacedOrNot) include Internships, CGPA, Employability Skills, and Branch, supported by instrument variables such as Certifications, Class Room Grade, and MOOCs Courses. Practical experiences like Projects and Hackathons influence Employability Skills and Internships, while Extracurricular Activities provide additional indirect contributions. Accommodation Type also plays a role in accessibility to career opportunities. The iterative falsification process ensures that the final graph structure is empirically robust, resistant to spurious correlations, and accurately reflects causal influences on employability outcomes. Figure 18 presents the falsification analysis of the Directed Acyclic Graph (DAG) to assess its validity in capturing causal relationships. The falsification summary confirms that the graph is falsifiable but not falsified, meaning that while it can be tested against alternative structures, it remains statistically valid. The Markov equivalence test shows that none of the 20 permutations fall within the equivalence class of the given DAG (p-value: 0.00), confirming its informativeness. Additionally, the DAG violates 25 out of 95 Local Markov Constraints (LMCs) but remains statistically superior to 95% of the permuted DAGs (p-value: 0.05), indicating a well-fitted model with minimal violations. The histogram visually compares the fraction of violations across permutations, where the given DAG exhibits fewer violations than most alternatives. Given that the p-value remains within the acceptable significance level (0.05) and the graph is informative, the final decision is not to reject the DAG, reinforcing its reliability in representing causal influences on employability factors.

5.6 Determining the feature relevance

Fitting causal mechanism of node Extracurricular: 100% 12/12 [00:00<00:00, 84.74it/s] Estimating Shapley Values. Average change of Shapley values in run 24 (120 evaluated permutations):

1.5642433697150502%: 100% 1/1 [10:20<00:00, 620.11s/it]

{('Branch', 'PlacedOrNot'): 0.09581752453266716,
('CGPA', 'PlacedOrNot'): 0.017797705580664305,
('Employability_Skills', 'PlacedOrNot'): 0.004209564247333569,
('Extracurricular', 'PlacedOrNot'): 0.0003637590066717129,
('Internships', 'PlacedOrNot'): 0.11635461991999982,
('Moocs_Courses', 'PlacedOrNot'): 0.002382087666668813,
('Projects', 'PlacedOrNot'): 0.007791442187337973} [0.16087654]

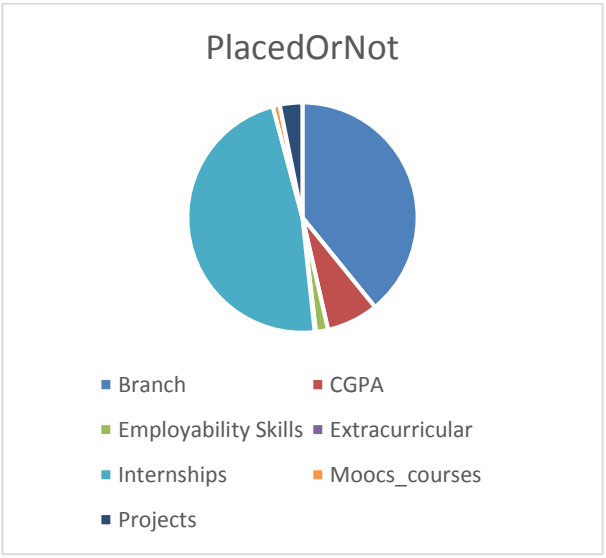


Figure 19. *Feature relevance with the outcome attribute*

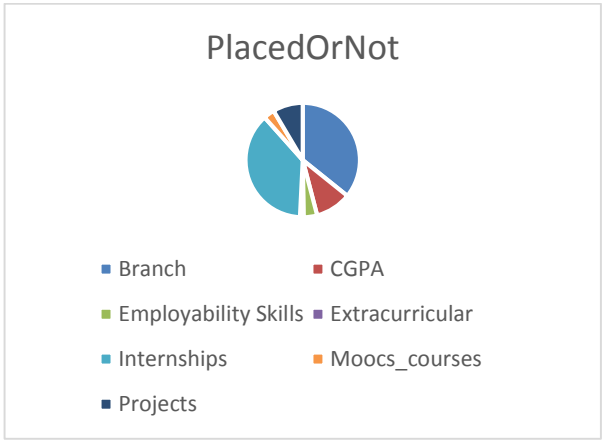


Figure 20. *Arrow strength of the other attributes with the outcome attribute*

5.7 Quantifying the Arrow Strength

Fitting causal mechanism of node Extracurricular: 100%|[redacted]| 12/12 [00:00<00:00, 91.66it/s]

{('Branch', 'PlacedOrNot'): 0.14766772177419357,
('CGPA', 'PlacedOrNot'): 0.042035757009345816,
('Employability_Skills', 'PlacedOrNot'): 0.015804014150943405,
('Extracurricular', 'PlacedOrNot'): 0.00416719349315069,
('Internships', 'PlacedOrNot'): 0.15526905468750002,
('Moocs_Courses', 'PlacedOrNot'): 0.012883459374999994,
('Projects', 'PlacedOrNot'): 0.035101821808510646

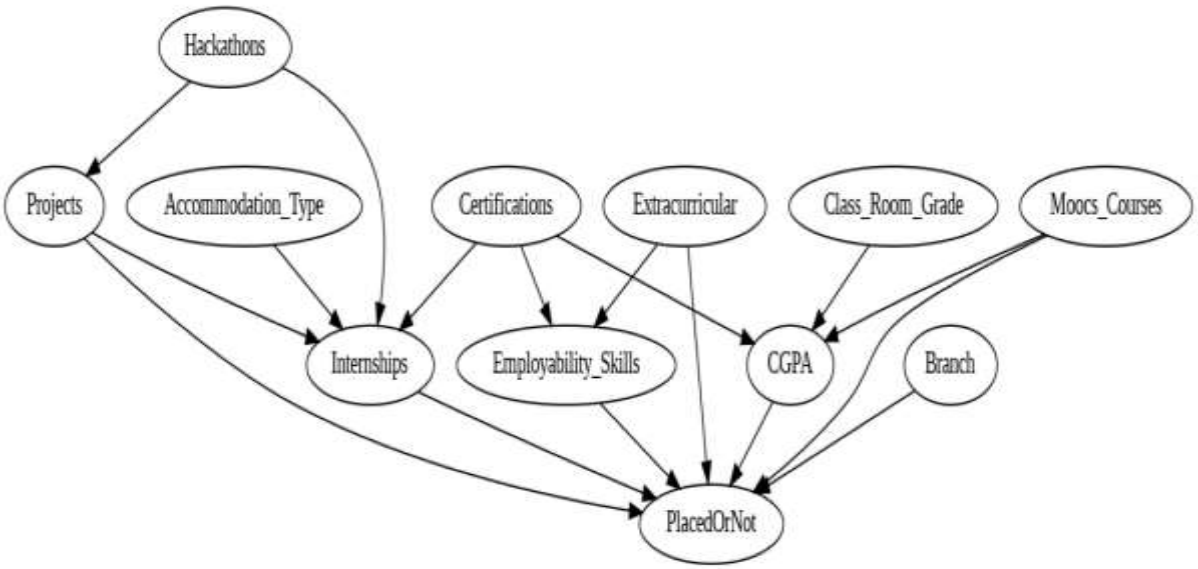


Figure 21*. Causal Model*

5.8 Creating the Causal Model

The process of determining feature relevance involves fitting the causal mechanism for the node Extracurricular, achieving a 100% completion rate with an average change in Shapley values of 1.56% over 120 evaluated permutations. The results indicate that Internships (0.1164) and Branch (0.0958) have the highest impact on Placement Outcomes (PlacedOrNot), followed by CGPA (0.0178) and Projects (0.0078), while Extracurricular Activities (0.00036) and MOOCs Courses (0.0024) contribute minimally. The quantification of arrow strength further refines these relationships, with Internships (0.1553) and Branch (0.1477) remaining the most influential, while CGPA (0.0420) and Projects (0.0351) show moderate effects, and Extracurricular (0.0042) and MOOCs Courses (0.0129) exert the weakest influences. These insights are visualized in Figures 19 and 20, illustrating the hierarchical impact of various attributes on placement outcomes. Finally, the causal model is constructed (Figure 21), synthesizing these results into a structured framework that captures the direct and indirect causal relationships influencing student employability.

6. Conclusions & Future Work

The proposed study is to use techniques of causal inference in observational data to determine the relationships between placement features and estimated the causal influence of a feature X on a later feature Y using the robustness checks like matching and regression. We applied this methodology to the 2020–2021 dataset that we acquired from the Chaitanya Bharathi Institute of Technology's Computer Science and Allied departments. This work forms the final validated Causal Graph (DAG) by using the different Causal Discovery Algorithms and analysing the resulting causal graphs using the Minimum Criteria and Maximum Differentiator (MCMD) notion of Domain Knowledge. The robust Causal Model is constructed, and the treatments, results, and confounders are identified. For this placement use case, the influence of treatments on the results has been estimated. To quantify the estimated treatment effect on the result, it is based on the mean value and the p-value (significance Level) computations. For strong outcomes, the effect estimate is also refuted. In the end, as decisions are made based on the expected outcomes, a robust decision-making model is developed. The study presents the findings from the several phases of modelling, identification, estimation, and refutation on the causal variables in the educational data set. To improve analysis and validation and help with decision-making, future study may consider data from other disciplines. Machine learning has been used and applied in different fields [48-57].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Jyothi, D. N., & Dulhare, U. N. (2023). Inferring causal relationships in student placements' performance using causal machine learning. *International Journal of Intelligent Systems and Applications in Engineering*, 11(2), 123-130.

[2] Kaur, P., Polyzou, A., & Karypis, G. (2019). *Causal inference in higher education: Building better curriculums*. arXiv preprint arXiv:1906.04698.

[3] Akshaya Kumar Mandal, Pedro Machado, & Eneko Osaba. (2025). Applying Coral Reef Restoration Algorithm for Quantum Computing in Genomic Data Analysis. *International Journal of Computer Engineering in Research Trends*, 12(1), 20–28

[4] Qin Yang, Claudio Castelnovo, & Florian Mölein. (2025). Utilizing Leaf Venation Network Model for Ethical AI Decision-Making in Financial Technologies. *International Journal of Computer Engineering in Research Trends*, 12(1), 39–48.

[5] Naresh Kumar Bhagavatham, Bandi Rambabu, Jaibir Singh, Dileep P, T. Aditya Sai Srinivas, M. Bhavsingh, & P. Hussain Basha. (2024). Autonomic Resilience in Cybersecurity: Designing the Self-Healing Network Protocol for Next-Generation Software-Defined Networking. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.640

[6] Akshaya Kumar Mandal, Pedro Machado, & Eneko Osaba. (2025). Applying Coral Reef Restoration Algorithm for Quantum Computing in Genomic

Data Analysis. *International Journal of Computer Engineering in Research Trends,* 12(1), 20–28.

[7] SumanPrakash, P., Ramana, K. S., CosmePecho, R. D., Janardhan, M., Churampi Arellano, M. T., Mahalakshmi, J., Bhavsingh, M., & Samunnisa, K. (2024). Learning-driven continuous diagnostics and mitigation program for secure edge management through zero-trust architecture. *Computer Communications,* 94–107. https://doi.org/10.1016/j.comcom.2024.04.007

[8] Krishna, U. V., Rao, G. S., Addepalli, L., Bhavsingh, M., S. D., V. S., & Jaime, L. M. (2024). Enhancing airway assessment with a secure hybrid network-blockchain system for CT & CBCT image evaluation. *International Research Journal of Multidisciplinary Technovation,* 6(2), 45–60. https://doi.org/10.54392/irjmt2425

[9] Onesimus John Waino, & Steven David. (2025). An Artificial Intelligence Model for Predicting Flooding and Drought in Bali Local Government Area of Taraba State, Nigeria. *International Journal of Computer Engineering in Research Trends*, 12(1), 1–19.

[10] Divyansh Awasthi, Zeinab Elngar, & Jeyarani Selvarajan. (2025). Implementing Bioluminescent Swarm Optimization to Enhance Blockchain Security in IoT Healthcare Systems. *International Journal of Computer Engineering in Research Trends,* 12(1), 29–38.

[11] Saranya, V. S., Subbarao, G., Balakotaiah, D., & Bhavsingh, M. (2024). Real-time traffic flow optimization using adaptive IoT and data analytics: A novel DeepStreamNet model. *International Journal of Advanced Research in Computer Science*, 15(10), 45–52

[12] Prakash, P. S., Janardhan, M., Sreenivasulu, K., Saheb, S. I., Neeha, S., & Bhavsingh, M. (2022). Mixed Linear Programming for Charging Vehicle Scheduling in Large-Scale Rechargeable WSNs. *Journal of Sensors,* 2022, 1–13. https://doi.org/10.1155/2022/8373343

[13] Lakshmi, M. S., Ramana, K. S., Pasha, M. J., Lakshmi, K., Parashuram, N., & Bhavsingh, M. (2022). Minimizing the Localization Error in Wireless Sensor Networks Using Multi-Objective Optimization Techniques. *International Journal on Recent and Innovation Trends in Computing and Communication,* 10(2s), 306–312. https://doi.org/10.17762/ijritcc.v10i2s.5948

[14] Qin Yang, Claudio Castelnovo, and Florian Mölein, (2025). Utilizing Leaf Venation Network Model for Ethical AI Decision-Making in Financial Technologies, *Int. J. Comput. Eng. Res. Trends*, 12(1);39–48.

[15] Burnett, J. W., & Blackwell, C. (2023). Graphical causal modelling: An application to identify and estimate cause-and-effect relationships. *Applied Economics,* 55(1), 1–15.

[16] Kaur, P., Polyzou, A., & Karypis, G. (2019). Causal inference in higher education: Building better curriculums. In *Proceedings of the 6th ACM Conference on Learning at Scale* (pp. 1–4).

[17] Ouaadi, I., & Ibourk, A. (2023). Causal discovery and features importance analysis: What can be inferred about at-risk students? In Proceedings of the *International Conference on Business Intelligence (pp. 134–145)*. Springer.

[18] Sharma, A., & Kiciman, E. (2020). DoWhy: An end-to-end library for causal inference. *arXiv Preprint, arXiv:2011.04216.*

[19] Shohei, S., Shimizu, S., & Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research,* 12, 1225–1248.

[20] Aragam, B. (2024). Greedy equivalence search for nonparametric graphical models. *arXiv Preprint*, arXiv:2406.17228.

[21] Forney, A., & Mueller, S. (2022). Causal inference in AI education: A primer. *Journal of Causal Inference*, 10(1), 141–173.

[22] Tadayon, M., & Pottie, G. (2021). Causal inference in educational systems: A graphical modeling approach. *arXiv Preprint,* arXiv:2108.00654.

[23] de Carvalho, W. F., & Zarate, L. E. (2019). Causality relationship among attributes applied in an educational data set. In Proceedings of the *34th ACM/SIGAPP Symposium on Applied Computing* (pp. 1271–1277).

[24] Sharma, A., & Kiciman, E. (2020). DoWhy: An end-to-end library for causal inference. *arXiv preprint* arXiv:2011.04216.

[25] Burnett, J. W., & Blackwell, C. (2023). Graphical causal modelling: An application to identify and estimate cause-and-effect relationships. *Applied Economics,* 55(1), 1–15.

[26] Shohei, S., Shimizu, S., & Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12, 1225.

[27] Kaur, P., Polyzou, A., & Karypis, G. (2019). Causal inference in higher education: Building better curriculums. In Proceedings of the *6th ACM Conference on Learning at Scale* (pp. 1–4).

[28] Tadayon, M., & Pottie, G. (2021). Causal inference in educational systems: A graphical modeling approach. *arXiv preprint* arXiv:2108.00654.

[29] Aragam, B. (2024). Greedy equivalence search for nonparametric graphical models. *arXiv preprint* arXiv:2406.17228.

[30] Ouaadi, I., & Ibourk, A. (2023). Causal discovery and features importance analysis: What can be inferred about at-risk students? In Proceedings of the *International Conference on Business Intelligence* (pp. 134–145). Springer.

[31] Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search (2nd ed.). *MIT Press.*

[32] Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models* (pp. 301–354). Springer.

[33] Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research,* 7, 2003–2030.

[34] de Carvalho, W. F., & Zarate, L. E. (2019). Causality relationship among attributes applied in an educational data set. In Proceedings of the *34th ACM/SIGAPP Symposium on Applied Computing* (pp. 1271–1277).

[35] Forney, A., & Mueller, S. (2022). Causal inference in AI education: A primer. *Journal of Causal Inference*, 10(1), 141–173.

[36] Dulhare, U. N., Jyothi, D. N., Balimidi, B., & Kesaraju, R. R. (2023). Classification models in education domain using PSO, ABC, and A2BC metaheuristic algorithm-based feature selection and optimization. In M. A. Jabbar, P. K. Reddy, & B. B. Chaudhuri (Eds.), *Machine Learning and Metaheuristics: Methods and Analysis* (pp. 255–270). Springer.

[37] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

[38] Kohavi, R., & Provost, F. (1998). Glossary of terms for AI & data mining. *Machine Learning*, 30(2–3), 271–274.

[39] Weidlich, W., Hicks, B., & Drachsler, H. (2023). Causal reasoning with causal graphs in educational technology research. *Educational Technology Research and Development*, 71, 1–19.

[40] Imbens, G., & Rubin, D. (2015). Causal inference for statistics, social, and biomedical sciences: *An introduction. Cambridge University Press.*

[41] Soo, H. N., Rahman, M. S., & Majumder, A. (2022). Machine learning-based student performance prediction: A systematic review. In Proceedings of *the IEEE International Conference on AI & Big Data* (AIBD) (pp. 1–6).

[42] Guo, R., Cheng, L., Li, J., Hahn, P., & Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4), 1–37.

[43] A. F. Wise, (2018). Learning Analytics: Using Data-Informed Decision-Making to Improve Teaching and Learning, *Contemporary Technologies in Education*, pp. 119–143, doi: 10.1007/978-3-319-89680-9_7.

[44] Dietterich, T. G. (2002). Machine learning for sequential data: A review. *In Proceedings of the International Conference on AI & Statistics* (pp. 15–23).

[45] Z. Zhang, (2019). Distinguishing between mediators and confounders is important for the causal inference in observational studies, *AME Medical Journal*, 4;35–35, doi: 10.21037/amj.2019.09.03.

[46] Y. Fang, (2024). Randomized Controlled Clinical Trials, *Causal Inference in Pharmaceutical Statistics*, pp. 19–39, doi: 10.1201/9781003433378-2.

[47] S. Shimizu, (2022). Correction to: Statistical Causal Discovery: LiNGAM Approach, pp. C1–C1, 2022, doi: 10.1007/978-4-431-55784-5_7.

[48] ZHANG, J. (2025). Artificial intelligence contributes to the creative transformation and innovative development of traditional Chinese culture. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.860

[49] M. Shanthalakshmi, & R.S. Ponmagal. (2025). An Intelligent Intrusion Detection System for VANETs Using Adaptive Fusion Models. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.935

[50] K.S. Praveenkumar, & R. Gunasundari. (2025). Optimizing Type II Diabetes Prediction Through Hybrid Big Data Analytics and H-SMOTE Tree Methodology. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.727

[51] Vijayadeep GUMMADI, & Naga Malleswara Rao NALLAMOTHU. (2025). Optimizing 3D Brain Tumor Detection with Hybrid Mean Clustering and Ensemble Classifiers. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.719

[52] Amjan Shaik, Bhuvan Unhelkar, & Prasun Chakrabarti. (2025). Exploring Artificial Intelligence and Data Science-Based Security and its Scope in IoT Use Cases. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.869

[53] S. Shankar, N. Padmashri, N. Shanmugapriya, S. Ramasamy, & P.S. Sruthi. (2025). IntelliFuzz: An Advanced Fuzzy Logic Framework for Dynamic Evaluation of Student Performance in Open-Ended Learning Tasks. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.911

[54] Sagiraju, S., Mohanty, J. R., & Naik, A. (2025). Hyperparameter Tuning of Random Forest using Social Group Optimization Algorithm for Credit Card Fraud Detection in Banking Data. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.777

[55] Badugu Sobhanbabu, & K.F. Bharati. (2025). Towards Precision Medicine with Genomics using Big Data Analytics. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.906

[56] I. Prathibha, & D. Leela Rani. (2025). Rainfall Forecasting in India Using Combined Machine Learning Approach and Soft Computing Techniques : A HYBRID MODEL. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.785

[57] Abu-Shaikha, M., & Nasereddin, S. (2025). Predicting Media Impact: A Machine Learning Framework for Optimizing Corporate Communication Strategies in Architectural Practices. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1032