**Al-Ishlah: Jurnal Pendidikan**
Vol.15, 2 (June, 2023), pp. 2481-2497
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v15i2.2831

# A Data-Driven Causal Modelling Analysis of Socio-Economic Factors and Its Impact on Student's Performance: A Case Study of a Junior High School in Bali

**Sabar Aritonang Rajagukguk[1], Dhomas Hatta Fudholi[2], Ahmad Rafie Pratama[3]**

[1] Universitas Islam Indonesia, Yogyakarta, Indonesia; 20917054@students.uii.ac.id
[2] Universitas Islam Indonesia, Yogyakarta, Indonesia; hatta.fudholi@uii.ac.id
[3] Universitas Islam Indonesia, Yogyakarta, Indonesia ahmad.rafie@uii.ac.id

**ARTICLE INFO**

***Keywords:***

Causal modelling;
NOTEARS;
Bayesian Networks;
student's performance analysis;
student socio-economic factors

***Article history:***

Received 2022-09-26
Revised 2022-10-30
Accepted 2023-06-23

**ABSTRACT**

Efforts to understand student performance have long been a highly-researched topic in the field of applied education computing. Current research in the field still places its focus on understanding and analyzing student performance using definitive variables such as the student's scores and their cognitive capabilities, which by themselves already explain the student's performance. The great diversity of Indonesian culture, which includes people from a wide range of socioeconomic origins, makes it all the more surprising that so little research has been done to examine the hidden socioeconomic aspects that may affect student performance. Research conducted on a single school may not be generalizable because of the diversity among them in terms of the elements that affect students' academic outcomes. In this investigation, we employ a causal modelling strategy that is data-driven to examine academic achievement. Data was retrieved from a public junior high school in Bali, Indonesia, and then processed with the Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure Learning (NOTEARS) and Bayesian Network algorithm in order to discover a latent causal structure and the effect between variables discovered from the structure. Findings show that the average skill score of a student is significantly influenced by the distance from school, the education level and income of parents, and their place in the family. Meanwhile, the average knowledge score is mainly influenced by the average skill score, the order in the family, and the parent's income level. The results of the study also show potential for practical implications where schools, researchers, and governments, can rethink their approach to education by analyzing data with the proposed approach. The limitations of this study include the quality of data to discover patterns and the limited number of variables used to study student performance factors. Future research may consider the use of more holistic, complete variables in order to discover more insights regarding student performance.



**Corresponding Author:**
Sabar Aritonang
Universitas Islam Indonesia, Yogyakarta, Indonesia; 20917054@students.uii.ac.id

## 1. INTRODUCTION

Understanding the factors that influence student performance has long been an interest of education researchers (Fleming & Malone, 1983; Rajagukguk, 2021; Rebai et al., 2020; Sandra et al., 2021). The main aim is to discover strategies to improve student learning performance (D. Liu et al., 2020; Rajagukguk, 2021), predict achievement (Anderkova et al., 2022; Bendangnuksung & Prabu, 2018; Data et al., 2021), and to determine appropriate managerial strategies in creating an ecosystem that can enhance performance (Albreiki et al., 2021; Bunce et al., 2017; Deng et al., 2019). So far, it has been found that socio-economic factors have an association with student achievement (Fleming & Malone, 1983; Ramaswami & Rathinasabapathy, 2012; Rebai et al., 2020). However, with the abundant data on students, most research has not used data-driven causal modelling approaches in an attempt to discover causality, rather than only associations, between variables (Anderkova et al., 2022; Ramaswami & Rathinasabapathy, 2012).

Previous research has explored a similar data-driven approach. In response to the challenges in predicting student performance, Ramaswami & Rahinasabapathy (2012) used a Network Augmented with Tree (TN) and a Bayesian Network approach (Ramaswami & Rathinasabapathy, 2012). Using six thousand data from three districts in India, which contained variables including the student's test scores, gender, body mass index (BMI), disability status, eating habits, living conditions, number of siblings, status in the family, vehicle ownership, parent profile, and school profile (Ramaswami & Rathinasabapathy, 2012). It was found that the discretization of test values into two categories, namely pass and fail, resulting in a model with higher accuracy than other discretization schemes (Ramaswami & Rathinasabapathy, 2012). Students' performance depended on the suitability of the residence, daily scores, the number of siblings, the type of transportation used, the type of outdoor sport preferred, and the father's income (Ramaswami & Rathinasabapathy, 2012).

Although Ramaswami & Rathinasabapathy's (2012) study has shown that socio-economic factors have an influence on student performance (Ramaswami & Rathinasabapathy, 2012), the study has not reported the effect or conditional probability distributions of the student's test scores, given the conditions of other variables. Furthermore, counterfactual or intervention analysis to view the effects of an intervention on certain variables of performance has also not been conducted. Also, as the study was only done on specific Indian schools, its verifiability and applicability in other schools that may or may not be in other countries are questionable.

Current research in the field of student performance still focuses on characterizing or predicting student performance based on definitive or concrete variables such as students' scores and capabilities, which by themselves can already be used to explain student performance. Little research has been done to understand student performance based on the student's socio-economic background, especially so in Indonesia. This is very unfortunate, as in Indonesia, with thousands of islands and diverse cultures, differing socio-economic backgrounds in various schools are a norm and can be very well related to each individual student's performance.

This study presents a data-driven framework of analysis using the Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS) and Bayesian network algorithm to provide deeper insights into the causality relationship between a student's performance and socio-economic factors. It also presents each variable's effect on the student's performance through the conditional probability distributions, which are then further analyzed and verified using intervention analysis. Using the proposed framework, schools and researchers alike are able to analyze their own case studies in order to discover insights and drive personalized strategic decisions to improve student performance. Furthermore, the results of this research have practical implications for government regulations, specifically for educational policies. Through the analysis, governments can, for example, personalize subsidies and create education programs for students with certain socio-economic backgrounds to improve their performance.

Students' performance is generally understood as an outcome of the learning process (Jihad, 2009; Sukmadinata, 2005). It is of general consensus that student performance is influenced and also influences various, such as their self-confidence and learning motivation (Park & Kim, 2022), leadership

styles of principals and teachers (Gümüş et al., 2022), socio-economic conditions (Bosch et al., 2021), gender (Alanzi, 2018), and self-regulation ability (Park & Kim, 2022). Besides latent factors, various statistical and machine learning approaches have also been proposed, such as the clustering of student characteristics and their performance (Helal et al., 2018; Tinuke Omolewa et al., 2019), classification of student's performance based on their characteristics, (Helal et al., 2018; Maheswari et al., 2021), and even the incorporation of social media analysis of students in relation to their performance (Dzogbenuku et al., 2022). However, noting that every student and every school are inherently diverse (Fleming & Malone, 1983), efforts to understand student performance can be made based on data instead (Anderkova et al., 2022; Y. Liu et al., 2021; Ramaswami & Rathinasabapathy, 2012). One approach is the data-driven construction of Bayesian networks (Ramaswami & Rathinasabapathy, 2012; Zheng et al., 2018) using various data obtained through educational data mining (Hernández-Blanco et al., 2019; Injadat et al., 2020).

Previous studies have extensively explored the use of Bayesian networks for student performance. It is shown that bayesian networks effectively predict students' performance and discover each student's weaknesses and strengths(Xing et al., 2021). In addition, bayesian networks are also used to classify student learning achievements (Sundar, 2013), predict performance based on one's activities on a Massive Open Online Course (MOOC) platform (Hao et al., 2022), and automate assessment results (Xing et al., 2021). Bayesian networks are typically represented as a Directed Acyclic Graph (DAG), with nodes denoting entities and edges denoting the connections between those entities (Kim & Jun, 2013; Wilson et al., 2016; Yakymenko et al., 2020). This approach is used to discover potential causal relationships with little-to-no theoretical basis through data (Yakymenko et al., 2020; Zheng et al., 2018). Bayes' theorem, the foundational theorem for Bayesian networks, is founded on the formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(A)$ is the probability of $A$, $P(B)$ is the probability of $B$ occurring, $P(B|A)$ is the probability of $B$ occurring given $A$, and $P(A|B)$ is the probability of $A$ occurring given $B$ (posterior probability) (Kim & Jun, 2013; Wilson et al., 2016; Yakymenko et al., 2020; Zheng et al., 2018).

Data availability in the digital era has made the implementation of Bayesian networks relevant to understanding the causal relationship behind the data (Mueller-Langer et al., 2020; Yakymenko et al., 2020; Zheng et al., 2018). Previous studies have attempted to find the relationship between variables in the context of student performance. However, none have used national education data to facilitate differences in the characteristics of each school. Researchers have made numerous attempts to determine a causal structure from data (Kuleshov and Ermon, 2021). However, all of these algorithms confront difficulties when attempting to estimate the structure of the directed acyclic graph (DAG) (Zheng et al., 2018). Due to the combinatorial nature of the DAG search space, the computing effort calls for resources that scale super exponentially with node or variable additions (Zheng et al., 2018). Therefore, the NOTEARS approach was used because the amount of processing required grows cubically rather than exponentially. Additionally, it has been demonstrated that its effectiveness and convenience of use outperform earlier approaches (Zheng et al., 2018).

Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS) is a score-based, structure learning algorithm used to discover directed acyclic graphs (DAGs) or Bayesian networks from data. It attempts to solve the problem of current structure learning methods, which are combinatorial in nature and requires computational resources that scale super exponentially with the growth in nodes. NOTEARS offers a new approach for score-based learning, where it transforms the original combinatorial optimization problem (see left of Figure 1) into a continuous optimization problem (see right of Figure 1) that avoids the combinatorial constraint (Zheng et al., 2018).

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to } G(W) \in \text{DAGs} \iff \min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to } h(W) = 0, \tag{1}$$

**Figure 1.** DAG Optimization Problem Equation (Zheng et al., 2018)

Where $d$ is the number of nodes, $F(W)$ is a score function, $G(W)$ is the graph representation of the weighted adjacency matrix $W$ with $d$ nodes, and $h$ is a smooth function over real matrices $\mathbb{R}^{d \times d}$, where when its level is set at zero characterizes acyclic graphs. By eliminating the combinatorial constraint, the structure of a DAG can be learned without traversing through the combinatorial space of DAGs. Furthermore, this allows for the use of standard numerical algorithms making implementation effortless. The continuous problem is then solved using the augmented Lagrangian method, which breaks the constrained problem into a series of unconstrained subproblems. These subproblems are then solved efficiently through standard numerical algorithms such as the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Zheng et al., 2018).

## 2. METHODS

The following figure describes the steps used in the implementation of this study.

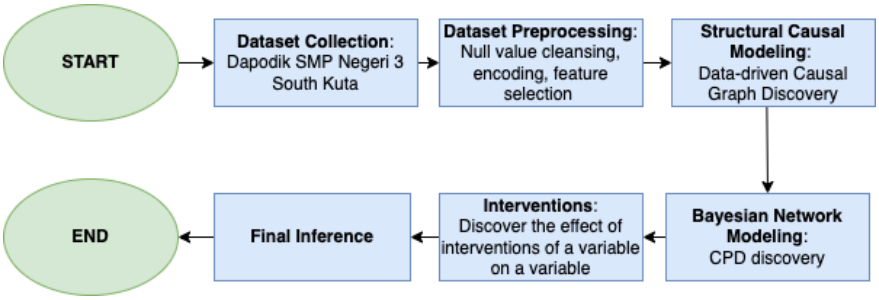


**Figure 2.** Illustration of the proposed methodology

The study begins with retrieving education data from Dapodik, an integrated nationwide data collection system used by the Ministry of Education in Indonesia. The data used for this study is data from a junior public high school in Bali. In order to guarantee that only clean data is utilized for analysis, data preprocessing is done after data collection. This includes removing empty values, encoding non-numerical variables into their numeric form, and feature selection. The NOTEARS algorithm is then used to analyse the preprocessed data for a structural causal modelling analysis. A Bayesian network model is then constructed using the generated causal structure to assess each variable's conditional probability distribution (presumably the effect) onto other variables. The model is then used to conduct interventions in order to determine how a change in a variable will affect other variables. The Python programming language and open-source libraries such as *pandas*, *numpy*, and *causalnex* were used to conduct all data preprocessing and analyses.

## 3. FINDINGS AND DISCUSSION

### *3.1 Dataset Collection*

In order to develop a causal model based on Dapodik, data collection has been carried out from a junior high school in Bali. The name of the school is not disclosed in order to protect the anonymity of the school. Overall, the total number of data collected is 292. The data to be used are students' identity (position in the family, whether they have special needs), socio-economic (distance from school,

parental data such as parent's education level, occupation, and income), and student achievement. The following table describes the variables used in detail.

**Table 1.** Dataset Variables

| Variable | Description |
|---|---|
| *Jenis Tinggal* (Staying With) | Who the student is staying with. It can be either parents or guardians. |
| *Penerima Kartu Perlindungan Nasional* (National Protection Card Recipient) | Whether the student has received the national protection card. |
| *Penerima Kartu Indonesia Pintar* (Smart Indonesian Card Recipient) | Whether the student has received the smart Indonesian card. |
| *Kebutuhan Khusus* (Special Needs) | List of the student's special needs. |
| *Anak ke-Berapa* (Child Order in the Family) | Self-explanatory. |
| *Tinggi Badan* (Height) | Self-explanatory. |
| *Berat Badan* (Weight) | Self-explanatory. |
| *Jarak Rumah ke Sekolah* (km) (Distance from School in kilometers) | Self-explanatory. |
| *Identitas Ayah – Tahun Lahir* (Father Identity – Year of Birth) | Self-explanatory. |
| *Identitas Ayah – Jenjang Pendidikan* (Father Identity – Education) | Self-explanatory. |
| *Identitas Ayah – Penghasilan* (Father Identity – Income) | Self-explanatory. |
| *Identitas Ayah – Berkebutuhan Khusus* (Father Identity – Special Needs) | List of the student's father's special needs. |
| *Identitas Ibu – Tahun Lahir* (Mother Identity – Year of Birth) | Self-explanatory. |
| *Identitas Ibu – Jenjang Pendidikan* (Mother Identity – Education) | Self-explanatory. |
| *Identitas Ibu – Penghasilan* (Mother Identity – Income) | Self-explanatory. |
| *Identitas Ibu – Berkebutuhan Khusus* (Mother Identity – Special Needs) | List of the student's mother's special needs. |
| *Identitas Wali – Tahun Lahir* (Guardian Identity – Year of Birth) | Self-explanatory. |
| *Identitas Wali – Jenjang Pendidikan* (Guardian Identity – Education) | Self-explanatory. |
| *Identitas Wali – Penghasilan* (Guardian Identity – Income) | Self-explanatory. |
| *Identitas Wali – Berkebutuhan Khusus* (Guardian Identity – Special Needs) | Self-explanatory. |
| *Riwayat Kesejahteraan – Jenis Kesejahteraan* (Wellness Benefit History – Type of Benefit) | List of wellness benefits given to the student. |
| *Riwayat Beasiswa – Jenis Beasiswa* (Scholarship History – Type of Scholarship) | List of scholarships given to the student. |
| *Nilai Sikap Sosial* (Social Attitude Score) | Self-explanatory. |

| Variable | Description |
|---|---|
| *Nilai Sikap Spiritual* (Spiritual Attitude Score) | Self-explanatory. |
| *Nilai Rata-rata Pengetahuan* (Average Knowledge Score) | Self-explanatory. |
| *Nilai Rata-rata Keterampilan* (Average Skill Score) | Self-explanatory. |

### *3.2 Dataset Preprocessing*


**Figure 3.** Preprocessing Steps

The collected data is then preprocessed through several procedures to ensure it is ready for modelling. Columns with a relatively large number of incomplete values are discarded first. After that, non-ordinal categorical data were transformed into their numeric form using the one-hot encoding process, where each to-be-encoded variable is replaced with the creation of a binary variable for each possible value of the variable. Ordinal categorical data, such as income levels, were converted in their respective order. As variables with a very high correlation (>0.8) can be represented by one or the other, removing highly correlated features was also done to reduce the computational complexity. The following table shows the final variables used from the dataset after preprocessing was conducted.

**Table 2.** Preprocessed Variables

| Variable | Status | Reason for Non-use |
|---|---|---|
| *Jenis Tinggal* (Staying With) | Not used | The variable's value did not vary as all students stayed with their parents. |
| *Penerima Kartu Perlindungan Nasional* (National Protection Card Recipient) | Not used | The variable's value did not vary as all students did not receive the national protection card. |
| *Penerima Kartu Indonesia Pintar* (Smart Indonesian Card Recipient) | Not used | The variable's value did not vary as all students did not receive the smart Indonesian card. |
| *Kebutuhan Khusus* (Special Needs) | Not used | The variable's value did not vary as all students did not have any special needs. |
| *Anak ke-Berapa* (Child Order in the Family) | Used | N/A |
| *Tinggi Badan* (Height) | Used | N/A |
| *Berat Badan* (Weight) | Used | N/A |
| *Jarak Rumah ke Sekolah* (km) (Distance from School in kilometers) | Used | N/A |
| *Identitas Ayah – Tahun Lahir* (Father Identity – Year of Birth) | Used | N/A |

| Variable | Status | Reason for Non-use |
|---|---|---|
| *Identitas Ayah – Jenjang Pendidikan* (Father Identity – Education) | Used | N/A |
| *Identitas Ayah – Penghasilan* (Father Identity – Income) | Used | N/A |
| *Identitas Ayah – Berkebutuhan Khusus* (Father Identity – Special Needs) | Used | N/A |
| *Identitas Ibu – Tahun Lahir* (Mother Identity – Year of Birth) | Used | N/A |
| *Identitas Ibu – Jenjang Pendidikan* (Mother Identity – Education) | Used | N/A |
| *Identitas Ibu – Penghasilan* (Mother Identity – Income) | Used | N/A |
| *Identitas Ibu – Berkebutuhan Khusus* (Mother Identity – Special Needs) | Used | N/A |
| *Identitas Wali – Tahun Lahir* (Guardian Identity – Year of Birth) | Not used | The variable's value was not used as all students stayed with their parents. Therefore, guardian-related columns were empty. |
| *Identitas Wali – Jenjang Pendidikan* (Guardian Identity – Education) | Not used | N/A |
| *Identitas Wali – Penghasilan* (Guardian Identity – Income) | Not used | N/A |
| *Identitas Wali – Berkebutuhan Khusus* (Guardian Identity – Special Needs) | Not used | N/A |
| *Riwayat Kesejahteraan – Jenis Kesejahteraan* (Wellness Benefit History – Type of Benefit) | Not used | The values for this variable did not vary as all students received a health guarantee for benefits. |
| *Riwayat Beasiswa – Jenis Beasiswa* (Scholarship History – Type of Scholarship) | Not used | The values for this variable did not vary as all students did not receive any type of scholarship. |
| *Nilai Sikap Sosial* (Social Attitude Score) | Used | N/A |
| *Nilai Sikap Spiritual* (Spiritual Attitude Score) | Used | N/A |
| *Nilai Rata-rata Pengetahuan* (Average Knowledge Score) | Used | N/A |

| Variable | Status | Reason for Non-use |
|---|---|---|
| *Nilai Rata-rata Keterampilan* (Average Skill Score) | Used | N/A |

### *3.3 Structural Causal Modeling*

After the data has been preprocessed, the data is now ready to be used in order to find the causal relationship between variables through causal modelling. The causal modelling approach used in this study is causal structure learning analysis using the NOTEARS algorithm to create a causal relationship structure graph (see Figure 3).

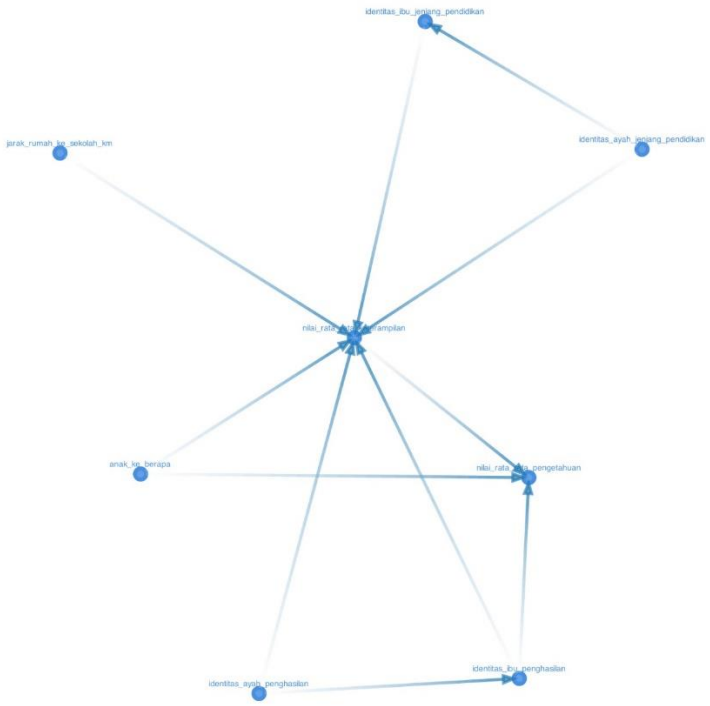


**Figure 4.** Causal graph structure of students' performance

Figure 3 shows that there is one dependent or most influenced variable, namely the average knowledge score. Attitude-related scores were not included in the model as their effect on other variables, and the effect of other variables on the attitude-related score is insignificant. This may be because all of the students received a B predicate for the attitude-related scores, making the values unvarying and, thus, the reason why it had no effect. Returning to the average knowledge score, we find that the mother's income strongly influences it, the average skill score, and the student's position in the household. Indirect factors include the student's father's income (through the mother's income), the student's parents' education levels (through the student's average skill score), and the students' parents' relative positions in the family (through the student's average skill score).

Other than that, the average skill score is directly influenced by the distance from school, the student's order in the family, and both parents' educational level and income. An interesting finding from this is that, firstly, the model is able to capture previously confirmed relationships, such as the influence of parents' educational level and income on the student's performance. Also, here, it can be seen that the average skill score becomes an important mediator for various variables in improving the average knowledge score. This indicates that in the case of data from the school used in this study, students' skills significantly influence their knowledge.

### *3.4 Bayesian Network Modeling*

A Bayesian network model was then constructed based on the created structure to examine the effect of each variable in the structure on student performance. Values that are continuous or have a very wide range of possible values were first discretized into the two categories of low and high because the library used to build the Bayesian network only supports discrete values. Each variable's conditional probability distributions (CPD) were mapped using the Bayesian network model after the data had undergone discretization, and the relationship between the variables was imposed using the built causal relationship model. Furthermore, interventions were conducted to discover how a change in one variable will affect the student's performance.

Figure 4 shows the CPD for the average knowledge score variable. It was found that the average knowledge score had the highest probability of being categorized as high if the average skill score was high, the mother's income was high, and if the student was the oldest child in the family (see Figure 4), with a CPD of 83,5%. This indicates that the eldest child or older child tends to have better knowledge scores and shows how the income of a student's parents plays a significant role in their performance. Furthermore, this also indicates that the youngest child tends to have a better average skill score.

CPD of: nilai_rata_rata_pengetahuan

| anak_ke_berapa | High | | | | Low | | | |
|---|---|---|---|---|---|---|---|---|
| identitas_ibu_penghasilan | High | | Low | | High | | Low | |
| nilai_rata_rata_keterampilan | High | Low | High | Low | High | Low | High | Low |
| nilai_rata_rata_pengetahuan | | | | | | | | |
| High | 0.666667 | 0.4 | 0.714286 | 0.333333 | 0.835821 | 0.10101 | 0.766667 | 0.049383 |
| Low | 0.333333 | 0.6 | 0.285714 | 0.666667 | 0.164179 | 0.89899 | 0.233333 | 0.950617 |

**Figure 5.** CPD of average skill score

It also appears that efforts to increase the average skill score need to be prioritized for students that are the eldest child in their families. Meanwhile, for the youngest child, various skill-based learning activities can be prioritized or further improved as it will significantly increase their knowledge score. Another finding in this study is that the higher the level of education and the higher the income of parents, the greater the chance that students that are the youngest child will get a high average skill score. The student's distance from school also plays a significant role, where the closer the student's house is to the school, the higher the chances of the student having a higher average skill score (see Figure 5).

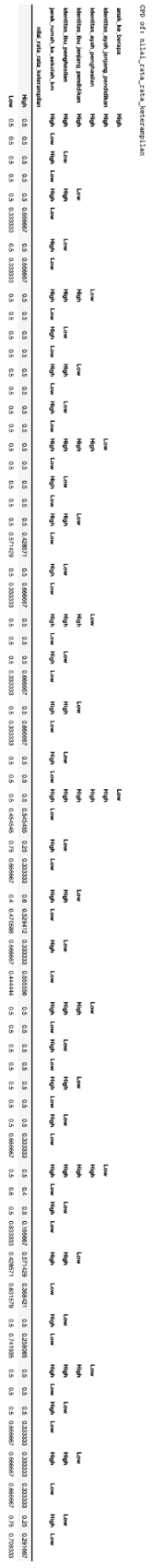

**Figure 6.** CPD of average knowledge score

Figure 5 also shows that for students that are the eldest child, the parent's education level does not significantly affect the average skill score; however, the parent's income still significantly affects the average skill score. Also, the distance from home does not significantly affect the average skill score.

### *3.5 Interventions*

The constructed Bayesian network model was also used to query data with various interventions (counterfactual analysis) to gain more insights. The goal of the intervention was to calculate the likelihood that each variable would fall into the very high category if another variable were set to high (100% probability; see Table 3). The formula $P(Y|do(X))$, which represents the probability distribution of $Y$ given $X$ is set to a specific value, provides the foundation for the probability calculation used in the intervention.

**Table 3.** Comparison of probabilities pre- and post-intervention

| Intervened Variable | Variable | Pre-intervention Probabilities (Very High) | Post-intervention Probabilities (Very High) |
|---|---|---|---|
| *Anak ke-Berapa* (Child Order in the Family) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 100% |
| | *Nilai Rata-rata Pengetahuan* (Average Knowledge Score) | 35,6% | 54,1% |
| | *Nilai Rata-rata Keterampilan* (Average Skill Score) | 37,3% | 52,6% |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |
| *Jarak Rumah ke Sekolah (km)* (Distance to School) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |
| | *Nilai Rata-rata Pengetahuan* (Average Knowledge Score) | 35,4% | 45,2% |
| | *Nilai Rata-rata Keterampilan* (Average Skill Score) | 37,3% | 50,4% |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 100% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |

| Intervened Variable | Variable | Pre-intervention Probabilities (Very High) | Post-intervention Probabilities (Very High) |
|---|---|---|---|
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |
| *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |
| | *Nilai Rata-rata Pengetahuan* (Average Knowledge Score) | 35,4% | 45,0% |
| | *Nilai Rata-rata Keterampilan* (Average Skill Score) | 37,3% | 50,5% |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 100% |
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |
| *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |
| | ***Nilai Rata-rata Pengetahuan* (Average Knowledge Score)** | **35,4%** | **35,2%** |
| | ***Nilai Rata-rata Keterampilan* (Average Skill Score)** | **37,3%** | **36,7%** |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |
| | ***Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education)** | **10,4%** | **100%** |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |

| Intervened Variable | Variable | Pre-intervention Probabilities (Very High) | Post-intervention Probabilities (Very High) |
|---|---|---|---|
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |
| *Identitas Ayah – Penghasilan* (Father Identity – Income) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |
| | ***Nilai Rata-rata Pengetahuan* (Average Knowledge Score)** | **35,4%** | **36,3%** |
| | ***Nilai Rata-rata Keterampilan* (Average Skill Score)** | **37,3%** | **37,9%** |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | ***Identitas Ayah - Penghasilan* (Father Identity – Income)** | **86,9%** | **100%** |
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |
| *Identitas Ibu – Penghasilan* (Mother Identity – Income) | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |
| | ***Nilai Rata-rata Pengetahuan* (Average Knowledge Score)** | **35,4%** | **40,4%** |
| | ***Nilai Rata-rata Keterampilan* (Average Skill Score)** | **37,3%** | **40,8%** |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |
| | ***Identitas Ibu - Penghasilan* (Mother Identity – Income)** | **59,8%** | **100%** |
| | *Anak ke-Berapa* (Child Order in the Family) | 0,4% | 0,4% |

| Intervened Variable | Variable | Pre-intervention Probabilities (Very High) | Post-intervention Probabilities (Very High) |
|---|---|---|---|
| *Nilai Rata-rata Keterampilan* (Average Skill Score) | ***Nilai Rata-rata Pengetahuan* (Average Knowledge Score)** | **35,4%** | **80,2%** |
| | ***Nilai Rata-rata Keterampilan* (Average Skill Score)** | **37,3%** | **100%** |
| | *Jarak Rumah ke Sekolah* (Distance to School) | 0,8% | 0,8% |
| | *Identitas Ayah - Jenjang Pendidikan* (Father Identity – Education) | 15,4% | 15,4% |
| | *Identitas Ibu - Jenjang Pendidikan* (Mother Identity – Education) | 10,4% | 10,4% |
| | *Identitas Ayah - Penghasilan* (Father Identity – Income) | 86,9% | 86,9% |
| | *Identitas Ibu - Penghasilan* (Mother Identity – Income) | 59,8% | 59,8% |

When the child order variable is intervened so that all students were the youngest child in the family, a significant increase in the probability of a higher average knowledge value and average skill score can be seen. Relating to the results found in the causal structural graph presented in Figure 3 and the CPD shown in Figure 4, it can be concluded that this increase is due to students who are "younger" in their families have a tendency to score well in the average skill scores, which directly affects the average knowledge score, thus explaining the large increase in both probabilities. The student's home distance from school also has an influence on their performance. For students whose house-to-school distance is shorter, the probability of their performance increasing is positively influenced, shown by an increase of 9.8% for the average knowledge score and 13.1% for the average skill score probabilities being high.

The educational level of a student's father also appears to significantly influence student performance. As seen in Table 3, there is an increase of 9.6% in the probability of the average knowledge score being very high and 13.7% for the average skill score. However, a different influence is found in the student's mother's education level. It is shown that there is no effect, and instead a decrease in the probability of students achieving high performance, by -0.2% and -0.6% for the average knowledge and skill scores, respectively. This is interesting because it contradicts the previous research findings by Ramaswami & Rathinasabapathy's (2012) study, which showed that both parents' educational levels should positively influence students' performance. This goes to show that every school in differing locations may have differences in the influence of each variable on the student's performance. This also implies that each school, ideally, should and can analyze data using the same proposed approach in order to build better learning approaches suitable for their students.

Next, the parental income of both parents seems to also positively influence student performance. As seen in the table, a higher father's income increases the probability of the student achieving high performance by 0,9% and 0,6% for the average knowledge score and average skill score, respectively. On the other hand, a higher mother's income increases the chance of higher performance by 5% and 3,5%, which is a significant increase when compared to the father's income. This may imply

that in regard to parental income, the mother's role in the school used in this study is of more importance than the father's.

It was also found that the intervention on the average skill score significantly influences the probability of the student getting a high average knowledge score, with a 44,8% increase from 35,4% to 80,2%. The average skill score can therefore be seen as an important variable that should be improved in order to significantly increase the average knowledge score and, in general, student performance. This finding contributes to the existing literature, which shows that not only do socio-economic factors influence student performance, but performance-related variables might affect one another as well. The question is, therefore, "why" and "how" the interrelated performance variable affect each other can be answered through future research.

## 4. CONCLUSION

The findings of this study showed that the average skill score is the main mediator in influencing the average knowledge score. Attitude values which are also part of the student's competence factors, do not significantly contribute to the causal model because the values used from the dataset were unvarying. This is unfortunate because an understanding of the causal relationship involving the attitude values of students cannot be found. The limitations of quality of the data processed and the usage of selected variables of student profiles and their family backgrounds are also limitations of this study.

In order to improve future analysis, schools may consider improving the implementation of assessment of knowledge, skills, and attitudes strategies so that further insight can be found. Future research can also consider the use of more diverse data, including variables related to the profile of teachers. This needs to be done to find out how the influence between teacher profiles and student performance based on data. Another thing that needs to be developed is the preparation of a model that schools can use to facilitate the diversity of data for each school; to discover causal relationships between variables in the data owned by each individual school.

## REFERENCES

Alanzi, K. A. (2018). Female accounting students and their academic performance: evidence from Kuwait. Journal of Islamic Accounting and Business Research, 9(5), 662–672. https://doi.org/10.1108/JIABR-10-2016-0128

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. Education Sciences, 11(9). https://doi.org/10.3390/educsci11090552

Anderkova, V., Adam, T., & Babic, F. (2022). Data-driven Student Performance Prediction. SAMI 2022 - IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics, Proceedings, 267–272. https://doi.org/10.1109/SAMI54271.2022.9780854

Bendangnuksung, & Prabu, D. (2018). Students' Performance Prediction Using Deep Neural Network. International Journal of Applied Engineering Research, 13(2), 1171–1176. http://www.ripublication.com

Bosch, E., Seifried, E., & Spinath, B. (2021). What successful students do: Evidence-based learning activities matter for students' performance in higher education beyond prior knowledge, motivation, and prior achievement. Learning and Individual Differences, 91(August), 102056. https://doi.org/10.1016/j.lindif.2021.102056

Bunce, L., Baird, A., & Jones, S. E. (2017). The student-as-consumer approach in higher education and its effects on academic performance. Studies in Higher Education, 42(11), 1958–1978. https://doi.org/10.1080/03075079.2015.1127908

Data, M. B., Zhao, L., Chen, K. U. N., Song, J. I. E., Zhu, X., Sun, J., Caulfield, B., & Namee, B. M. A. C. (2021). Academic Performance Prediction Based on.

https://doi.org/10.1109/ACCESS.2020.3002791

Deng, H., Wang, X., Guo, Z., Decker, A., Duan, X., Wang, C., Alex Ambrose, G., & Abbott, K. (2019). PerformanceVis: Visual analytics of student performance data from an introductory chemistry course. Visual Informatics, 3(4), 166–176. https://doi.org/10.1016/j.visinf.2019.10.004

Dzogbenuku, R. K., Doe, J. K., & Amoako, G. K. (2022). Social media information and student performance: the mediating role of hedonic value (entertainment). Journal of Research in Innovative Teaching & Learning, 15(1), 132–146. https://doi.org/10.1108/JRIT-12-2020-0095

Fleming, M. L., & Malone, M. R. (1983). THE RELATIONSHIP OF STUDENT CHARACTERISTICS AND STUDENT PERFORMANCE IN SCIENCE AS VIEWED BY META-ANALYSIS RESEARCH*. In JOURNAL OF RESEARCH IN SCIENCE TEACHING (Vol. 20, Issue 5).

Gümüş, S., Bellibaş, M. Ş., & Pietsch, M. (2022). School leadership and achievement gaps based on socioeconomic status: a search for socially just instructional leadership. Journal of Educational Administration, 60(4), 419–438. https://doi.org/10.1108/JEA-11-2021-0213

Hao, J., Gan, J., & Zhu, L. (2022). MOOC performance prediction and personal performance improvement via Bayesian network. Education and Information Technologies, 27(5), 7303–7326. https://doi.org/10.1007/s10639-022-10926-8

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. Knowledge-Based Systems, 161(December 2017), 134–146. https://doi.org/10.1016/j.knosys.2018.07.042

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining. Complexity, 2019. https://doi.org/10.1155/2019/1306039

Injadat, M. N., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. Knowledge-Based Systems, 200, 105992. https://doi.org/10.1016/j.knosys.2020.105992

Jihad, A. (2009). Evaluasi Pembelajaran. Multi Pressindo.

Liu, D., Zhang, Y., Zhang, J. U. N., Li, Q., Zhang, C., & Yin, Y. U. (2020). Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction. 8. https://doi.org/10.1109/ACCESS.2020.3033200

Liu, Y., Tian, K., Wang, H., Liu, H., Wu, Y., & Chen, X. (2021). Data-driven based student programming competition award prediction via machine learning models. ICCSE 2021 - IEEE 16th International Conference on Computer Science and Education, 463–468. https://doi.org/10.1109/ICCSE51940.2021.9569407

Maheswari, K., Priya, A., Balamurugan, A., & Ramkumar, S. (2021). Analyzing student performance factors using KNN algorithm. Materials Today: Proceedings, xxxx. https://doi.org/10.1016/j.matpr.2020.12.1024

Park, S., & Kim, N. H. (2022). University students' self-regulation, engagement and performance in flipped learning. European Journal of Training and Development, 46(1–2), 22–40. https://doi.org/10.1108/EJTD-08-2020-0129

Rajagukguk, S. A. (2021). Tinjauan Pustaka Sistematis: Prediksi Prestasi Belajar Peserta Didik Dengan Algoritma Pembelajaran Mesin. Jurnal SNATi, 1.

Ramaswami, M., & Rathinasabapathy, R. (2012). Student Performance Prediction Modelling: A Bayesian Network Approach Framework of Int elligent Recommendat ion Syst em for a Privat e Tert iary Inst it ut ion in Nigeria Student Performance Prediction.

Rebai, S., ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70(August 2018), 100724. https://doi.org/10.1016/j.seps.2019.06.009

Sandra, L., Lumbangaol, F., & Matsuo, T. (2021). Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review. TEM Journal, 10(4), 1919–1927. https://doi.org/10.18421/TEM104-56

Sukmadinata, N. S. (2005). Landasan Psikologi Proses Pendidikan. Remaja Rosda Karya.

Sundar, P. V. P. (2013). A COMPARATIVE STUDY FOR PREDICTING STUDENT'S ACADEMIC PERFORMANCE USING BAYESIAN NETWORK CLASSIFIERS. IOSR Journal of Engineering, 03(02), 37–42. https://doi.org/10.9790/3021-03213742

Tinuke Omolewa, O., Taye Oladele, A., Adekanmi Adeyinka, A., & Roseline Oluwaseun, O. (2019). Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions. Journal of Engineering and Applied Sciences, 14(22), 8254–8260. https://doi.org/10.36478/jeasci.2019.8254.8260

Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2021). Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network Model. Journal of Educational Computing Research, 59(2), 230–256. https://doi.org/10.1177/0735633120960422

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. Advances in Neural Information Processing Systems, 2018-Decem(1), 9472–9483.