Tecnológico de Monterrey

Evidencia 1. Artículo de investigación. Redes bayesianas caso discreto

Johanna C. Willis-Ruiz (A01741070)[1], Victoria González-González (A01737594)[2], Sara Rivera-Méndez (A01068365)[3], Cristobal M. Meza (A01643121)[4], René A. Calzadilla-Calderón (A01246501)[5] and Valeria Aguilar-Meza (A01741304)[5]

[1] *Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Campus Guadalajara*

Fecha: 19/08/2024

Abstract—La pandemia de COVID-19 tuvo un impacto significativo en los sistemas educativos, generando nuevos desafíos para comprender y analizar las dinámicas y condiciones educativas experimentadas por los estudiantes al regresar a las escuelas. Este estudio utiliza redes bayesianas para explorar los datos de la Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE) 2021 en México. El objetivo es identificar y analizar los factores que influyen en las experiencias educativas de los estudiantes durante el ciclo escolar 2021-2022, centrándose en cómo variables como género, edad y nivel educativo se correlacionan con resultados clave, tales como cambios de escuela, bajas y métodos de evaluación. Los hallazgos proporcionan una visión de los desafíos continuos para lograr una educación inclusiva y equitativa, especialmente en el contexto de la recuperación post-pandemia.

Keywords—Redes bayesianas, ENAPE, encuestas, educación,

I. Introducción

El acceso a la educación es un derecho esencial que debe estar garantizado para todas las personas, ya que es un componente crucial para el desarrollo justo y equitativo de cualquier sociedad. Sin embargo, asegurar que este derecho se ejerza de manera igualitaria entre todos los sectores de la población sigue siendo uno de los mayores retos en el presente. Tras la pandemia, la reapertura de las escuelas y el regreso gradual de niñas, niños, adolescentes y jóvenes a las aulas, hizo evidente la necesidad de entender las nuevas dinámicas de estudio y las condiciones en las que se encontraban al retomar sus actividades escolares. Consecuentemente, la Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE) 2021 surge como una herramienta fundamental para identificar a los distintos grupos poblacionales en edad escolar, además del nivel educativo alcanzado y las condiciones en las que se encuentran [1]. El presente artículo tiene como objetivo explorar los resultados de esta encuesta, con la finalidad de extraer y analizar la información relevante que nos permita entender mejor la situación educativa actual y los desafíos que persisten en la búsqueda de una educación inclusiva y equitativa.

II. Métodos

Se utilizarán redes bayesianas para mostrar relaciones de probabilidad y causalidad entre variables con la finalidad de brindar respuesta a las siguientes queries:

- Probabilidad de que la razón principal por la que un hombre se cambió de escuela en el ciclo escolar (2021-2022) haya sido por motivos de la pandemia COVID
- ¿Es más probable que se hayan usado exámenes como método de evaluación dado que la asistencia también haya sido un mérodo de evaluación o que se hayan usado proyectos como método de evaluación dado que la asistencia también haya sido un mérodo de evaluación?
- Probabilidad de que una persona sea mayor de edad (18 o más) y haya estudiado hasta segundo año de secundaria
- ¿Es más probable que la persona de sexo femenino haya dejado de estudiar debido a un casamiento/embarazo/dedicarse a tareas del hogar o por que tenía que trabajar?

Como observación al momento de trabajar la base de datos al tener demasiados datos vacíos y no poderlos descartar, el código nos marcaba error por lo que rellenamos todos los datos vacíos con 50.

a. Redes Bayesianas

Para la resolución de esta situación problema, fue indispensable el uso de Redes Bayesianas, las cuales son un modelo probabilístico gráfico cuya función es representar un conjunto de variables aleatorias y sus dependencias a través de un grafo dirigido acíclico.

*1. **Grafo Dirigido Acíclico***

Un grafo dirigido acíclico (DAG) es aquel que es finito y cuya característica principal es que no permite ciclos, es decir, no es posible regresar a un mismo nodo por medio de sus arcos (relaciones directas entre nodos).

Los DAG se componen de cierto número de "nodos padre", los cuales no dependen de ningún otro nodo. Posteriormente el grafo se expandirá por medio de sus "nodos hijo", que son aquellos que sí dependen de otro u otros nodos y su relación se representa a través de los arcos.

Una propiedad interesante de los DAG es que su estructura puede ser reducida a un punto óptimo en que su recorrido cumpla con todas las relaciones especificadas en el mismo sin ninguna pérdida. Básicamente significa que es posible reducir las relaciones de los nodos hasta un punto mínimo en que dicha reducción no afecta la capacidad de verificar la información de ningún nodo en ningún momento[2].

*2. **Redes Bayesianas Multinomiales***

Durante el desarrollo de este artículo de investigación, se utilizaron en específico las Redes Bayesianas Multinomiales; éstas se definen como un modelo cuya red asociada contiene nodos cuya distribución es multinomial y representan variables categóricas.

b. Hill-climbing en R

Hill-climbing (HC) es un algoritmo basado en valores en un espacio de grafos dirigidos e incluye un método de búsqueda heurística que funciona de manera voraz evaluando el nivel de ajuste de cada estado posible para cada paso [3]. Además, este algoritmo se basa en puntuaciones que maximizan la estructura de una DAG.

*1. **Simple Hill Climbing***

Se utiliza para encontrar una solución cercana a un óptimo local en un espacio de búsqueda.

*2. **Steepest-Ascent Hill Climbing***

Explora todas las opciones a partir de un estado antes de avanzar, eligiendo siempre la opción más beneficiosa.

III. APLICACIÓN

Una vez leídas las preguntas correspondientes a nuestro equipo, se buscó en el catálogo de la base de datos las columnas necesarias para responder las mismas ($EDAD$, $SEXO$, $PB3\_4$, $PC3\_3\_1$, $PC3\_3\_2$, $PA3\_5$, $PA3\_8\_6$, $PA3\_8\_2$, $PA3\_8\_1$) y se le asignó a cada una su respectiva abreviación con base en su(s) grafema(s) inicial(es):

Variable	**Valor**
E	Edad
S	Sexo
C	Razón por la que se cambió de escuela
G	Nivel de escolaridad
NG	Número de grado escolar
DB	Razón por la que se dio de baja
A	Se usó la asistencia como método de evaluación
EX	Se usaron exámenes como método de evaluación
P	Se usaron tareas como método de evaluación

TABLE 1: VALORES CORRESPONDIENTES A CADA VARIABLE

1. ***DAGs***

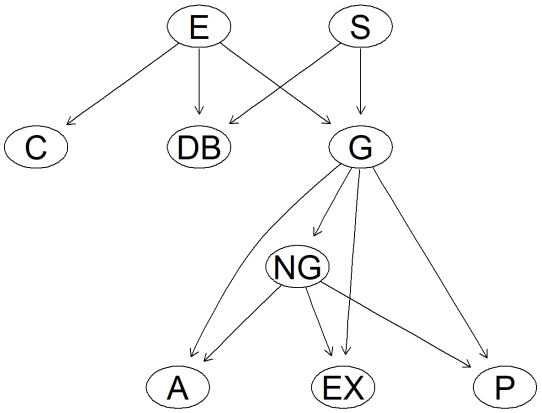

Fig. 1: DAG 1

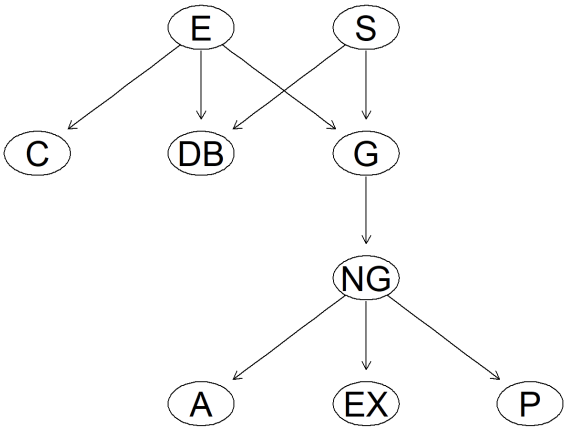

Fig. 2: DAG 2

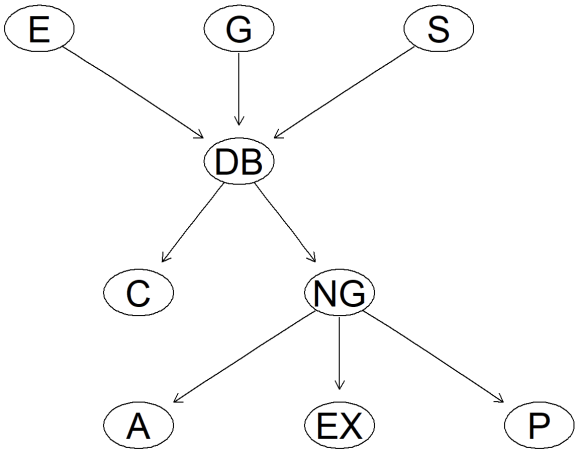

Fig. 3: DAG 3

2. ***Redes bayesianas utilizando las DAGs***

```
bn_mle1 = bn.fit(dag1, data = data_dag, method = "mle")
bn_mle2 = bn.fit(dag2, data = data_dag, method = "mle")
bn_mle3 = bn.fit(dag3, data = data_dag, method = "mle")
```

*3. **Significancia de las relaciones de dependencia***

From	To	Strength
E	C	0.0000000
S	G	1.0000000
E	G	0.0000000
G	NG	0.0000000
S	DB	1.0000000
E	DB	0.7186835
G	A	1.0000000
NG	A	1.0000000
G	EX	1.0000000
NG	EX	1.0000000
G	P	1.0000000
NG	P	1.000000

TABLE 2: DAG 1

From	To	Strength
S	G	1.0000000
E	G	0.0000000
G	NG	0.0000000
S	DB	1.0000000
E	DB	0.7186835
E	C	0.0000000
NG	EX	0.0000000
NG	A	0.0000000
NG	P	0.0000000

TABLE 3: DAG 2

From	To	Strength
S	DB	1.0000000
E	DB	1.0000000
G	DB	1.0000000
DB	NG	0.0000000
DB	C	1.0000000
NG	EX	0.0000000
NG	A	0.0000000
NG	P	0.0000000

TABLE 4: DAG 3

*4. **Red bayesiana con mejor ajuste***

Una vez creadas nuestras DAG´s, se utilizó el criterio de información bayesiano (BIC) para seleccionar el mejor modelo para nuestras preguntas. Con base en esto, los resultados para cada una de nuestras DAG´s fueron:

1. -240858.6
2. -238534
3. -315832.8

Y debido a la naturaleza de la librería utilizada:

```
library(bnlearn)
```

Se debe elegir el resultado "más negativo", por lo que la red bayesiana con mejor ajuste es la tercera.

5. *Uso del algoritmo hill-climbing para obtener la mejor estructura de la DAG para los datos*

```
best_dag = hc(data_dag)
modelstring(best_dag)
graphviz.plot(best_dag, shape = "ellipse")
```

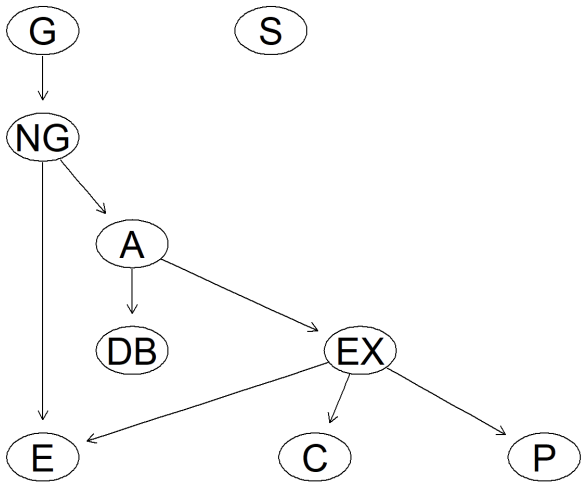

Fig. 4: Mejor DAG

6. *Pertinencia de las relaciones de la mejor DAG*

En esta DAG obtenida por medio del algoritmo hill-climbing podemos notar que hay una variable que es indepediente del resto y además no influye en ninguna de las otras variables, i.e. S. Por otro lado, G determina NG, lo cuál tiene sentido dado que G representa el nivel de escolaridad y NG el nivel del grado, e.g. G = Secundaria y NG = 2 $\rightarrow$ Segundo de Secundaria. Asimismo, es interesante que en esta DAG G y NG afectan a la edad, E; ya que intuitivamente se podría pensar que esta relación de dependencia sucede de manera inversa.

En cuanto a las relaciones entre los métodos de evaluación, podemos notar que es extraño que haya un arco entre A y EX, esto significa que el uso de exámenes como método de evaluación depende de si se usaron las asistencias como método. Por otro lado, el arco $EX \rightarrow P$ sugiere que el uso de tareas/proyectos como evaluación depende de los exámenes.

Por último, vemos que el uso de exámenes (EX) determina la razón por la que se cambió de escuela en el ciclo escolar 2021-2022 (C), lo cual parece tener más sentido dado que si los exámenes fueron complicados y la persona en cuestión reprobó, eso podría implicar un cambio de escuela. Asimismo, las asistencias (A), análogo al caso anterior, podría afectar la razón por la que se dio de baja de la escuela en tal ciclo escolar (DB). Es plausible que la gente no pudiera asistir y que esto causara una baja de calificaciones y que finalmente se diera de baja de la escuela.

7. *Respuesta a las queries*

Con una confianza en las predicciones de nuestra DAG, se procedió a responder las queries planteadas en base a nuestro modelo.

Estos queries fueron resueltos con la función de cpquery, una función que estima la probabilidad condicional de un evento dada la evidencia. Para este evento fe utilizado una red Bayesiana y el método de Monte Carlo para generar 1,000,000 de simulaciones.

Query 1:

$$\mathbb{P}(S = 1 \mid C = 10) = 0.000333 \tag{1}$$

Donde $S = 1$ simboliza los valores que corresponden a hombres y $C = 10$ a que se haya cambiado de escuela en el ciclo escolar (2021-2022) por motivos de la pandemia COVID

Query 2:

$$\mathbb{P}(EX = 1|A = 1) > \mathbb{P}(P = 1|A = 1) \Rightarrow 0.7199356 \not> 0.8990911 \tag{2}$$

Donde $EX = 1$ simboliza que fueron utilizados los exámenes como método de evaluación durante la pandemia, mientras que por su parte $P = 1$ simboliza que los trabajos escritos sean utilizados. $A = 1$ simboliza el hecho que se haya utilizado la

asistencia como método de evaluación.

Query 3:

$$\mathbb{P}(E \geq 18, G = 3, NG = 2) = 0.005165 \tag{3}$$

Donde E es edad, G es nivel de escolaridad y NG es el número de grado.

Query 4:

$$\mathbb{P}(S = 2 \mid DB = 10 \cup DB = 11) > \mathbb{P}(S = 2 \mid DB = 2) \Rightarrow 0.461039 \not> 0.475043 \tag{4}$$

Donde $S = 2$ es sexo femenino, $DB = 10 \cup DB = 11$ es que se haya dejado de estudiar debido a un casamiento/embarazo/dedicarse a tareas del hogar y $DB = 2$ representa que la persona dejó de estudiar por trabajo.

IV. Conclusiones

Los datos nos mostraron diferentes hechos sobre la relación entre las variables en cuestión. De la primera pregunta concluimos que los datos muestran que el porcentaje de hombres que se dieron de baja principalmente por motivos de pandemia es de 0.033%.

Por otro lado, encontramos que dado el uso de la asistencia como método de evaluación, se usaron más los proyectos y tareas con un porcentaje de 89% que los exámenes (71%) como complementarios en la evaluación.

Asimismo, podemos ver que la probabilidad de que una persona mayor de 18 esté en segundo de secundaria es extremedamente baja, i.e. 0.5165%.

Además, es ligeramente más probable que una persona que dejó sus estudios por razones laborales (46.1%) o profesionales sea mujer a que una persona que haya dejado los estudios por razones de a un casamiento/embarazo/dedicarse a tareas del hogar sea mujer (47.5%).

Referencias

[1] I. N. de Estadística y Geografía (INEGI), "Encuesta nacional sobre acceso y permanencia en la educación (enape) 2021," https://www.inegi.org.mx/programas/enape/2021/, 2021, consultado el: 19 de agosto de 2024.

[2] M. José, "¿qué es un dag?" *bit2me ACADEMY*, 2020.

[3] R. P. Adhitama and D. R. Saputro, "Hill climbing algorithm for bayesian network structure," in *Proceedings of the AIP Conference*, vol. 2479, no. 1. Surakarta, Indonesia: AIP Publishing, July 2022, p. 020035, presented at the International Conference on Mathematics Research and Learning (ICOMER) 2021, August 11, 2021. [Online]. Available: https://doi.org/10.1063/5.0099793