Acceso y Permanencia en la Educación en México

Adara L. Pulido Sánchez[1] and Adamaris Leticia De Dios Ramos[2]

[1]Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias

Abstract

This study uses Bayesian networks to analyze the access and the educational retention in Mexico based on a database from INEGI. Bayesian network models help to examine the relationships between variables and make certain predictions. The reduction of the database and the modification of variables according to the catalog of each question are essential for the creation of DAGs. The Bayesian Information Criterion (BIC) is used to evaluate different DAGs and select the best option. The different queries provide results on the relationships between certain variables to better understand educational persistence and offer information on access and retention in education in Mexico.

Keywords: Grafos acíclicos dirigidos (DAGs), Redes bayesianas, Algoritmo hill-climbing, Criterio de Información Bayesiano (BIC), Análisis de dependencia, Modelización probabilística.

1 Introducción

La educación es un fenómeno que ha sido la base de todas las sociedades a lo largo de la historia, actuando como un puente que conecta a cada generación. A través de la educación se transmiten costumbres, valores y conocimientos indispensables para el desarrollo y permanencia de cada cultura. Este proceso también juega un papel crucial en la reducción de las desigualdades sociales y la promoción de la igualdad de género, empoderando a los individuos con las herramientas necesarias para alcanzar su máximo potencial. Asimismo, la educación fomenta la tolerancia entre las personas.

A pesar de los avances en el ámbito educativo, persisten problemas de acceso y permanencia. Por ello, este artículo examina la Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE) realizada por el INEGI en 2021, cuyo objetivo es generar información estadística sobre el acceso y permanencia de la población de 0 a 29 años en el Sistema Educativo Nacional.

2 Metodología

Esta metodología combina los enfoques descriptivos y matemáticos para analizar el Acceso y Permanencia en la Educación en México. Se sigue un proceso estructurado que abarca la construcción de variables, la modelización mediante grafos acíclicos dirigidos (DAGs), el uso de redes bayesianas, y la optimización de estas estructuras utilizando el algoritmo hill-climbing. A continuación, se detalla cada paso de manera integrada.

2.1 Definición y Construcción de Variables

El primer paso consiste en identificar y definir un conjunto de variables $X = \{X_1, X_2, \ldots, X_n\}$ relevantes para el estudio del acceso y la permanencia en la educación en México. Estas variables incluyen factores socioeconómicos, demográficos, y características institucionales que se asume influyen en los resultados educativos.

Cada variable X_i se selecciona con base en una revisión teórica y se construye a partir de los datos disponibles. Estas variables forman la base para la creación de los modelos probabilísticos y las relaciones de dependencia que serán analizadas.

2.2 Propuesta y Construcción de Diferentes DAGs

Un **Grafo Acíclico Dirigido (DAG)**, denotado como $G = (V, E)$, es una estructura que representa las relaciones de dependencia entre las variables de interés. Se proponen tres diferentes DAGs, donde:

- V es el conjunto de variables X que representan los nodos del grafo.
- E es el conjunto de aristas dirigidas que indican una relación de influencia directa entre las variables.

Cada DAG refleja una hipótesis distinta sobre las interdependencias entre las variables. La ausencia de ciclos en el DAG asegura que no existan bucles de retroalimentación, manteniendo una estructura causal clara.

2.3 Construcción de Redes Bayesianas

Con base en los DAGs propuestos, se construyen redes bayesianas para modelar la distribución conjunta de las variables X. La red bayesiana se define matemáticamente como:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{pa}(X_i))$$

donde $\text{pa}(X_i)$ son los padres de X_i en el DAG. Cada red bayesiana permite evaluar cómo las probabilidades condicionales de las variables se relacionan entre sí, proporcionando una forma estructurada de inferencia sobre el conjunto de datos.

2.4 Análisis de la Significancia de las Relaciones de Dependencia

Una vez construidas las redes bayesianas, se realiza un análisis de la significancia de las relaciones de dependencia en los DAGs propuestos. Esto implica calcular la **información mutua** $I(X_i; X_j)$ entre variables conectadas:

$$I(X_i; X_j) = \sum_{x_i} \sum_{x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

Este análisis permite identificar qué relaciones tienen una influencia significativa y cuáles podrían ser descartadas o reevaluadas en función de su baja dependencia.

2.5 Selección de la Mejor Red

Para determinar la red bayesiana que mejor se ajusta a los datos, se evalúan las estructuras propuestas mediante la **verosimilitud** de los datos. La verosimilitud es una medida de cuán bien un modelo probabilístico G explica los datos observados D. Matemáticamente, se define como:

$$\mathcal{L}(G \mid D) = \prod_{j=1}^{m} \prod_{i=1}^{n} P(x_i^{(j)} \mid \text{pa}(x_i^{(j)}))$$

donde $P(x_i^{(j)} \mid \text{pa}(x_i^{(j)})$ es la probabilidad condicional de la j-ésima observación de la variable X_i, dado el valor de sus padres en el DAG. La verosimilitud refleja la probabilidad de observar los datos D bajo el modelo G. Un modelo con una mayor verosimilitud es considerado mejor, ya que sugiere que es más probable que los datos observados provengan de dicho modelo.

Además, se utiliza el **Criterio de Información Bayesiano (BIC)** para evaluar y comparar las estructuras de DAGs:

$$\text{BIC} = k \log n - 2 \log \mathcal{L}(G \mid D)$$

donde k es el número de parámetros libres en el modelo, y n es el tamaño de la muestra. El DAG que maximice la verosimilitud o minimice el BIC es seleccionado como el modelo que mejor representa las relaciones de dependencia en los datos. También se utilizan criterios de ajuste como el **Criterio de Información Bayesiano (BIC)**:

$$\text{BIC} = k \log n - 2 \log \mathcal{L}(G \mid D)$$

El DAG que maximice la verosimilitud o minimice el BIC es seleccionado como el modelo que mejor representa las relaciones de dependencia en los datos.

2.6 Optimización mediante el Algoritmo Hill-Climbing

El **algoritmo hill-climbing** se emplea para optimizar la estructura del DAG. Inicia con un DAG G_0 y, en cada iteración, se realiza una modificación (como agregar, eliminar o invertir una arista) para obtener un nuevo DAG G_{t+1}:

$$G_{t+1} = \arg\max_{G'} S(G')$$

donde $S(G')$ es la función de puntuación, generalmente basada en la log-verosimilitud. Este proceso se repite hasta que se alcanza un máximo local en la función de puntuación, proporcionando la estructura de DAG más adecuada para los datos.

2.7 Discusión de la Pertinencia de la DAG Optimizada

La pertinencia de la DAG optimizada se discute en términos de la consistencia y relevancia de las relaciones de dependencia identificadas. Se analiza cómo esta estructura refleja las dinámicas subyacentes en el acceso y permanencia en la educación en México, y se evalúan las implicaciones para la política educativa y futuras investigaciones.

2.8 Respuesta a las Queries Asignadas

Finalmente, se utilizan las redes bayesianas optimizadas para responder a las queries asignadas. Las respuestas se basan en inferencias probabilísticas obtenidas de las redes, proporcionando una comprensión detallada y cuantitativa de los factores que afectan el acceso y la permanencia en la educación en México.

3 Aplicación

Para la realización de la encuesta se contó con un grupo de 101 preguntas. La base de datos utilizada fue *conjunto_de_datos_tmodulo_*2021, contiene información de la Encuesta Nacional sobre Acceso y Permanencia en la Educación (ENAPE). La base de datos contiene las preguntas de la encuesta, las preguntas en la encuesta se encuentran separadas por ciertas categorías: I Residentes y carácterísticas de la viviendas, II Identificación de personas de 0 a 9 años, III Personas de 0 a 29 años, $APARTADOA$ Inscripción en el ciclo escolar 2020-2021, $APARTADOB$ Inscripción en el ciclo escolar 2021-2022, $APARTADOC$ Población no inscrita en el ciclo escolar 2020-2021 ni en el ciclo escolar 2021-2022, $APARTADOD$ Participación económica y consecuencia de dejar de trabajar y $SECCIONIV$ Opinión sobre el valor de la educación.

Se realizó una limpieza de la base de datos, se tomaron las variables necesarias para los queries, para así poder hacer una reducción de la base de datos. Con esto, se realizó un cambio de variables en función al catálogo de cada pregunta y conseguir un buen manejo de esta para las DAGs. En el presente trabajo solo se utilizaron 12 preguntas que se tomarán como variables.

- Pregunta PA3.1 ¿(NOMBRE) estuvo inscrita(o) el pasado año o ciclo escolar (2020-2021)?
- Pregunta PA3.2 ¿La escuela donde (NOMBRE) estuvo inscrita(o) (asistió) el pasado ciclo escolar fue...?
- Pregunta PA3.4 ¿(NOMBRE) concluyó el grado (semestre, cuatrimestre o módulo) en que estuvo inscrita(o) el pasado año o ciclo escolar (2020-2021)?
- Pregunta PA3.5 ¿Cuál fue la razón principal por la que (NOMBRE) no concluyó el grado (semestre, cuatrimestre o módulo)?
- Pregunta PB3.1 ¿(NOMBRE) está inscrita(o) en el actual año o ciclo escolar (2021-2022) que inició en agosto/septiembre de 2021?
- Pregunta PB3.3 ¿La escuela donde (NOMBRE) está inscrita(o) es...? Pregunta PB3.4 ¿Cuál fue la razón principal por la que (NOMBRE) se cambió de escuela en el actual año o ciclo escolar (2021-2022)?
- Pregunta PB3.5.NIVEL Actualmente, ¿en qué grado (semestre, cuatrimestre o módulo) escolar está inscrita(o) (NOMBRE)?
- Pregunta PB3.11.4 ¿Qué medio o medios utiliza(n) la(s) maestra(s) o maestro(s) de (NOMBRE), para informarle sobre las actividades escolares, impartir sus clases o para la entrega de trabajos? - Lo hacen de manera presencial
- Pregunta PC3.3.1 ¿(NOMBRE) hasta qué año y grado aprobó en la escuela? - NIVEL

- Pregunta PC3.6 ¿Cuál es la razón principal por la que (NOMBRE) dejó de asistir o interrumpió la escuela?

Para practicidad se renombraron las variables de la siguiente forma

- PA3.1: IPre
- PA3.2: EP
- PA3.4: GCP
- PA3.5: RGNC
- PB3.1: IA
- PB3.3: EA
- PB3.4: RC
- PB3.5: N
- PB3.11.4: P
- PC3.3.1: NM
- PC3.6: RDE

Estas variables se tomaron debido a que tiene relación con la respuesta de un banco de preguntas o queries que buscan obtener información de la base de datos.

- ¿Es más probable que una persona se cambie de escuela debido a un cambio de residencia, al cierre de la escuela o en busca de una mejor calidad educativa?
- ¿Es más probable que la persona de sexo femenino no haya concluido el grado debido a la pandemia COVID-19 o por trabajo ?
- ¿Es más probable que los estudiantes de niveles educativos más bajos estudien en modalidad presencial en comparación con aquellos en niveles educativos superiores?
- ¿Es más probable que una persona no haya concluido la secundaria dado que haya tenido que entrar a trabajar o haya tenido problemas personales?

Retomando lo dicho en la metodología, con el fin de representar las relaciones de independencia entre las variables utilizadas, se propusieron tres diferentes DAGs.

3.1 Primer Grafo Acíclico Dirigido

En este primer grafo se partió de la idea de que todas las variables dependen del "SEXO", así mismo el nivel máximo de estudios (NM) depende directamente de si la persona estuvo inscrita en el curso previo (IPre) y en el actual (IA), para ser nodo padre de la razón por la cual dejo de estudiar(RDE).

Por otro lado la razon de cambio (RC) depende de la escuela pasada (EP) y de la actual (EA), la razón por la que una persona no concluyó el grado (RGNC) depende de si la persona concluyó el mismo(GCP). Por último, la modalidad presencial (P) de las clases depende del nivel de estudios actuales.

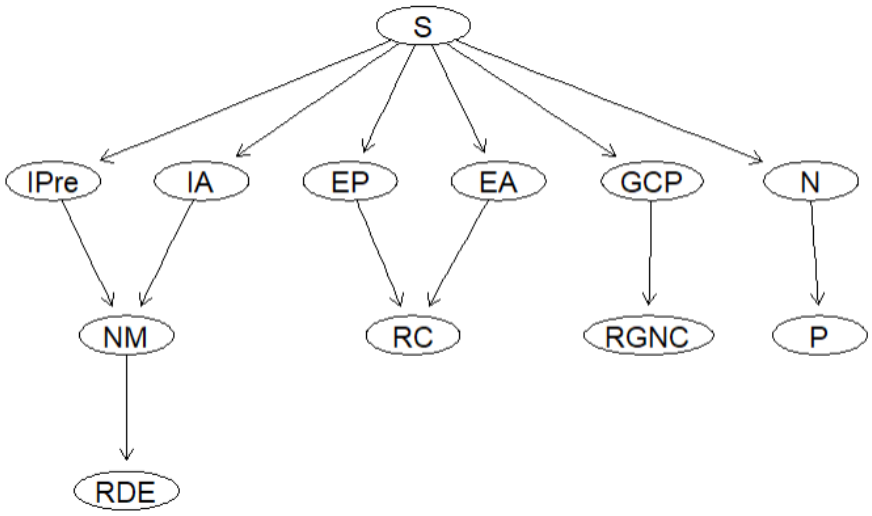

Figure 1: Primer Grafo Acíclico Dirigido

Con el objetivo de conocer la significancia de las relaciones de dependencia de esta Primera DAG, se calculan la fuerza de cada uno de los arcos de la red bayesiana.

	from <chr>	to <chr>	strength <dbl>
1	S	EP	0.202636158
2	S	EA	0.031948290
3	S	GCP	0.006726542
4	S	IPre	0.197028827
5	S	IA	0.051264436
6	S	N	0.023775696
7	EP	RC	0.000000000
8	EA	RC	0.000000000
9	GCP	RGNC	0.000000000
10	IPre	NM	0.000000000

Figure 2: Primer DAG - Significancias

Como se puede observar en la tabla, los arcos con mayor significancia son EP a RC, GCP a RGNC, IPre a NM, S a N y S a EA. Sin embargo para conocer el puntaje de esta red

bayesiana se utiliza el Criterio de información bayesiano el cual resulta con un valor de -257507.2. De igual manera, el Criterio de información de Akaike deriva un puntaje de -255516

3.2 Segundo Grafo Acíclico Dirigido

Por el contrario, en esta nueva DAG, se plantea la posible dependencia de la escuela del ciclo pasado (EP) y la del actual (EA), con la razón de cambio (RC). De igual forma, si la persona concluyó el mismo (GCP) depende de la razón por la que una persona no concluyó el grado (RGNC).

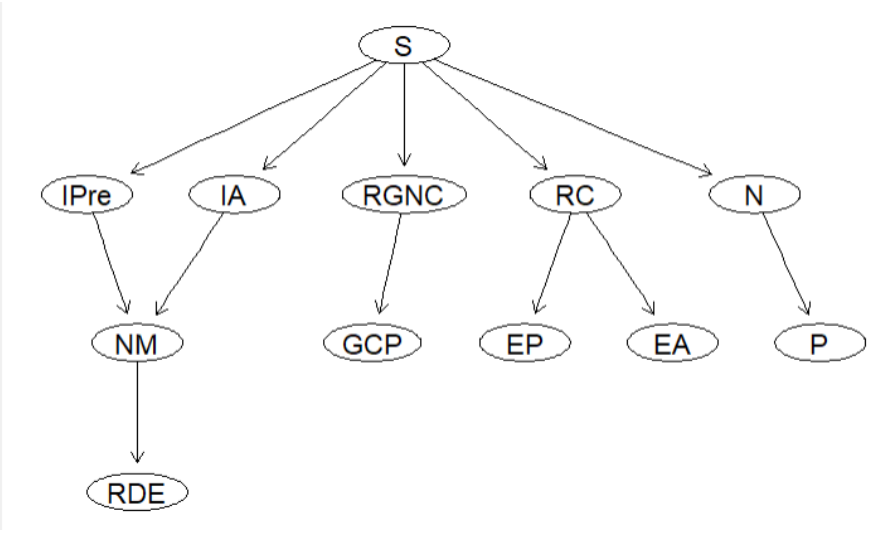

Figure 3: Segundo Grafo Acíclico Dirigido

Se calculan nuevamente la fuerza de cada uno de los arcos de la red bayesiana para conocer la significancia de las relaciones de dependencia de esta segunda DAG.

	from <chr>	to <chr>	strength <dbl>
1	S	RC	5.117963e-01
2	S	RGNC	9.130484e-04
3	S	IPre	1.970288e-01
4	S	IA	5.126444e-02
5	S	N	2.377570e-02
6	RC	EP	9.441633e-139
7	RC	EA	1.309014e-162
8	RGNC	GCP	0.000000e+00
9	IPre	NM	0.000000e+00
10	IA	NM	0.000000e+00

Figure 4: Segunda DAG - Significancias

De acuerdo a lo observado en la tabla, se entiende que todos los arcos tienen una dependencia significativa entre sus nodos padres e hijos. No obstante se vuelven a realizar el criterio de

información bayesiano resultando −258466.5 y el el Criterio de información de Akaike con un puntaje de −256580.1.

3.3 Tercer Grafo Acíclico Dirigido

En este tercer grafo, se decidió volver a la idea de partir de la variable "SEXO" como padre de las demás variables, sin embargo, para este caso decidimos proponer a la variable de Razón por la que se dejó la escuela, "RDE", como el nodo padre de las variables IPre e IA.

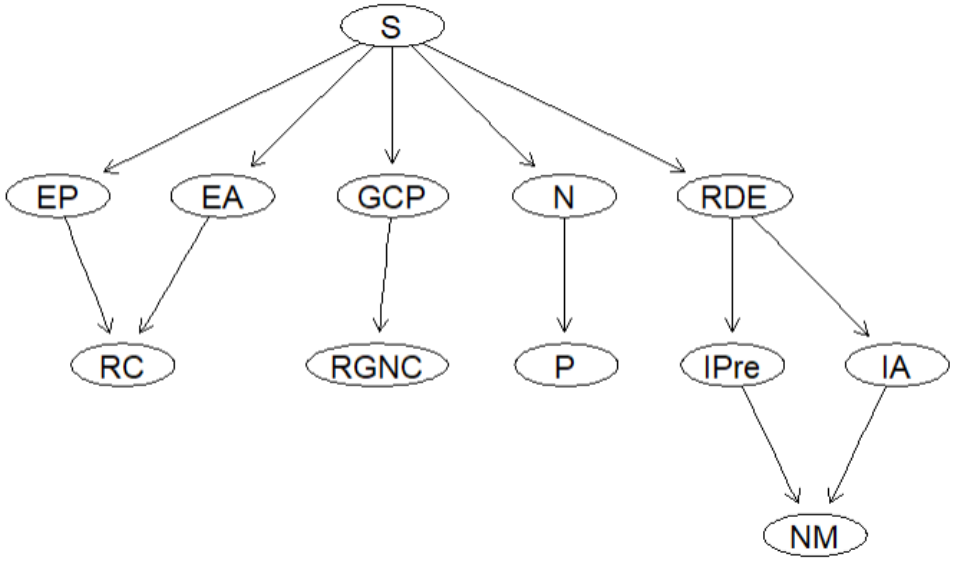

Figure 5: Tercer Grafo Acíclico Dirigido

Con el objetivo de conocer la significancia de las relaciones de dependencia de esta Tercera DAG, se calculan la fuerza de cada uno de los arcos de la red bayesiana.

	from <chr>	to <chr>	strength <dbl>
1	S	EP	2.026362e-01
2	S	EA	3.194829e-02
3	S	GCP	6.726542e-03
4	S	RDE	1.093235e-127
5	S	N	2.377570e-02
6	EP	RC	0.000000e+00
7	EA	RC	0.000000e+00
8	GCP	RGNC	0.000000e+00
9	IPre	NM	0.000000e+00
10	IA	NM	0.000000e+00

Figure 6: Tercer DAG - Significancias

Haciendo una revisión de los resultados obtenidos en la tabla de significancia, podemos observar, una vez más, que todos los arcos tienen una dependencia significativa entre sus nodos padres e hijos. Además de que, nuevamente, realizado el criterio de información

bayesiano resultando −252319.5 y el el Criterio de información de Akaike con un puntaje de −250697.2.

3.4 Mejor estructura del Grafo Acíclico Dirigido

Con la finalidad de optimizar la estructura de la DAG para esta base de datos se utiliza el algoritmo Hill-climbing obteniendo lo siguiente como mejor estructura.

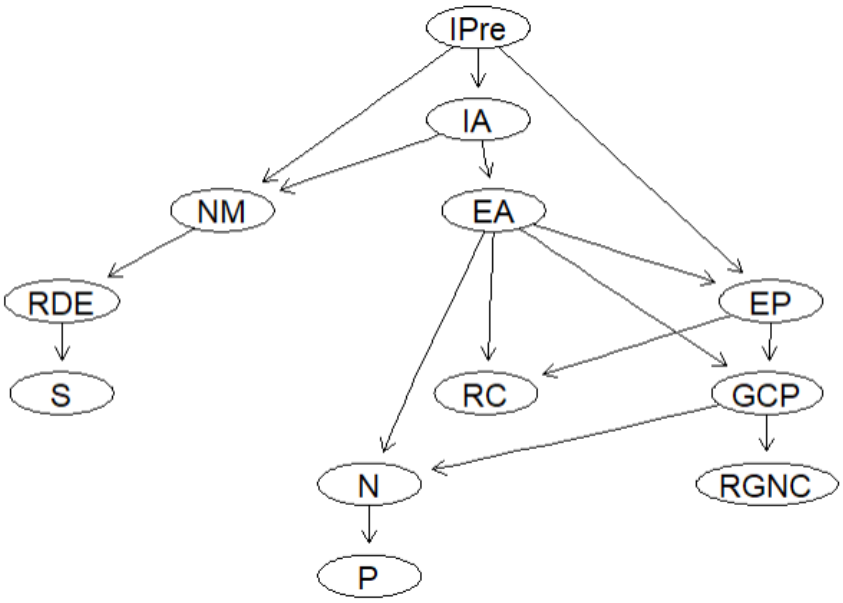

Figure 7: Mejor Grafo Acíclico Dirigido

En este caso observamos como la variable "IPRE" toma el lugar de la variable padre, así como también, la variable "IA" pasa a ser un nodo padre para las variables "NM", "EA" y "EP", dejando a "S", "RC", "P" y "RGNC" como las variables más codependientes de la estrucura.

Al observar sus relaciones las relaciones de dependencia entre sus nodos se analiza que todos estan en cero, lo que quiere decir que cada una de las relaciones establecidas tienen una alta significancia.

3.5 Queries

A partir de las siguientes preguntas:

- ¿Es más probable que una persona se cambie de escuela debido a un cambio de residencia, al cierre de la escuela o en busca de una mejor calidad educativa?

	from <chr>	to <chr>	strength <dbl>
1	EA	N	0.000000e+00
2	N	P	0.000000e+00
3	IA	EA	0.000000e+00
4	IPre	EP	0.000000e+00
5	EP	GCP	0.000000e+00
6	NM	RDE	0.000000e+00
7	IPre	IA	0.000000e+00
8	IPre	NM	0.000000e+00
9	EA	EP	0.000000e+00
10	IA	NM	0.000000e+00

Figure 8: Mejor DAG - Significancias

	Calidad	cierre escuela	costo
32020	87	9	53
Covid-19	Edad	falta de espacio	gusto
20	5	25	5
motivos personales	organizacion	Residencia	universidad
93	13	9	4

Figure 9: Querie 1

Al observar la tabla de las razones de cambio se deduce que es más probable que alguien se cambie de escuela por la calidad educativa

- ¿Es más probable que la persona de sexo femenino no haya concluido el grado debido a la pandemia COVID-19 o por trabajo ?

```
, , = Covid-19

   No Si
F  0  0  0
M  0  1  0
```

Figure 10: Querie 2

Gracias a la tabla y a la función cpquery se puede observar que es más probable que

		No	Si
F	0	34	0
M	0	54	0

, , = trabajo

Figure 11: Querie 2

no haya concluido el grado debido al trabajo

- ¿Es más probable que los estudiantes de niveles educativos más bajos estudien en modalidad presencial en comparación con aquellos en niveles educativos superiores?

		No se declaro	Si
	12688	0	0
bachillerato tecnologico	0	538	263
doctorado	0	7	3
especialidad	0	22	7
maestria	0	76	18
Preparatoria	4	1655	889
Prescolar	4	1106	931
Primaria	4	3820	3105
profesional	0	2617	803
Profesional tecnico	0	23	20
Secundaria	4	2047	1591
tecnico sup universitario	0	66	32

Figure 12: Querie 3

En esta tercer pregunta si se declara como nulos los no declarados se puede rechazar que es más probable que los niveles más bajos estudien de manera presencial que de niveles superiores

- ¿Es más probable que una persona haya concluido la secundaria dado que haya tenido que entrar a trabajar o haya tenido problemas personales?

	otro	problemas personales	reprobado	trabajo
	0	0	0	0
bachillerato tecnologico	17	12	8	119
especialidad	0	0	0	1
maestria	0	0	0	12
Ninguno	2	0	0	1
Preparatoria	47	35	32	532
Prescolar	6	1	0	0
Primaria	13	23	5	48
profesional	33	10	12	438
Profesional tecnico	2	0	0	28
Secundaria	31	43	25	289
tecnico sup universitario	2	1	0	41

Figure 13: Querie 4

Con la útlima tabla se puede apreciar que es mucho mnás probable que una persona no haya concluido sus estudios por entrar a trabajar que por problemas personales

4 Conclusiones

Conclusión

Desde el punto de vista metodológico, el uso de redes bayesianas demostró ser una herramienta efectiva para modelar y entender las complejas interrelaciones entre las diversas variables educativas consideradas. La aplicación del algoritmo de *hill-climbing* permitió una construcción eficiente de la estructura de la red, identificando de manera óptima las dependencias más significativas presentes en los datos. Este enfoque facilitó la detección de patrones y relaciones que podrían no ser evidentes mediante métodos estadísticos tradicionales.

En esta investigación, se ha desarrollado un modelo de red bayesiana para analizar los factores que influyen en el acceso y la permanencia en la educación en México. Los resultados obtenidos revelan que la calidad educativa es la principal razón por la que los estudiantes deciden cambiar de escuela, destacando la importancia de mejorar las condiciones académicas para fomentar la estabilidad educativa. Además, se observó que las mujeres tienen una mayor probabilidad de no concluir su grado escolar debido a la necesidad de incorporarse al mercado laboral, lo que subraya la necesidad de implementar políticas que apoyen la continuidad educativa femenina. También se identificó que los niveles educativos más bajos tienen menos acceso a clases presenciales comparados con niveles superiores, evidenciando desigualdades que deben ser abordadas para garantizar una educación equitativa.

El enfoque adoptado no solo permitió manejar de manera eficaz la incertidumbre inherente a los datos educativos, sino que también ofreció una flexibilidad notable para incorporar diferentes tipos de variables y relaciones. Esto es especialmente relevante en contextos educativos, donde los factores influyentes suelen ser múltiples y están interconectados de maneras complejas.

Sin embargo, es importante reconocer que los resultados y conclusiones están sujetos a las limitaciones de los datos utilizados, incluyendo posibles sesgos y la calidad de la información recolectada. Futuras investigaciones podrían ampliar este estudio incorporando más variables y utilizando conjuntos de datos más extensos y diversos para validar y desarollar de mejor manera los hallazgos actuales.

A Apéndice

`https://github.com/LAK3SHORE/Bayesian-Networks-Study-Case`

[1] [2] [3]

References

[1] R. G. Cowell et al. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, 2007.

[2] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[3] M. Mora-Olate. "Educación como disciplina y como objeto de estudio: aportes para un debate". In: *Desde el Sur* 12.1 (2020), pp. 201–211.