Artículo 1 - Redes Bayesianas multinomiales

Barush Caliel C. Luna Diego B. Castillo
Pedro Emilio S. Rodríguez Vidal Alejandro G. López

2025-08-31

Abstract

This study explores the application of Bayesian networks to model urban mobility patterns using large-scale transport survey data. We evaluate multiple structural learning approaches, including expert-defined DAGs and constrained algorithmic optimization using Hill Climbing methods. Our methodology demonstrates how domain knowledge can be integrated with machine learning techniques to create interpretable yet statistically robust models. The research highlights both the capabilities and limitations of data-driven transport modeling, particularly regarding geographical representation in official surveys. The findings contribute to methodological discussions in transportation research and offer insights into balancing statistical optimization with theoretical coherence in complex systems modeling.

Table of contents

1 Introducción

Este artículo presenta un análisis de movilidad urbana mediante el uso de redes bayesianas (DAGs) aplicadas a datos de la Encuesta Origen-Destino 2017 del INEGI. La investigación se centra en la Zona Metropolitana del Valle de México y combina métodos de aprendizaje automático con conocimiento sustantivo para modelar las complejas relaciones entre variables demográficas, características de viaje y elección de modos de transporte.

2 Metodología

La metodología consistió en la construcción y análisis de DAGs utilizando RStudio como herramienta, así como el paquete de *bnlearn*. Inicialmente se propusieron DAGs basadas en el criterio individual de los autores, para luego ser optimizadas por medio del algoritmo de Hill Climbing, tanto con restricciones estructurales como sin ellas, buscando maximizar la distribución de la red de los datos. Para descartar DAGs, se aplicaron pruebas de hipótesis y se evaluó la fuerza de los arcos con la función *arc.strength*(). Finalmente, se emplearon queries en R para extraer relaciones clave y realizar observaciones relevantes, integrando conocimiento sobre el tema, optimización y análisis estadístico para construir modelos explicativos basados en DAGs.

3 Resultados (aplicación)

3.1 Base de datos

Posterior a la selección de variables según el contenido de los queries, se seleccionaron las variables y unieron en un solo dataset, el cual servió como la

base para construir las DAG's, analizar las dependencias y realizar las cuatro consultas asignadas. Se muestra a continuación un fragmento de la misma.

```
head_df = readRDS("objetos/head_base_final.rds")
head_df
```

```
  id_soc p5_3 p5_6 p5_7_7 p5_14_01 p5_14_07 p5_14_12 p5_14_14 p5_14_19 p5_14_20
1   1268    1   01     09         2         2         2         1         2         2
2   1268    1   03     09         2         2         2         1         2         2
3   1268    2   01     09         2         2         2         1         2         2
4   1268    2   07     09         2         2         2         2         2         2
5   1269    1   01     09         2         2         2         1         2         2
6   1269    1   03     09         2         2         2         1         2         2
  sexo edad ent niv
1    2   26   9   7
2    2   26   9   7
3    2   26   9   7
4    2   26   9   7
5    2   22   9   6
6    2   22   9   6
```

3.2 Diagramas de DAG's utilizadas y analizadas

3.2.0.0.1 DAG propuesta 1

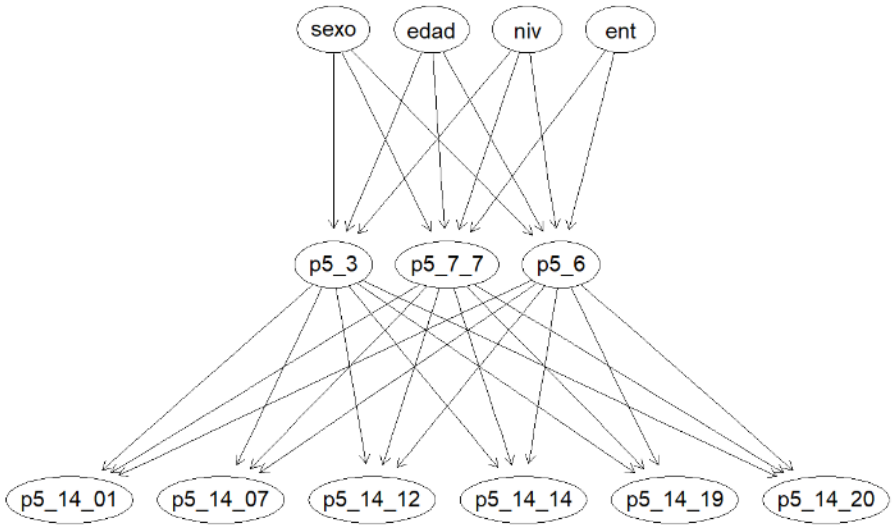

Figure 1: DAG 1

3.2.0.0.2 DAG propuesta 2

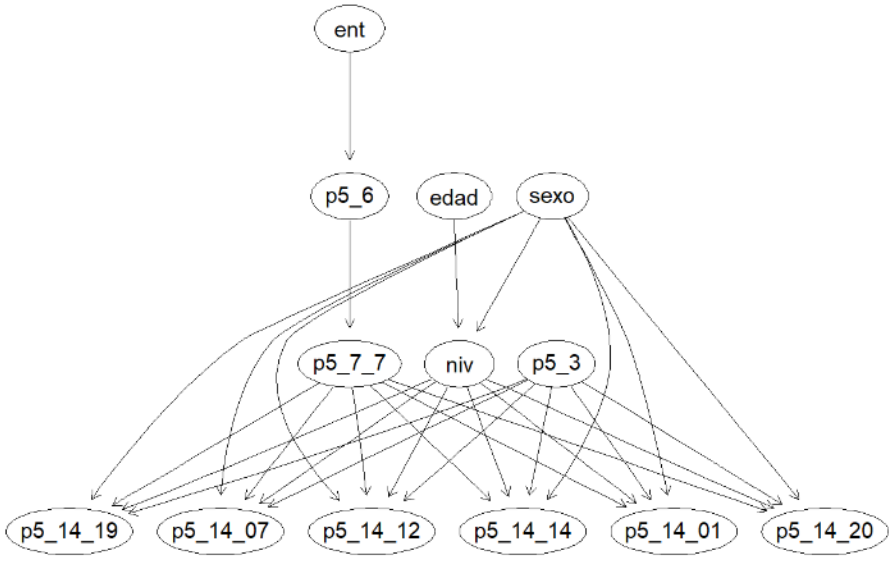

Figure 2: DAG 2

3.2.0.0.3 DAG propuesta 3

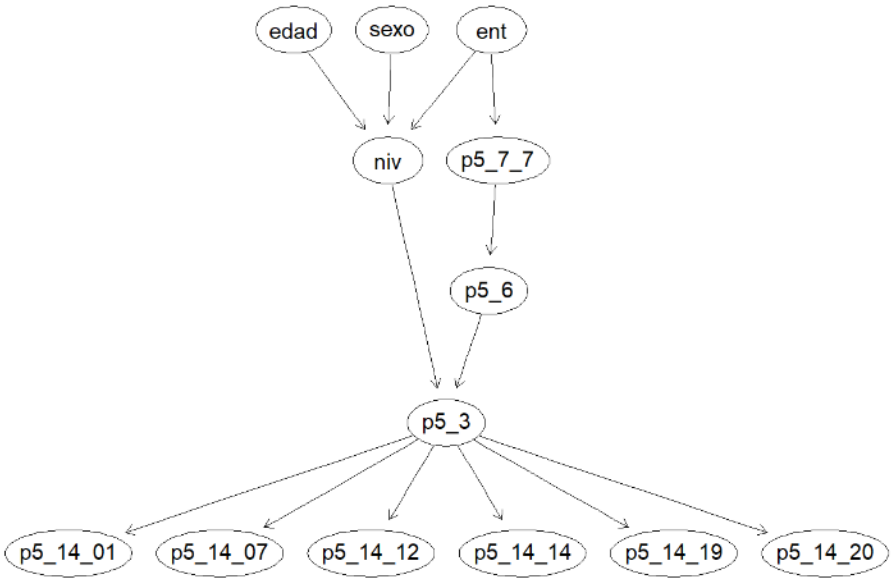

Figure 3: DAG 3

3.2.0.0.4 DAG con HC sin restricciones

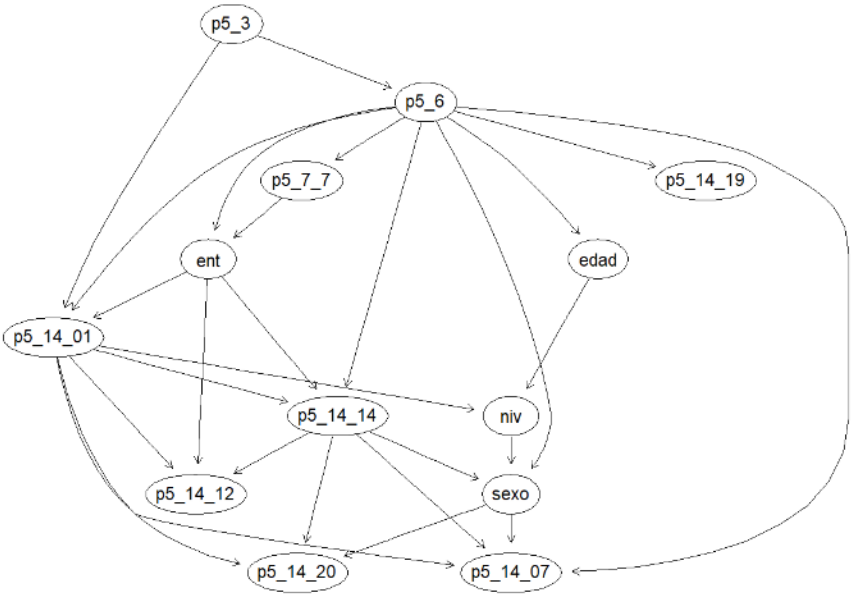

Figure 4: DAG con HC sin restricciones

3.2.0.0.5 DAG HC con restricciones número 0

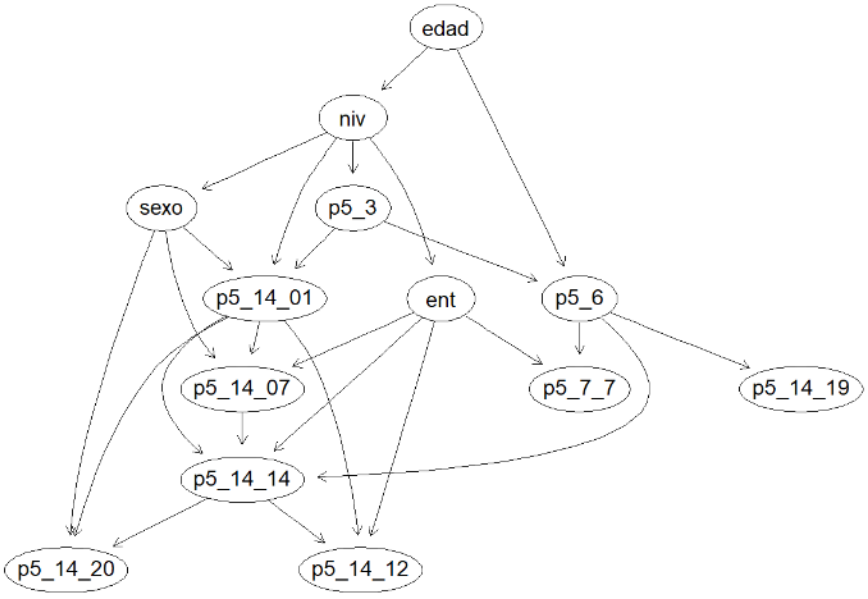

Figure 5: DAG HC con restricciones 0

3.2.0.0.6 DAG HC con restricciones final

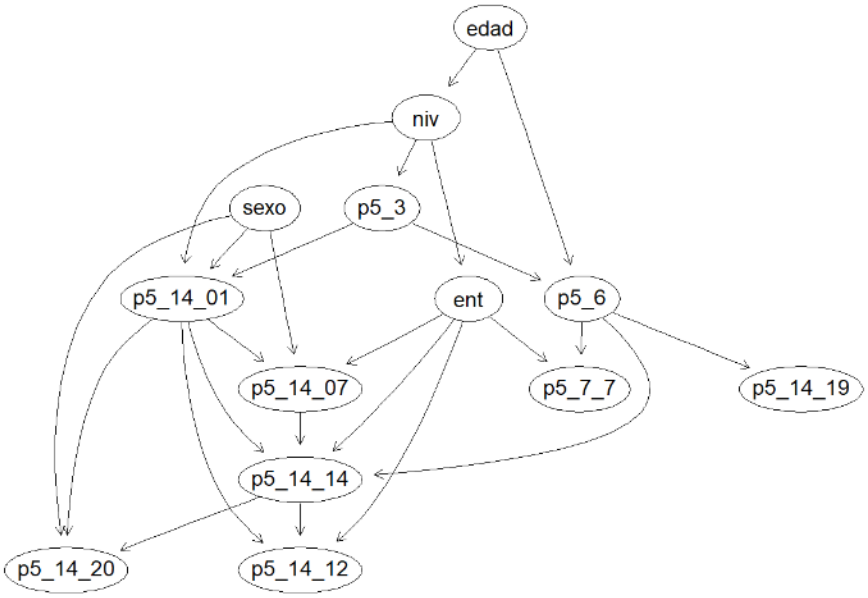

Figure 6: DAG HC con restricciones final

Tabla 1. *BIC y AIC de las DAG propuestas y con el algoritmo HC.*

DAG	BIC	AIC
DAG 1	-7,479,102	-6,150,732
DAG 2	-6,089,081	-6,041,948
DAG 3	-5,819,506	-5,784,574
DAG HC sin restricciones	-5,586,194	-5,556,406
DAG HC con restricciones	-5,603,767	-5,573,527

3.3 Resultados de prueba arc.strength()

Una vez seleccionadas la DAG 3, la DAG con HC sin restricciones y la DAG HC con restricciones, se les hicieron una prueba por medio del método arc.strength, el cual busca medir la fuerza de la relación probabilística entre dos nodos conectados por un arco. Los resultados fueron estos:

3.3.0.0.1 DAG 3 propuesta

```
prueba_strength_dag3 = readRDS("objetos/prueba_strength_dag3.rds")
prueba_strength_dag3
```

```
     from       to     strength
1    sexo      niv 0.000000e+00
2    edad      niv 0.000000e+00
3     ent      niv 0.000000e+00
4     ent   p5_7_7 0.000000e+00
5  p5_7_7     p5_6 0.000000e+00
6     niv     p5_3 0.000000e+00
7    p5_6     p5_3 0.000000e+00
8    p5_3 p5_14_01 0.000000e+00
9    p5_3 p5_14_07 1.841923e-07
10   p5_3 p5_14_12 2.608496e-02
11   p5_3 p5_14_14 0.000000e+00
12   p5_3 p5_14_19 8.300152e-03
13   p5_3 p5_14_20 1.116759e-02
```

3.3.0.0.2 DAG con HC sin restricciones

```
prueba_strength_best_dag = readRDS("objetos/prueba_strength_best_dag.rds")
prueba_strength_best_dag
```

```
       from       to      strength
1    p5_7_7      ent  0.000000e+00
2      edad      niv  0.000000e+00
3  p5_14_01 p5_14_14  0.000000e+00
4      p5_3     p5_6  0.000000e+00
5      p5_6     edad  0.000000e+00
6      p5_6      ent  0.000000e+00
7      p5_6     sexo  0.000000e+00
8      p5_6   p5_7_7  0.000000e+00
9  p5_14_14     sexo  0.000000e+00
10 p5_14_01 p5_14_07  0.000000e+00
11     p5_6 p5_14_01  0.000000e+00
12      niv     sexo  0.000000e+00
13     p5_6 p5_14_14  0.000000e+00
14     p5_6 p5_14_07 7.295157e-303
15      ent p5_14_12 8.026779e-269
16     p5_3 p5_14_01  0.000000e+00
17 p5_14_01      niv  0.000000e+00
18      ent p5_14_01  0.000000e+00
19      ent p5_14_14  0.000000e+00
```

```
20 p5_14_14 p5_14_07  0.000000e+00
21     sexo p5_14_07  0.000000e+00
22 p5_14_01 p5_14_12  5.334871e-92
23 p5_14_14 p5_14_20 4.782076e-176
24 p5_14_01 p5_14_20 1.488990e-156
25     p5_6 p5_14_19  1.095026e-85
26     sexo p5_14_20  2.762316e-46
27 p5_14_14 p5_14_12  7.275552e-18
```

3.3.0.0.3 DAG HC con restricciones

```
prueba_strength_restricted = readRDS("objetos/prueba_strength_restricted.rds")
prueba_strength_restricted
```

```
       from       to      strength
1       ent   p5_7_7  0.000000e+00
2      edad      niv  0.000000e+00
3  p5_14_01 p5_14_14  0.000000e+00
4      p5_3     p5_6  0.000000e+00
5      p5_6   p5_7_7  0.000000e+00
6      edad     p5_6  0.000000e+00
7       niv p5_14_01  0.000000e+00
8  p5_14_07 p5_14_14  0.000000e+00
9       niv      ent  0.000000e+00
10 p5_14_01 p5_14_07  0.000000e+00
11     sexo p5_14_01  0.000000e+00
12     sexo p5_14_07  0.000000e+00
13     p5_3 p5_14_01  0.000000e+00
14     p5_6 p5_14_14  0.000000e+00
15      ent p5_14_12 8.026779e-269
16      niv     p5_3 1.386580e-205
17      ent p5_14_07 2.246968e-199
18      ent p5_14_14  0.000000e+00
19 p5_14_01 p5_14_12  5.334871e-92
20 p5_14_14 p5_14_20 4.782076e-176
21 p5_14_01 p5_14_20 1.488990e-156
22     p5_6 p5_14_19  1.095026e-85
23     sexo p5_14_20  2.762316e-46
24 p5_14_14 p5_14_12  7.275552e-18
```

3.4 Prueba de p-value sobre DAG con HC

Resultado del análisis de los valores del BIC y AIC, así como de la prueba arc.strength, se seleccionó la DAG HC con restricciones para analizar si sus

relaciones de dependencia son significativas:

```
resultados_p_hipotesis = readRDS("objetos/resultados_p_hipótesis.rds")
resultados_p_hipotesis
```

```
$ent_to_p5_7_7

    Mutual Information (disc.)

data:  ent ~ p5_7_7 | p5_6
mi = 564118, df = 748, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$edad_to_niv

    Mutual Information (disc.)

data:  edad ~ niv
mi = 380935, df = 910, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$p5_14_01_to_p5_14_14

    Mutual Information (disc.)

data:  p5_14_01 ~ p5_14_14 | p5_6 + p5_14_07 + ent
mi = 221628, df = 102, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$p5_3_to_p5_6

    Mutual Information (disc.)

data:  p5_3 ~ p5_6 | edad
mi = 68019, df = 1472, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$p5_6_to_p5_7_7

    Mutual Information (disc.)
```

```
data:  p5_6 ~ p5_7_7 | ent
mi = 74686, df = 1056, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$edad_to_p5_6

    Mutual Information (disc.)

data:  edad ~ p5_6 | p5_3
mi = 85171, df = 2912, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$niv_to_p5_14_01

    Mutual Information (disc.)

data:  niv ~ p5_14_01 | p5_3 + sexo
mi = 41455, df = 40, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$p5_14_07_to_p5_14_14

    Mutual Information (disc.)

data:  p5_14_07 ~ p5_14_14 | p5_6 + p5_14_01 + ent
mi = 31580, df = 102, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$niv_to_ent

    Mutual Information (disc.)

data:  niv ~ ent
mi = 13997, df = 20, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0


$p5_14_01_to_p5_14_07

    Mutual Information (disc.)

data:  p5_14_01 ~ p5_14_07 | sexo + ent
```

```
mi = 5912, df = 6, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0
```

3.5 Probabilidad de las queries

3.5.0.1 Primer Query

Probabilidad de que una persona haya salido de un hospital en Yucatán haya utilizado transporte de personal o automóvil como medio de transporte 0

3.5.0.2 Segundo Query

Probabilidad de que una persona de 20 años y de sexo femenino use el tren ligero como medio de transporte dado que es de CDMX 0.008258929

3.5.0.3 Tercera Query

Probabilidad de que un hombre joven universitario en Jalisco utilice tren ligero durante los días entre semana 0

Este resultado se debe a la baja cantidad de datos registrados de viajes dentro de Jalisco (solamente 13 registros), lo que genera una representatividad insuficiente para capturar esta combinación específica de características.

3.5.0.4 Cuarto Query

Probabilidad de que un hombre use un vehículo no motorizado dado que el origen es desde su hogar; escuela; comercio, mercado, tienda o centro comercial; restaurante, bar, cafetería; deportivo, gimnasio; o vía pública 0.2962361

4 Análisis de los resultados

4.1 Selección de DAG:

La tabla 1 compara los estadísticos BIC y AIC de las cinco DAG's propuestas. Por las características del BIC y AIC de ser estadísticos relativos y que bajo el funcionamiento de la librería *bnlearn*, un valor mayor denota un mejor modelo, las DAG's que destacan son la DAG 3, DAG con HC sin restricciones y DAG HC con restricciones. No obstante, La DAG 3, al ser seleccionada por hipótesis humanas, pueden presentar un sesgo o no capturar la complejidad de los datos. Para ello, se depuró de cinco DAG's a solamente dos: DAG con HC sin restricciones y DAG HC con restricciones.

Gracias al algoritmo hill-climbing, la versión de estas DAG's son óptimas al menos en un máximo local (no siempre global). La DAG sin restricciones tiene

el mayor valor, no obstante, presenta falta deinterpretabilidad con arcos y nodos que no forman sentido en la vida real. Un ejemplo de ello es que, según la DAG con HC sin restricciones (BIC $= -5,586,194$ y un \$AIC = -5,556,406 \$), el sexo depende del nivel educativo o que la edad depende del origen del viaje. Por estas observaciones, es que se escogió la DAG HC con restrcciones, la cual presenta mayor interpretabilidad y está en un óptimo local con un BIC $= -5,603,767$ y un AIC $= -5,573,527$

4.2 Análisis de fuerza de arcos

Los resultados de la prueba `arc.strength()` muestran que, en la DAG 3, la gran mayoría de sus arcos están cerca de 0, con excepción de las relaciones de p5_3 con p5_14_07 ($1.841923e-07$) y p5_3 con p5_14_12 ($2.608496e-02$). De esta manera, aún cuando estos valores son mayores a los demás, siguen siendo relaciones de dependencia significativas.

El comportamiento es similar para las DAG's con HC utilizando y sin restricciones. La variante en este caso es que todos los arcos tienen valores cercanos o iguales a cero ($0.000000e+00$). Por ello, también es posible confirmar que las relaciones de dependencia son significativas, y por ende, no es necesario removerlas.

4.3 Pruebas de hipótesis de dependencia

La prueba de independencia condicional basada en información mutua discreta evaluó los arcos de la DAG HC con restricciones. Se puede observar que el p-value $< 2.2e-16$ en cada uno de los arcos, lo que indica que se rechaza la hipótesis nula de independencia. Entonces se puede afirmar que sí existen relaciones de dependencia son estadísticamente significativas, incluso cuando las relaciones están condicionadas por otras variables. Asimismo, estos resultados junto con las pruebas de fuerza de arcos basados en información mutua reafirman el análisis sobre la DAG HC con restricciones.

4.4 Probabilidad de queries

Los resultados se dividieron en dos: las probabilidades diferentes e iguales a 0. Para las diferentes a 0, el query 2: la probabilidad de que una persona de 20 años y de sexo femenino use el tren ligero como medio de transporte dado que es de CDMX fue muy baja, cerca del 0.83%, sugiriendo que el uso del tren ligero por este grupo de mujeres es muy poco común bajo el modelo de los datos analizados. Para el query cuatro, la probabilidad fue del 29.48%. Esta es una probabilidad baja, no obstante, con más número de frecuencias, por tanto, podría ser poco frecuente, mas no improbable.

Por otro lado, en el caso del query 1, esto se debe a que no existen registros de personas que salgan de un hospital en Yucatán en los datos disponibles. Igualmente, la query tres tuvo un valor nulo, donde la baja cantidad de datos registrados de viajes dentro de Jalisco (solamente 13 registros) es lo que generó una representatividad insuficiente para capturar esta combinación específica de características. En ambos casos, el valor nulo refleja limitaciones de la muestra, y no indican que tales eventos sean improbables.

5 Conclusiones

Este proyecto mostró que la combinación de algoritmos de optimización con un previo conocimiento básico es una estrategia efectiva para construir redes bayesianas aplicadas al análisis de bases de datos (en este caso sobre movilidad urbana). Los resultados confirman que, aunque los algoritmos de Hill Climbing logran optimizar métricas estadísticas, la inclusión de restricciones en la estructura es esencial para asegurar que los modelos resultantes tengan sentido e interpretabilidad.

Al mismo tiempo, las limitaciones presentes durante la investigación dejan aprendizajes importantes para futuros trabajos. La experiencia con herramientas de control de versiones, en particular el uso inicial de GitHub por parte del equipo, creó algunos retos en la coordinación del trabajo colaborativo. Esto muestra la importancia de familiarizarse previamente con estas herramientas en proyectos de investigación.

La principal limitación del estudio se relaciona con el alcance geográfico de la base de datos, ya que está centrada únicamente en la zona metropolitana del Valle de México. Esto redujo la capacidad de responder queries sobre otras regiones del país y limitó la posibilidad de generalizar los resultados.

Para futuras investigaciones, es recomendable trabajar con bases de datos que incluyan mayor diversidad geográfica y demográfica, además de dedicar una fase más amplia a conocer el contexto y las características de los datos antes de formular queries específicas.

6 Referencias

Anthropic. (2024). Claude 3.5 Sonnet [Large language model]. Anthropic. https://claude.ai

INEGI. (2017). Encuesta Origen Destino en Hogares de la Zona Metropolitana del Valle de México (EOD) 2017. Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/programas/eod/2017/#datos_abiertos

RCoder. (2024, 2 enero). Exportar datos en R. RCODER. https://r-coder.com/exportar-datos-r/

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3), 1-22. https://doi.org/10.18637/jss.v035.i03