

TALLER 4: Modelos estadísticos que permitan comprender los mecanismos de formación de la red

Apellidos	Nombre	Código	Login
Arenas Solano	Karol Daniela	202014996	k.arenas
Fonseca Najar	Haider Yesid	202116856	h.fonseca
Urrea Lopez	Juan Sebastian	201914710	Js.urrea

- 1) Algunas de las métricas que se pueden utilizar para analizar la evolución de la red en el tiempo incluyen la densidad de la red, la longitud promedio del camino más corto, el coeficiente de clustering, el grado de los nodos, la cantidad de componentes conectados, entre otros.
 - a. **¿Está la red altamente conectada en cada instante de tiempo?** No. De hecho, la red se encuentra desconectada en 5 componentes distintos en $t=1$. La densidad de la red sí aumenta en el tiempo, pasando del 6,8% al 7,1%. También, la red se conecta en un único componente en $t=2$, lo que muestra que la conectividad mejoró en el tiempo. A pesar de esto, el coeficiente de clustering disminuye del 52% al 40% entre ambos períodos.
 - b. **¿La transmisión de información entre los estudiantes se puede dar de manera rápida?** En el instante $t=1$ NO, puesto que la red se encuentra desconectada. Existen 4 clusters pequeños que pueden representar pequeños grupos interconectados de amigos que solo interactúan entre ellos. Dentro de estos clusters, la conectividad es alta, siendo que en muchas ocasiones el camino más corto entre cualquier par de nodos de dichos componentes es 1. Sin embargo, al considerar el componente más grande (33 nodos), se observa que la longitud promedio del camino más corto es de 3,45 nodos. Por su parte, en el instante $t=2$ la longitud promedio del camino más corto es de 4,48 nodos, pero todo el grafo se encuentra conectado, por lo que la información puede transmitirse de forma más efectiva por la red.
 - c. **¿La cantidad de conexiones entre los estudiantes aumenta o disminuye con el tiempo?** Aumenta. En el instante $t=1$, el grado promedio de los nodos es de 3,149 conexiones, mientras que en $t=2$ es de 3,375. Para este cálculo se tuvo en cuenta que la red es No Dirigida.
- 2)
 - a. Construyan una tabla de frecuencia para cada una de las dos variables seleccionadas (Sport y Smoke). En este punto deberán poder determinar la cantidad de estudiantes que realizan o no cada acción.

Tabla de Frecuencias para T=1

	No Fuma Cigarrillo	Fuma Regularmente	Total
No hace actividad física	10	3	13
Hace Actividad Física regularmente	28	9	37
Total	38	12	50

Tabla de Frecuencia para T=2

	No fuma cigarrillo	Fuma regularmente	Total
No hace actividad física	11	10	21
Hace actividad física regularmente	22	7	29
Total	33	17	50

b) A partir de la tabla de frecuencia, determinen las siguientes probabilidades:

- i. Probabilidad de hacer ejercicio y fumar
- ii. Probabilidad de no hacer ejercicio y fumar
- iii. Probabilidad de hacer ejercicio y no fumar
- iv. Probabilidad de no hacer ejercicio y no fumar

Primera Tabla (T=1):

Probabilidad de hacer ejercicio y fumar

= Frecuencia de "Hace actividad física regularmente" y "Fuma regularmente" / Total
= 9 / 50

Probabilidad de no hacer ejercicio y fumar

= Frecuencia de "No hace actividad física" y "Fuma regularmente"
/ Total = 3 / 50

Probabilidad de hacer ejercicio y no fumar

= Frecuencia de "Hace actividad física regularmente" y "No fuma cigarrillo" / Total
= 28 / 50

Probabilidad de no hacer ejercicio y no fumar

= Frecuencia de "No hace actividad física" y "No fuma cigarrillo"
/ Total = 10 / 50

Segunda Tabla (T=2):

Probabilidad de hacer ejercicio y fumar

= Frecuencia de "Hace actividad física regularmente" y "Fuma regularmente" / Total
= 7 / 50

Probabilidad de no hacer ejercicio y fumar

= Frecuencia de "No hace actividad física" y "Fuma regularmente"
/ Total = 10 / 50

Probabilidad de hacer ejercicio y no fumar
= Frecuencia de "Hace actividad física regularmente" y "No fuma cigarrillo" / Total
= 22 / 50

Probabilidad de no hacer ejercicio y no fumar
= Frecuencia de "No hace actividad física" y "No fuma cigarrillo"
/ Total = 11 / 50

c) Calculen y concluyan sobre los siguientes odds de:

i. Hacer ejercicio tal que el estudiante fuma

En T = 1:

$$= (9/50) / (12/50) = 9/12 = 3/4$$

En T = 2 :

$$= (7/50) / (17/50) = 7/17$$

ii. Hacer ejercicio tal que el estudiante no fuma

En T = 1:

$$= (28/50) / (38/50) = 28/38 = 14/19$$

En T = 2:

$$= (22/50) / (33/50) = 22/33$$

Hacer ejercicio tal que el estudiante fuma:

En ambas ventanas de tiempo (T=1 y T=2), los resultados sugieren que, en general, los estudiantes que fuman tienen una probabilidad moderada de participar en actividad física. En la primera ventana (T=1), los estudiantes fumadores tienen una probabilidad de 3/4 de hacer ejercicio en comparación con los no fumadores. Sin embargo, en la segunda ventana (T=2), esta probabilidad disminuye, y los estudiantes fumadores tienen una probabilidad de 7/17 de hacer ejercicio. Esto indica que la relación entre fumar y hacer ejercicio puede variar en diferentes momentos, y puede haber una disminución en la probabilidad de hacer ejercicio entre los estudiantes fumadores en la segunda ventana de tiempo.

Hacer ejercicio tal que el estudiante no fuma:

En ambas ventanas de tiempo, los resultados sugieren que los estudiantes que no fuman tienen una probabilidad más alta de participar en actividad física en comparación con los estudiantes que fuman. En la primera ventana (T=1), los estudiantes no fumadores tienen una probabilidad de 14/19 de hacer ejercicio, y en la segunda ventana (T=2), esta probabilidad es de 22/33. Esto indica que, en general, los estudiantes no fumadores muestran una tendencia a tener una probabilidad más alta de hacer ejercicio en ambas ventanas de tiempo.

d) Evalúen e interpreten la relación entre las dos variables. Para ello deberán calcular los odds ratio relacionados a hacer ejercicio y fumar y hacer ejercicio y no fumar. Deben interpretar sus resultados contestando a preguntas como ¿en cuánto aumenta la posibilidad de que un estudiante haga ejercicio dado que fuma cigarrillos?

Hacer ejercicio tal que el estudiante fuma:

$$OR_{T1} = 9/28$$

$$OR_{T2} = 7/22$$

Conclusión: En ambas ventanas de tiempo (T=1 y T=2), la relación es negativa. El hecho de fumar hace que sea menos posible que un estudiante haga ejercicio.

Hacer ejercicio tal que el estudiante no fuma:

$$OR_{T1_no_fumar} = 28/22$$

$$OR_{T2_no_fumar} = 22/17$$

Conclusión: En ambas ventanas de tiempo (T=1 y T=2), la relación es positiva. Los estudiantes que no fuman tienen más posibilidades de hacer ejercicio.

No hacer ejercicio tal que el estudiante fuma:

$$OR_{T1_no_ejercicio} = 3/22$$

$$OR_{T2_no_ejercicio} = 9/17$$

Conclusión: En ambas ventanas de tiempo (T=1 y T=2), la relación es negativa. El hecho de fumar hace que sea menos posible que un estudiante no haga ejercicio.

No hacer ejercicio tal que el estudiante no fuma:

$$OR_{T1_no_ejercicio_no_fumar} = 22/3$$

$$OR_{T2_no_ejercicio_no_fumar} = 17/9$$

Conclusión: En ambas ventanas de tiempo (T=1 y T=2), la relación es positiva. Los estudiantes que no fuman tienen más posibilidades de no hacer ejercicio.

3)

- a. Elegimos el modelo Exponential Random Graph Model (ERGM) debido a su capacidad para modelar redes sociales complejas. ERGM nos permite analizar en detalle cómo ciertos atributos están relacionados con la formación de amistades, lo cual es crucial para abordar el problema del consumo de sustancias en adolescentes.
- b. Nuestro análisis reveló relaciones significativas entre los siguientes hábitos y la formación de amistades:
 - i. **Consumo de Alcohol:** Encontramos una clara homofilia en el consumo de alcohol. Los estudiantes que consumen alcohol tienden a ser amigos entre sí. Esto sugiere que el consumo de alcohol es un factor de cohesión en la formación de amistades.
 - ii. **Consumo de Cigarrillos:** En cuanto al consumo de cigarrillos, también observamos una tendencia a la homofilia. Los estudiantes que fuman regularmente tienden a asociarse con otros fumadores. Esta homofilia en el consumo de cigarrillos puede deberse a la afinidad entre personas con hábitos similares.
 - iii. **Actividad Física:** En el caso de la actividad física, encontramos que los estudiantes que hacen ejercicio regularmente tienden a formar amistades con otros estudiantes activos. Esto sugiere que la actividad física es un factor de cohesión en la formación de amistades, posiblemente relacionado con la participación en actividades deportivas conjuntas.
- c.
 - i. Utilizamos datos de redes sociales de adolescentes.
 - ii. Aplicamos un Modelo de Grafos de Respuesta Exponencial (ERGM) para analizar las redes de amistad en dos momentos (T1 y T2) y su

- relación con el consumo de alcohol, drogas, tabaco y la actividad física.
- iii. Seleccionamos métricas ERGM, como reciprocidad, transitividad y homofilia en hábitos.
 - iv. Ajustamos el modelo y evaluamos la significancia de las métricas en la formación de amistades.
- 4) Para construir la matriz de adyacencia, la matriz de atributos y la variable dependiente (predisposición en el uso de drogas) se utilizó el siguiente código de Python:

```
import networkx as nx
import pandas as pd

# Obtener la matriz de adyacencia en t=2
adjacency_matrix_t2 = nx.adjacency_matrix(G_t2)

# adjacency_matrix_t2 ahora contiene la matriz de adyacencia en t=2

# Combinar los DataFrames relevantes en attributes_data_t2
attributes_data_t2 = pd.DataFrame({"Node": alcohol_data["Node"]})

# Agregar las columnas de las variables de interés de t=2
attributes_data_t2["Alcohol"] = alcohol_data["t=2"]
attributes_data_t2["Drugs"] = drugs_data["t=2"]
attributes_data_t2["Smoke"] = smoke_data["t=2"]
attributes_data_t2["Sport"] = sport_data["t=2"]

attributes_data_t2.set_index("Node", inplace=True)

# Crear una matriz de atributos organizada
attributes_matrix_t2 = attributes_data_t2[["Smoke", "Sport"]].copy()

# attributes_matrix_t2 ahora contiene la matriz de atributos organizada en t=2

def transform_drugs(row):
    if row["Drugs"] in [1, 2]:
        return "Baja"
    else:
        return "Alta"

predisposition = attributes_data_t2.apply(transform_drugs, axis=1)
```

Seguidamente, se exportaron en Python para ser importados en R. Luego, se aplicó el modelo NAM:

```
# Exportar la matriz de adyacencia (W) en t=2 a un archivo CSV
adjacency_matrix_t2.toarray() # Convertir la matriz dispersa de NetworkX a una matriz densa
W_file = "W_matrix_t2.csv"
pd.DataFrame(adjacency_matrix_t2.toarray()).to_csv(W_file, index=False)

# Exportar la matriz de atributos (X) en t=2 a un archivo CSV
X_file = "X_matrix_t2.csv"
attributes_matrix_t2.to_csv(X_file)

# Exportar la columna 'Predisposition' a un archivo CSV
y_file = "y_column_t2.csv"
predisposition.to_csv(y_file, index=False)
```

```

# Importar la matriz de adyacencia (W) en R
W_file <- "W_matrix_t2.csv"
W <- as.matrix(read.csv(W_file, header = TRUE))

# Importar la matriz de atributos (X) en R
X_file <- "X_matrix_t2.csv"
X <- read.csv(X_file, row.names = 1, header = TRUE)
X <- as.matrix(X)

# Importar la columna 'Predisposition' en R
y_file <- "y_column_t2.csv"
y <- read.csv(y_file, header = TRUE)
y <- ifelse(y == "Alta", 1, 0)

m1<- lnam(y = y, x = X, W1 = W)
summary(m1)

```

Se obtiene como resultados:

```
> summary(m1)
```

Call:

```
lnam(y = y, x = X, W1 = W)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.55316	-0.13662	-0.02166	0.06107	0.89368

Coefficients:

	Estimate	Std. Error	Z value	Pr(> z)
Smoke	0.27843	0.06507	4.279	1.88e-05 ***
Sport	-0.17212	0.05023	-3.426	0.000612 ***
rho1.1	0.13114	0.03212	4.083	4.44e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	Estimate	Std. Error
Sigma	0.2724	0.001

Goodness-of-Fit:

Residual standard error: 0.3105 on 47 degrees of freedom (w/o Sigma)
Multiple R-Squared: 0.4192, Adjusted R-Squared: 0.3822
Model log likelihood: -7.59 on 46 degrees of freedom (w/Sigma)
AIC: 23.18 BIC: 30.83

Null model: meanstd

Null log likelihood: -25.14 on 48 degrees of freedom
AIC: 54.27 BIC: 58.1

AIC difference (model versus null): 31.09

Heuristic Log Bayes Factor (model versus null): 27.27

Los resultados obtenidos del modelo NAM (Network Autocorrelation Model) proporcionan información importante sobre la influencia social en la predisposición de los estudiantes a consumir drogas, utilizando las variables de atributos Smoke y Sport en la matriz de atributos.

Coefficientes para Smoke y Sport:

Smoke (0.27843): Un coeficiente positivo significativo indica que el consumo de cigarrillos en un estudiante está relacionado positivamente con la predisposición a consumir drogas. En otras palabras, los estudiantes que fuman cigarrillos tienen una mayor predisposición a consumir drogas.

Sport (-0.17212): Un coeficiente negativo significativo sugiere que la participación en actividad física regular está relacionada negativamente con la predisposición a consumir drogas. Los estudiantes que hacen ejercicio regularmente tienden a tener una predisposición menor a consumir drogas.

rho1.1 (0.13114): El coeficiente rho1.1 mide la influencia de la red de amistades en la predisposición de los estudiantes a consumir drogas. Un valor positivo y significativo (4.083) indica que la red de amistades tiene un efecto positivo en la predisposición a consumir drogas. En otras palabras, la predisposición a consumir drogas tiende a ser similar entre amigos, lo que respalda la idea de influencia social en el consumo de drogas.

Sigma (0.2724): Sigma representa el error estándar del modelo. Un valor bajo indica que el modelo se ajusta bien a los datos.

Goodness-of-Fit:

El modelo tiene un buen ajuste a los datos, ya que el coeficiente de determinación (R^2) múltiple es 0.4192, lo que significa que el modelo explica una parte significativa de la variabilidad en la predisposición a consumir drogas.

AIC (Criterio de Información de Akaike) y BIC (Criterio de Información Bayesiano):

Los valores AIC y BIC son medidas de la calidad del modelo en términos de ajuste y complejidad. Un modelo con un AIC y BIC más bajos se considera mejor. En este caso, el modelo NAM tiene un AIC de 23.18 y un BIC de 30.83, lo que sugiere que es un modelo adecuado.

En resumen, el modelo NAM indica que el consumo de cigarrillos (Smoke), la participación en actividad física regular (Sport), y la influencia de la red de amistades (rho1.1) son factores significativos en la predisposición de los estudiantes a consumir drogas. Los estudiantes que fuman cigarrillos tienden a tener una mayor predisposición, mientras que aquellos que hacen ejercicio regularmente tienen una menor predisposición. Además, la red de amistades ejerce una influencia positiva en la predisposición de los estudiantes a consumir drogas, lo que sugiere la presencia de influencia social en este comportamiento.

Dado que el consumo de sustancias (tabaco y alcohol) muestra una correlación significativa con la predisposición a consumir drogas, el Ministerio debería implementar programas de prevención del consumo de sustancias en las escuelas. Estos programas podrían incluir charlas educativas, campañas de concientización y apoyo a estudiantes que deseen dejar de fumar o consumir alcohol. La idea es reducir la influencia negativa de estos hábitos en la predisposición de los estudiantes a consumir drogas.

Dado que la actividad física está relacionada negativamente con la predisposición a consumir drogas, el Ministerio debe enfocarse en fomentar la actividad física en las escuelas. Esto se puede lograr mediante la implementación de programas deportivos, actividades extracurriculares relacionadas con el ejercicio, y la creación de espacios adecuados para la práctica de deportes. Incentivar la participación activa de los estudiantes en deportes y actividades físicas puede ayudar a reducir su predisposición al consumo de drogas.

Dado que la influencia de la red de amistades es un factor significativo en la predisposición de los estudiantes a consumir drogas, el Ministerio podría promover la formación de relaciones saludables y amistades positivas. Esto podría incluir programas de mentoría, talleres de desarrollo personal y social, y actividades que fomenten la cooperación y la construcción de relaciones sólidas y saludables.