Universidad de San Carlos de Guatemala

Facultad de Ingeniería

Escuela de Estudios de Postgrado

Maestría en ingeniería para la industria con especialización en ciencia de la computación

PROYECTO MINERIA DE DATOS PARTE 1

Alvaro Andrés Díaz De León
999013647

Guatemala, Julio de 2024

I. INDICE

1. RESUMEN

Este proyecto se centra en el análisis de patrones de gasto y migración en diferentes contextos socioeconómicos y geográficos en Guatemala. Utilizando técnicas de minería de datos, específicamente las reglas de asociación Apriori y FP-Growth, se identificaron patrones clave en los gastos de hogares en extrema pobreza en áreas rurales y de hogares no pobres en áreas urbanas metropolitanas. Además, se analizaron patrones de migración tanto en áreas rurales como urbanas, enfocándose en la cantidad de familiares viviendo en el extranjero y los destinos de migración. Estos análisis permiten comprender mejor las prioridades y comportamientos de diferentes grupos poblacionales en Guatemala, proporcionando una base sólida para generar propuestas concretas para mejorar las condiciones de vida y abordar problemáticas específicas identificadas en el análisis.

2. REGLAS DE ASOCIACION APRIORI

Las reglas de asociación Apriori son un método de minería de datos que identifica relaciones entre variables en grandes bases de datos. Este algoritmo encuentra conjuntos de elementos que frecuentemente aparecen juntos y genera reglas del tipo "Si A, entonces B". Evalúa la relevancia de las reglas mediante medidas como soporte y confianza.

2.1. PATRON 1

Para el primer patrón se delimita el dataset a las personas que se encuentran en el dominio del área rural que están en extrema pobreza, se busca definir en que gastan estas personas y cuanto es lo que gastan por producto

2.1.1. Diccionario

Información de variable

Variable	Etiqueta
DOMINIO	Dominio de Estudio
NO_HOGAR	Número de hogar
POBREZA	Clasificación de hogar (Pobreza)
ID_GASTOSMP	ID GASTOMP
DESC_GASTOSMP	DESC GASTOMP
P13B05	5. El mes pasado, ¿usted o alguna persona del hogar gastaron dinero en…
P13B06	6. ¿Cuánto gastaron en total durante el mes pasado en (...)?

Valores de variable

Valor		Etiqueta
DOMINIO	1	Urbano Metropolitano
POBREZA	1,00	Pobre extremo
P13B05	1	Si

2.1.2. Tabla codigo

```
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("POCUPA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEI"))]

datamsc <- subset(datagastos, P13B05 == 1)

datamsc <- subset(datamsc, DOMINIO == 3)

datamsc <- subset(datamsc, POBREZA == 1)

datamsc2 <- datamsc[, c(5,10,11,13)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 50))

datamsc2
```

2.1.3. Ilustración del resultado

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{}	=>	{P13B06=[14,30)}	0.3184151	0.3184151	1	1	1117
[2]	{}	=>	{NO_HOGAR=[836,7.11e+03)}	0.3315279	0.3315279	1	1	1163
[3]	{}	=>	{P13B06=[1,14)}	0.3318130	0.3318130	1	1	1164
[4]	{}	=>	{NO_HOGAR=[8e+03,1.1e+04]}	0.3340935	0.3340935	1	1	1172
[5]	{}	=>	{NO_HOGAR=[7.11e+03,8e+03)}	0.3343786	0.3343786	1	1	1173
[6]	{}	=>	{P13B06=[30,500]}	0.3497719	0.3497719	1	1	1227
[7]	{}	=>	{ID_GASTOSMP=[8,29]}	0.3557583	0.3557583	1	1	1248
[8]	{}	=>	{ID_GASTOSMP=[4,8)}	0.3868301	0.3868301	1	1	1357

8 rows

2.1.4. Discusión del resultado

En base al diccionario podemos encontrar que la mayor concentración de gastos está en los siguientes productos:

rhs	support	confidence
{ID_GASTOSMP=[4,8)}	0.3868301	0.3868301

Los cuales son los siguientes:

ID_GASTOSMP	DESC_GASTOSMP
4	Jabón de baño, shampú, acondicionador, etc.?
5	Pasta dental, cepillo dental, hilo dental, enjuague bucal, etc.?
6	Papel higiénico y toallas sanitarias, protectores femeninos, servilletas, toallas desechables, etc.?
7	Cepillos para el cabello, peines, peinetas, ganchos, diademas, colas, tubos, etc.?

También se puede ver precios desde Q30 hasta Q500 en total por cada producto adquirido el mes pasado

rhs	support	confidence
{P13B06=[30,500]}	0.3497719	0.3497719

2.1.5. Interpretación de resultados

El análisis muestra que una alta concentración de gastos se destina a productos de higiene personal, con un soporte de 0.3868301 (38.7% de las transacciones). Además, los montos gastados en estos productos oscilan entre Q30 y Q500 mensuales, con un soporte de 0.3497719 (35% de las transacciones). Esto sugiere que, incluso en extrema pobreza, las personas asignan una parte considerable de su presupuesto a artículos de higiene personal.

2.2. PATRON 2

Para el segundo patrón se delimita el dataset a las personas que se encuentran en el dominio del área urbano metropolitano que no están en pobreza, se busca definir en que gastan estas personas y cuanto es lo que gastan por producto

2.2.1. Diccionario

Información de variable

Variable	Etiqueta
DOMINIO	Dominio de Estudio
NO_HOGAR	Número de hogar
POBREZA	Clasificación de hogar (Pobreza)
ID_GASTOSMP	ID GASTOMP
P13B05	5. El mes pasado, ¿usted o alguna persona del hogar gastaron dinero en…
P13B06	6. ¿Cuánto gastaron en total durante el mes pasado en (...)?

Valores de variable

Valor		Etiqueta
DOMINIO	1	Urbano Metropolitano
POBREZA	3,00	No pobre
P13B05	1	Si

2.2.2. Tabla código

```
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("POCUPA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEI"))]

datamsc <- subset(datagastos, P13B05 == 1)

datamsc <- subset(datamsc, DOMINIO == 1)

datamsc <- subset(datamsc, POBREZA == 3)

datamsc2 <- datamsc[, c(5,10,13)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 20))

datamsc
```

2.2.3. Ilustración del resultado

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{}	=>	{NO_HOGAR=[1,238)}	0.3306915	0.3306915	1	1	1669
[2]	{}	=>	{NO_HOGAR=[600,888]}	0.3340598	0.3340598	1	1	1686
[3]	{}	=>	{NO_HOGAR=[238,600)}	0.3352487	0.3352487	1	1	1692
[4]	{}	=>	{P13B06=[75,3e+03]}	0.3455518	0.3455518	1	1	1744
[5]	{}	=>	{P13B06=[30,75)}	0.3550624	0.3550624	1	1	1792
[6]	{}	=>	{ID_GASTOSMP=[5,13)}	0.3566475	0.3566475	1	1	1800
[7]	{}	=>	{ID_GASTOSMP=[13,30]}	0.3970676	0.3970676	1	1	2004

2.2.4. Discusión del resultado

En base al diccionario podemos encontrar que la mayor concentración de gastos está en los siguientes productos:

rhs	support	confidence
{ID_GASTOSMP=[13,30]}	0.3970676	0.3970676

Los cuales son los siguientes:

ID_GASTOSMP	DESC_GASTOSMP
12	Guantes para lavar y de cocina, esponjas, lazos, ganchos para colgar ropa, limpiadores, escurridor de platos, etc.?
13	Desinfectantes para piso y baños, desodorantes ambientales e insecticidas, limpiavidrios, limpiadores de muebles, repelentes, etc.?
14	Aceite de bebé, hisopos, mamones, pepes, pachas, pañales desechables y/o de tela, camisetas, baberos, frazaditas para bebé, etc.?
15	Hilos para coser, lanas, botones, elásticos, zippers y similares, etc.? (para uso del hogar)
16	Libros y revistas (no incluya los textos escolares)
17	Colonias, desodorantes, lociones, talcos, perfumes, gel para el cabello, vaselina, tratamiento para el cabello, etc.?
18	Alka seltzer, sal andrews, aspirinas, alcohol, etc.? (medicinas para primeros auxilios y medicamentos comprados sin receta)
19	Comidas para mascotas? (alpiste, concentrados, etc.)
20	Lavado planchado y reparación de prendas de vestir fuera del hogar?
21	Recreación, diversión como: espectáculos públicos, cine, fútbol, compra de casettes, CD`s, DVD`s, etc.?
22	Barbería (corte de pelo y afeitada), salón de belleza, (peinado, rizado, manicure, pedicure, maquillaje)
23	Servicio de empleada doméstica, lavandera, planchadora, chofer, jardinero, guardaespaldas que viven en el hogar?
24	Servicio de empleada doméstica, lavandera, planchadora, chofer, jardinero, guardaespaldas que NO viven en el hogar?
25	Peaje (uso de autopistas)?
26	Gimnasio, sauna, baño turco, masajes, etc.?
27	Pagos por pensión alimenticia?
28	Gastos por pago de parqueo para vehículos del hogar?
29	Gastos por pasajes extraurbanos (rutas largas)?
30	Gastos por servicios de vigilancia, guardias de seguridad?

También se puede ver precios desde Q75 hasta Q3000 en total por cada producto adquirido el mes pasado

rhs	support	confidence
{P13B06=[75,3000]}	0.3455518	0.3455518

2.2.5. Interpretación de resultados

Estos patrones de gasto indican que las personas en áreas urbanas metropolitanas que no están en pobreza asignan una parte significativa de su presupuesto a una amplia variedad de productos, desde artículos de limpieza y desinfectantes hasta servicios de entretenimiento y

cuidado personal. Esta diversidad en el gasto sugiere un mayor nivel de consumo y una prioridad en mantener tanto el hogar como el bienestar personal y familiar.

2.3. PATRON 3

Para el tercer patrón se centra en la migración enfocado en las personas que más tienen familiares viviendo en el exterior para ver de qué departamento provienen.

2.3.1. Diccionario

Información de variable

Variable	Etiqueta
DEPTO	Departamento
P01I01A	1.a En los ultimos 5 años ¿Alguna persona que vivía en este hogar, vive actualmente en otro país?
P01I01B	1.b ¿Cuántas personas del hogar vive actualmente en otro país ahora?

Valores de variable

Valor		Etiqueta
P01I01A	1	Si

2.3.2. Tabla código

```
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("PEA"))]
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("POCUPA"))]
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("PEI"))]

datamsc <- subset(datamigracion, P01I01A == 1)


datamsc2 <- datamsc[, c(1,10)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 20))

datamsc2
```

2.3.3. Ilustración del resultado

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	{}	=>	{DEPTO=[1,10)}	0.3144186	0.3144186	1.0000000	1	338
[2]	{}	=>	{DEPTO=[16,22]}	0.3386047	0.3386047	1.0000000	1	364
[3]	{}	=>	{DEPTO=[10,16)}	0.3469767	0.3469767	1.0000000	1	373
[4]	{}	=>	{P01I01B=[1,7]}	1.0000000	1.0000000	1.0000000	1	1075
[5]	{DEPTO=[1,10)}	=>	{P01I01B=[1,7]}	0.3144186	1.0000000	0.3144186	1	338
[6]	{P01I01B=[1,7]}	=>	{DEPTO=[1,10)}	0.3144186	0.3144186	1.0000000	1	338
[7]	{DEPTO=[16,22]}	=>	{P01I01B=[1,7]}	0.3386047	1.0000000	0.3386047	1	364
[8]	{P01I01B=[1,7]}	=>	{DEPTO=[16,22]}	0.3386047	0.3386047	1.0000000	1	364
[9]	{DEPTO=[10,16)}	=>	{P01I01B=[1,7]}	0.3469767	1.0000000	0.3469767	1	373
[10]	{P01I01B=[1,7]}	=>	{DEPTO=[10,16)}	0.3469767	0.3469767	1.0000000	1	373

2.3.4. Discusión del resultado

El siguiente patrón indica que las personas de los hogares que se entrevistaron en los siguientes departamentos tiene alrededor de 1 a 7 personas en el extranjero

lhs		rhs	support	confidence
{DEPTO=[10,16)}	=>	{P01I01B=[1,7]}	0.3469767	1

Los cuales son los siguientes departamentos:

Valor		Etiqueta
DEPTO	10	Suchitepéquez
	11	Retalhuleu
	12	San Marcos
	13	Huehuetenango
	14	Quiché
	15	Baja Verapaz

2.3.5. Interpretación de resultados

El análisis revela que los departamentos con mayor migración hacia el extranjero son Suchitepéquez, Retalhuleu, San Marcos, Huehuetenango, Quiché, y Baja Verapaz. En estos departamentos, el 34.7% de las familias tienen entre 1 y 7 miembros viviendo fuera del país. Esta tendencia sugiere una alta incidencia de migración internacional, especialmente en los departamentos del pais fronterizos como México, en busca de mejores oportunidades.

2.4. PATRON 4

El cuarto patrón se enfoca en la migración también, pero esta vez en las personas no pobres del área urbano metropolitano que migraron del país.

2.4.1. Diccionario

Información de variable

Variable	Etiqueta
DOMINIO	Dominio de Estudio
POBREZA	Clasificación de hogar (Pobreza)
P01I02	2. Sexo
P01I03	3. ¿Qué edad tenia al irse?
P01I05	5. ¿En qué país se encuentra actualmente?

Valores de variable

Valor		Etiqueta
DOMINIO	1	Urbano Metropolitano
POBREZA	3,00	No pobre

2.4.2. Tabla código

```
datamigracion2 <- datamigracion2[ , !(names(datamigracion2) %in%
c("PEA"))]
datamigracion2 <- datamigracion2[ , !(names(datamigracion2) %in%
c("POCUPA"))]
datamigracion2 <- datamigracion2[ , !(names(datamigracion2) %in%
c("PEI"))]

datamsc <- subset(datamigracion2, DOMINIO == 1)

datamsc3 <- subset(datamsc, POBREZA == 3)

datamsc2 <- datamsc3[, c(10, 11,13)]

reglas <- apriori(datamsc2, parameter = list(support=0.3,
confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 60))

datamsc3
```

2.4.3. Ilustración del resultado

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[11]	{P01I03=[23,33.3)}	=>	{P01I02=[1,2]}	0.3611111	1.0000000	0.3611111	1	13
[12]	{P01I02=[1,2]}	=>	{P01I03=[23,33.3)}	0.3611111	0.3611111	1.0000000	1	13
[13]	{P01I05=[3.01e+03,8.1e+03]}	=>	{P01I02=[1,2]}	0.3611111	1.0000000	0.3611111	1	13
[14]	{P01I02=[1,2]}	=>	{P01I05=[3.01e+03,8.1e+03]}	0.3611111	0.3611111	1.0000000	1	13
[15]	{P01I05=[3e+03,3.01e+03)}	=>	{P01I02=[1,2]}	0.6388889	1.0000000	0.6388889	1	23
[16]	{P01I02=[1,2]}	=>	{P01I05=[3e+03,3.01e+03)}	0.6388889	0.6388889	1.0000000	1	23

2.4.4. Discusión del resultado

En la discusión se encuentra que los hombres y mujeres con buena calidad de vida en su mayoría viajan ya sea a norte america o a los alrededores del país

lhs		rhs	support	confidence
{P01I02=[1,2]}	=>	{P01I05=[3000,3010)}	0.6388889	0.6388889

Los países son los siguientes:

Valor		Etiqueta
P01I05	0	PAÍS DESCONOCIDO
	3001	ESTADOS UNIDOS
	3002	CANADÁ
	3003	MÉXICO
	3004	EL SALVADOR
	3005	BELICE
	3006	COSTA RICA
	3007	HONDURAS
	3008	NICARAGUA

2.4.5. Interpretación de resultados

El patrón identificado tiene un soporte de 0.6388889, indicando que el 63.9% de las personas migrantes pertenecen a este grupo. Esto se puede dar a muchos factores, ya que el estilo de vida de los países aledaños es similar con un estilo de vida mas barato en algunos casos como lo es México por ejemplo y en el caso de Norte América puede ser por oportunidades laborares o turismo.

3. REGLAS DE ASOCIACION FP-GROWTH

Las reglas de asociación FP-Growth son un método de minería de datos que identifica relaciones entre variables en grandes bases de datos. Este algoritmo utiliza una estructura de árbol (FP-tree) para encontrar conjuntos de elementos frecuentes sin generar candidatos, lo que lo hace más eficiente que Apriori. FP-Growth genera reglas del tipo "Si A, entonces B", evaluando su relevancia con medidas como soporte y confianza.

3.1. PATRON 1

El primer patrón intenta determinar la causa de robo de las personas que utilizan el cajero automático

3.1.1. Diccionario

Información de variable

Variable	Etiqueta
ID_SEGURIDAD	ID Seguridad
DESC_SEGURIDAD	Descripción de seguridad ciudadana
P02A04B	4.b ¿Dónde se encontraba (...) cuando fue víctima de (...)
P02A07	7. ¿Cuál fue la razón principal para no presentar la denuncia del/la (...)?

Valores de variable

Valor		Etiqueta
P02A04B	6	Al salir del banco o cajero automático

3.1.2. Tabla código

```
datamsc <- subset(data, P02A04B == 6)
```

```
datamsc2 <- datamsc[, c(10,11,23)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules",
supp = .2, conf = .2)

inspect(head(sort(reglas, by = "lift"), 20))
```

3.1.3. Ilustración del resultado

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[11]	{ID_SEGURIDAD=[6,15]}	=>	{P02A07.=[6,8]}	0.75	0.7500000	1.000000	3
[12]	{}	=>	{P02A07.=[6,8]}	0.75	0.7500000	1.000000	3
[13]	{DESC_SEGURIDAD=Fraude bancario}	=>	{ID_SEGURIDAD=[6,15]}	0.50	1.0000000	1.000000	2
[14]	{ID_SEGURIDAD=[6,15]}	=>	{DESC_SEGURIDAD=Fraude bancario}	0.50	0.5000000	1.000000	2
[15]	{DESC_SEGURIDAD=Fraude bancario, P02A07.=[6,8]}	=>	{ID_SEGURIDAD=[6,15]}	0.50	1.0000000	1.000000	2
[16]	{}	=>	{DESC_SEGURIDAD=Fraude bancario}	0.50	0.5000000	1.000000	2
[17]	{DESC_SEGURIDAD=Estafa}	=>	{ID_SEGURIDAD=[6,15]}	0.50	1.0000000	1.000000	2
[18]	{ID_SEGURIDAD=[6,15]}	=>	{DESC_SEGURIDAD=Estafa}	0.50	0.5000000	1.000000	2
[19]	{DESC_SEGURIDAD=Estafa, P02A07.=[6,8]}	=>	{ID_SEGURIDAD=[6,15]}	0.25	1.0000000	1.000000	1
[20]	{}	=>	{DESC_SEGURIDAD=Estafa}	0.50	0.5000000	1.000000	2

11-20 of 20 rows Previous 1 2 Next

3.1.4. Discusión del resultado

El patron muestra que la mayoría de robos al salir del banco o algún cajero automático ha sido por medio de fraude o estafa

lhs		rhs	support	confidence
{DESC_SEGURIDAD=Fraude bancario}	=>	{ID_SEGURIDAD=[6,15]}	0.5	1
{DESC_SEGURIDAD=Estafa}	=>	{ID_SEGURIDAD=[6,15]}	0.5	1

3.1.5. Interpretación de resultados

Este resultado es curioso porque usualmente se tiene la creencia de que los robos son a mano armada pero también hay pruebas de que en los últimos años los cajeros han sido alterados para robar información y realizar estafas.

3.1. PATRON 2

En el segundo patrón se busca encontrar en que lugar es donde se concentra la mayor parte de violencia y de que tipo es

3.1.1. Diccionario

Información de variable

Variable	Etiqueta
ID_SEGURIDAD	ID Seguridad
P02A02	2. En los últimos doce meses ¿cuántas veces fueron víctimas de (...)
P02A04B	4.b ¿Dónde se encontraba (...) cuando fue víctima de (...)
P02A05	5. ¿Quiénes fueron los agresores?

3.1.2. Tabla código

```
datamsc <- subset(data, P02A02 >= 10)

datamsc2 <- datamsc[, c(10,15, 16)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules",
supp = .2, conf = .2)

inspect(reglas[])
```

3.1.3. Ilustración del resultado

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1]	{}	=>	{P02A05=[3,4]}	0.7391304	0.7391304	1.0000000	17
[2]	{P02A04B=[12,13]}	=>	{P02A05=[3,4]}	0.5652174	0.8125000	1.0992647	12
[3]	{P02A05=[3,4]}	=>	{P02A04B=[12,13]}	0.5652174	0.7647059	1.0992647	12
[4]	{}	=>	{P02A04B=[12,13]}	0.6956522	0.6956522	1.0000000	16
[5]	{ID_SEGURIDAD=[9,10)}	=>	{P02A05=[3,4]}	0.3478261	0.7272727	0.9839572	8
[6]	{P02A05=[3,4]}	=>	{ID_SEGURIDAD=[9,10)}	0.3478261	0.4705882	0.9839572	8
[7]	{ID_SEGURIDAD=[9,10), P02A04B=[12,13]}	=>	{P02A05=[3,4]}	0.3043478	0.7777778	1.0522876	7
[8]	{ID_SEGURIDAD=[9,10), P02A05=[3,4]}	=>	{P02A04B=[12,13]}	0.3043478	0.8750000	1.2578125	7
[9]	{P02A04B=[12,13], P02A05=[3,4]}	=>	{ID_SEGURIDAD=[9,10)}	0.3043478	0.5384615	1.1258741	7
[10]	{ID_SEGURIDAD=[9,10)}	=>	{P02A04B=[12,13]}	0.3913043	0.8181818	1.1761364	9

1-10 of 22 rows Previous 1 2 3 Next

3.1.4. Discusión del resultado

Los resultados nos indican que el lugar donde mayor se percata la violencia es dentro de los hogares y en el trabajo, por personas desconocidas y que la violencia es de tipo amenaza

lhs		rhs	support	confidence
{P02A04B=[12,13], P02A05=[3,4]}	=>	{ID_SEGURIDAD=[9,10)}	0.3043478	0.5384615

Lugar donde se realiza la violencia:

P02A04B	12	En la vivienda
	13	En el trabajo

Persona quien realiza el acto de violencia:

P02A05	3	Personas desconocidas
	4	Familiares

Tipo de violencia:

ID_SEGURIDAD	DESC_SEGURIDAD
9	Amenazas

3.1.5. Interpretación de resultados

Usualmente se tiene la percepción que afuera del hogar o trabajo es donde hay mas violencia y se asocia mucho al robo armado, pero realmente basado en los resultados obtenido se obtiene que donde más se sufre es en los lugares supuestamente más seguros como lo es el lugar de empleo o el hogar y es por medio de amenazas que se violenta a las personas

3.1. PATRON 3

Para el patrón numero 3 se busca identificar cuales son los tipos de violencia para la población no pobre.

3.1.1. Diccionario

Información de variable

Variable	Etiqueta
DEPTO	Departamento
ID_SEGURIDAD	ID Seguridad
P02A02	2. En los últimos doce meses ¿cuántas veces fueron víctimas de (...)
P02A04B	4.b ¿Dónde se encontraba (...) cuando fue víctima de (...)
P02A05	5. ¿Quiénes fueron los agresores?

Valores de variable

Valor		Etiqueta
POBREZA	3,00	No pobre

3.1.2. Tabla código

```
datamsc <- subset(data, POBREZA == 3)

datamsc2 <- datamsc[, c(10,1,13, 15, 16)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules",
supp = .1, conf = .1)

inspect(reglas[])
```

3.1.3. Ilustración del resultado

	lhs <chr>		rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[11]	{DEPTO=[5,12)}	=>	{ID_SEGURIDAD=[1,6)}	0.1079520	0.3330201	0.9993414	9215
[12]	{ID_SEGURIDAD=[1,6)}	=>	{DEPTO=[5,12)}	0.1079520	0.3239471	0.9993414	9215
[13]	{DEPTO=[5,12)}	=>	{ID_SEGURIDAD=[6,11)}	0.1079872	0.3331285	1.0009334	9218
[14]	{ID_SEGURIDAD=[6,11)}	=>	{DEPTO=[5,12)}	0.1079872	0.3244632	1.0009334	9218
[15]	{DEPTO=[5,12)}	=>	{ID_SEGURIDAD=[11,15]}	0.1082215	0.3338513	0.9997270	9238
[16]	{ID_SEGURIDAD=[11,15]}	=>	{DEPTO=[5,12)}	0.1082215	0.3240721	0.9997270	9238
[17]	{}	=>	{DEPTO=[5,12)}	0.3241606	0.3241606	1.0000000	27671
[18]	{DEPTO=[1,5)}	=>	{ID_SEGURIDAD=[1,6)}	0.1068977	0.3332359	0.9999890	9125
[19]	{ID_SEGURIDAD=[1,6)}	=>	{DEPTO=[1,5)}	0.1068977	0.3207832	0.9999890	9125
[20]	{DEPTO=[1,5)}	=>	{ID_SEGURIDAD=[6,11)}	0.1067220	0.3326882	0.9996102	9110

11-20 of 24 rows Previous 1 2 3 Next

3.1.4. Discusión del resultado

Este patrón nos indica que los departamentos alrededor del departamento de Guatemala sufren principales actos de violencia como lo son robo y daños a la propiedad privada.

lhs		rhs	support	confidence
{DEPTO=[1,5)}	=>	{ID_SEGURIDAD=[1,6)}	0.1068977	0.3332359

Valor		Etiqueta
DEPTO	1	Guatemala
	2	El Progreso
	3	Sacatepéquez
	4	Chimaltenango

ID_SEGURIDAD	DESC_SEGURIDAD
1	Robo de vehículos o autopartes
2	Robo de motocicletas o autopartes
3	Robo a viviendas
4	Robo con violencia
5	Robo sin violencia (hurto)

3.1.5. Interpretación de resultados

El análisis revela que en los departamentos alrededor del departamento de Guatemala, la población no pobre sufre principalmente de robos y daños a la propiedad privada, destacando robos de vehículos, motocicletas, viviendas, y robos con o sin violencia. Este patrón muestra que el 10.7% de las transacciones en el dataset incluyen estos tipos de violencia, y en un 33.3% de los casos en los departamentos mencionados se sufre alguno de estos robos.

3.1. PATRON 4

El cuarto patrón busca encontrar la razón del porque las denuncias de violencia no son realizadas

3.1.1. Diccionario

Información de variable

Variable	Etiqueta
DEPTO	Departamento
ID_SEGURIDAD	ID Seguridad
P02A07	7. ¿Cuál fue la razón principal para no presentar la denuncia del/la (...)?

Valores de variable

Valor		Etiqueta
P02A01	1	Si

3.1.2. Tabla código

```
datamsc <- subset(data, P02A01 == 1)

datamsc2 <- datamsc[, c(1,10, 23)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules",
supp = .2, conf = .2)

inspect(head(sort(reglas, by = "lift"), 10))
```

3.1.3. Ilustración del resultado

	lhs <chr>	<chr>	rhs <chr>	support <dbl>	confidence <dbl>	lift <dbl>	count <int>
[1]	{}	=>	{ID_SEGURIDAD=[9,15]}	0.4261566	0.4261566	1	479
[2]	{}	=>	{P02A07.=[6,97]}	0.3959075	0.3959075	1	444
[3]	{}	=>	{DEPTO=[5,13)}	0.3603203	0.3603203	1	405
[4]	{}	=>	{DEPTO=[13,22]}	0.3487544	0.3487544	1	392
[5]	{}	=>	{ID_SEGURIDAD=[4,9)}	0.3140569	0.3140569	1	353
[6]	{}	=>	{DEPTO=[1,5)}	0.2909253	0.2909253	1	327
[7]	{}	=>	{P02A07.=[3,6)}	0.2775801	0.2775801	1	312
[8]	{}	=>	{P02A07.=[1,3)}	0.2749110	0.2749110	1	309
[9]	{}	=>	{ID_SEGURIDAD=[1,4)}	0.2597865	0.2597865	1	292

9 rows

3.1.4. Discusión del resultado

Según el resultado obtenido las personas no realizan la denuncia debido a una gran variedad de razones de las que mas resaltan son miedo, desconocimiento, desagrado. Adicional, a poca competencia de las autoridades.

rhs	support	confidence
{P02A07.=[6,97]}	0.3959075	0.3959075

P02A07	6	La policía/autoridad competente no hubiera hecho nada
	7	Desagrado o miedo a la policía/autoridades/no quería nada que ver con la policía/autoridades
	8	No me atreví por miedo a represalias
	9	El proceso burocrático es muy complicado
	10	No conozco el procedimiento para denunciar delitos
	11	El costo del procesimiento es caro
	97	No sabe/no contesta

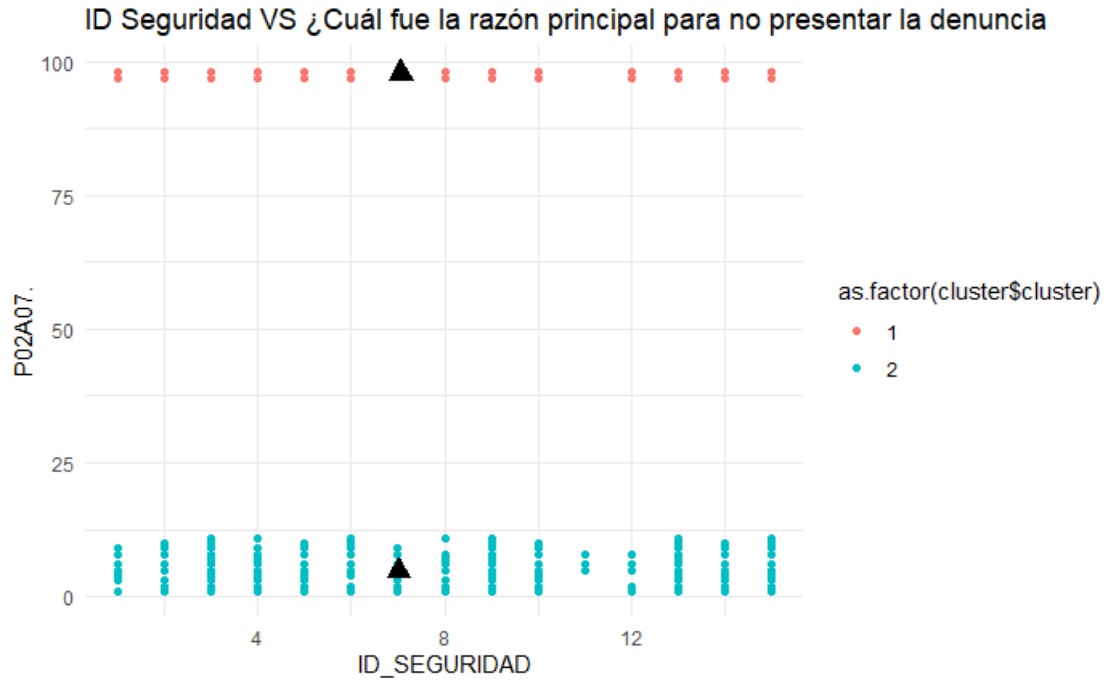
3.1.5. Interpretación de resultados

El cuarto patrón revela que las principales razones por las cuales las personas no presentan denuncias de violencia incluyen la percepción de que la policía o autoridades competentes no actuarán (39.6%), miedo o desagrado hacia las autoridades, temor a represalias, desconocimiento del procedimiento, y la complicación o costo del proceso burocrático. Este hallazgo subraya la necesidad de mejorar la confianza en las autoridades y simplificar el proceso de denuncia para fomentar una mayor presentación de denuncias por parte de las víctimas.

4. ANÁLISIS DE CLÚSTER

En el análisis de los clúster se encuentra 2 patrones interesantes el primero indica que hay una cantidad considerable en las personas que no saben como poner una denuncia (clúster 1) por lo que nos indica que se debe de incrementar las campañas de información acerca del procedimiento del mismo.

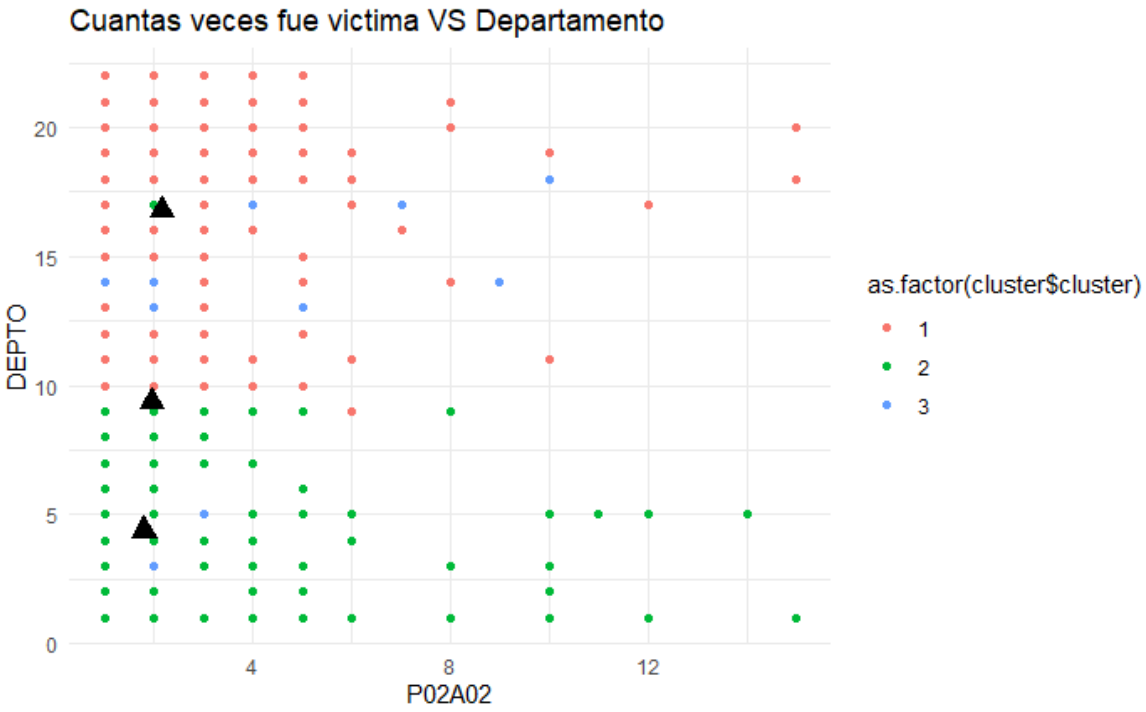
```
datamsc <- subset(data, P02A01 == 1)

datamsc2 <- datamsc[, c(1,10, 23)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules", supp = .2,
conf = .2)

inspect(head(sort(reglas, by = "lift"), 10))


datamsc2 <- na.omit(datamsc2)
datamsc2
cluster <- kmeans(datamsc2, centers=2)

ggplot(datamsc2, aes(x=ID_SEGURIDAD ,y=P02A07., color =
as.factor(cluster$cluster)))+
  geom_point()+
  geom_point(data = as.data.frame(cluster$centers), aes(x=ID_SEGURIDAD
,y=P02A07.), , color="black", size =4, shape=17)+
  labs(title = "ID Seguridad VS ¿Cuál fue la razón principal para no
presentar la denuncia")+
  theme_minimal()
```



Hay una correlación interesante entre los departamentos, las personas que viven actualmente fuera del país y el nivel de violencia de los departamentos lo que puede indicar que las personas que migran pueden ser por una posible violencia aparte de la pobreza como tal como puede verse en el artículo “La migración y las múltiples violencias en Guatemala”

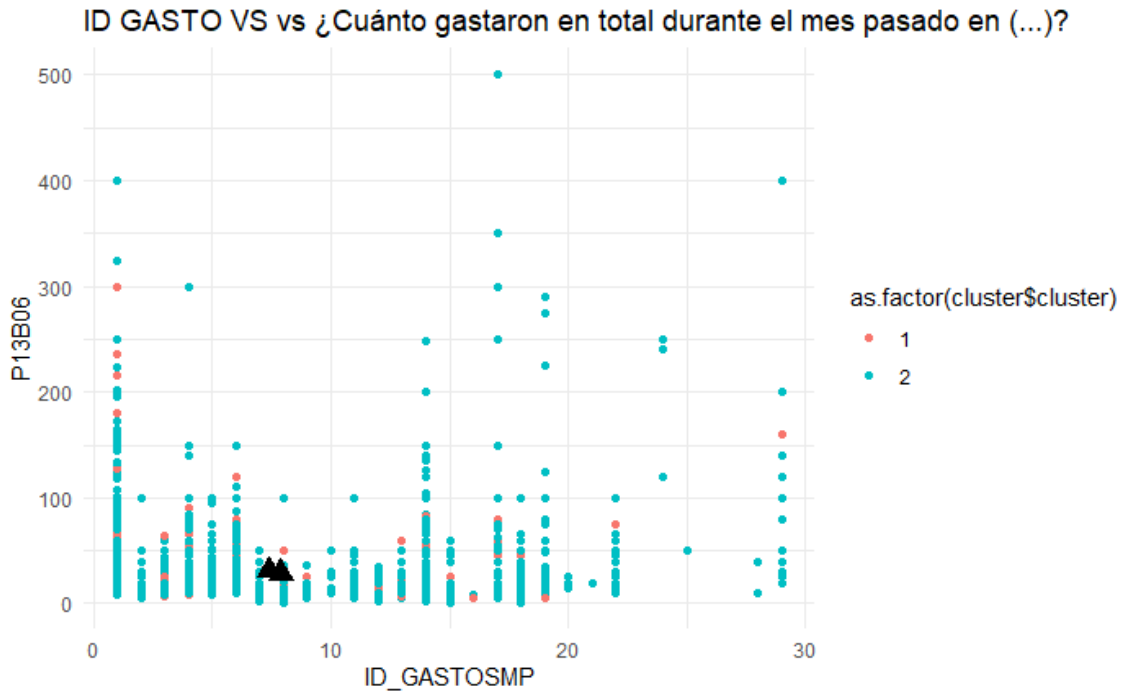
```
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("PEA"))]
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("POCUPA"))]
datamigracion <- datamigracion[ , !(names(datamigracion) %in% c("PEI"))]

datamsc <- subset(datamigracion, P01I01A == 1)


datamsc2 <- datamsc[, c(1,10)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 20))

datamsc2

cluster <- kmeans(datamsc2, centers=3)

ggplot(datamsc2, aes(x=P01I01B ,y=DEPTO, color =
as.factor(cluster$cluster)))+
  geom_point()+
  geom_point(data = as.data.frame(cluster$centers), aes(x=P01I01B
,y=DEPTO), , color="black", size =4, shape=17)+
  labs(title = "Cuántas personas del hogar vive actualmente en otro país
ahora vs Departamento")+
  theme_minimal()
```

```
datamsc2 <- datamsc[, c(10,1,13, 15, 16)]

reglas <- fim4r(datamsc2, method = "fpgrowth", target = "rules", supp = .1,
conf = .1)

inspect(reglas[])


datamsc2 <- na.omit(datamsc2)
datamsc2
cluster <- kmeans(datamsc2, centers=3)

ggplot(datamsc2, aes(x=P02A02 ,y=DEPTO, color =
as.factor(cluster$cluster)))+
  geom_point()+
  geom_point(data = as.data.frame(cluster$centers), aes(x=P02A02 ,y=DEPTO),
, color="black", size =4, shape=17)+
  labs(title = "Cuantas veces fue victima VS Departamento")+
  theme_minimal()
```

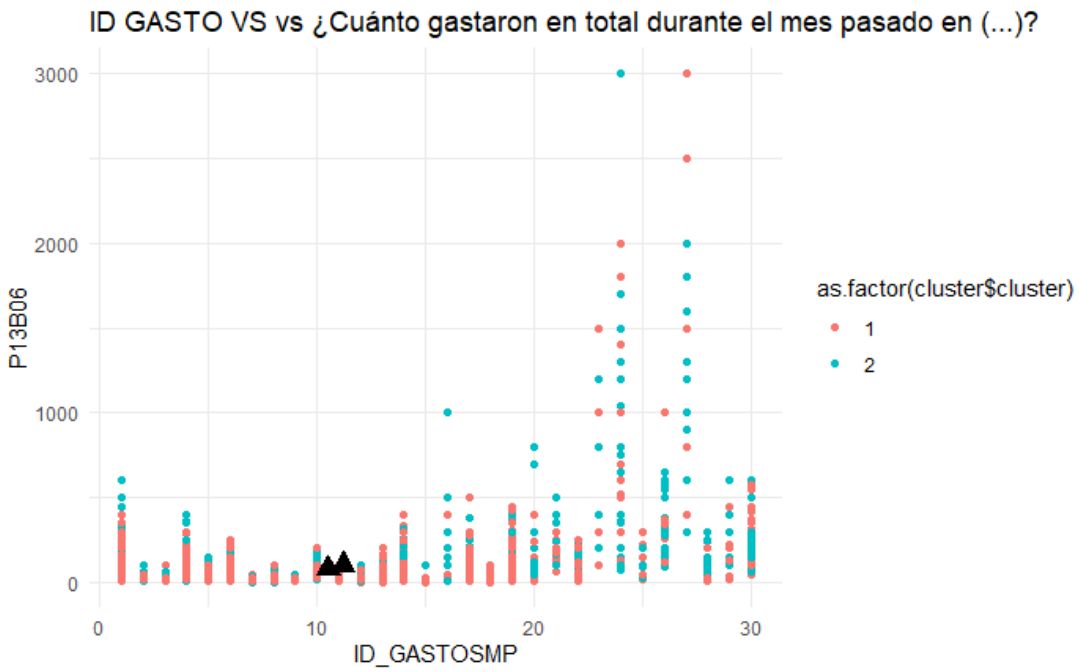


En las siguientes 2 graficas se puede ver bastante claro el contraste del gasto de una persona contra los gastos de una persona no pobre, sobre todo hay un gran espacio vacío en los gastos de las personas de escasos recursos y viene siendo en la parte de recreación y ciertos pagos de servicios como de limpieza o vigilancia

```
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("POCUPA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEI"))]

datamsc <- subset(datagastos, P13B05 == 1)

datamsc <- subset(datamsc, DOMINIO == 3)

datamsc <- subset(datamsc, POBREZA == 1)

datamsc2 <- datamsc[, c(5,10,13)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 50))


datamsc2 <- na.omit(datamsc2)

datamsc2

cluster <- kmeans(datamsc2, centers=2)

ggplot(datamsc2, aes(x=ID_GASTOSMP ,y=P13B06, color =
as.factor(cluster$cluster)))+
  geom_point()+
  geom_point(data = as.data.frame(cluster$centers), aes(x=ID_GASTOSMP
,y=P13B06), , color="black", size =4, shape=17)+
  labs(title = "ID GASTO VS vs ¿Cuánto gastaron en total durante el mes
pasado en (...)?")+
  theme_minimal()
```



```
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("POCUPA"))]
datagastos <- datagastos[ , !(names(datagastos) %in% c("PEI"))]

datamsc <- subset(datagastos, P13B05 == 1)

datamsc <- subset(datamsc, DOMINIO == 1)

datamsc <- subset(datamsc, POBREZA == 3)

datamsc2 <- datamsc[, c(5,10,13)]

reglas <- apriori(datamsc2, parameter = list(support=0.3, confidence=0.3))

inspect(head(sort(reglas, by = "lift", decreasing = TRUE), 20))

datamsc2

datamsc2 <- na.omit(datamsc2)

cluster <- kmeans(datamsc2, centers=2)

ggplot(datamsc2, aes(x=ID_GASTOSMP ,y=P13B06, color =
as.factor(cluster$cluster)))+
  geom_point()+
  geom_point(data = as.data.frame(cluster$centers), aes(x=ID_GASTOSMP
,y=P13B06), , color="black", size =4, shape=17)+
  labs(title = "ID GASTO VS vs ¿Cuánto gastaron en total durante el mes
pasado en (...)?")+
  theme_minimal()
```



5. PROPUESTAS

5.1. Concienciación acerca de la violencia doméstica y laboral:

Es muy difícil hacer una propuesta acerca de la violencia domestica y laboral ya que depende de muchos factores externos, por ejemplo, las principales agresiones son por amenaza en el trabajo que vienen la mayoría por parte de jefes directos y es función de Recursos Humanos evitar que eso pase pero hay muchos casos de corrupción en donde se vuelve muy complicada la situación y en el hogar por la principal cabeza de la familia es quien realiza la agresión, Como propuesta se pueden realizar campañas en ambos ámbitos en redes lo que lo hace mas rentable y accesible a todos para promover un cambio.

5.2. Seguridad en cajeros automáticos

Como propuesta para fortalecer la seguridad en cajeros automáticos se recomienda, siempre revisar el cajero antes de insertar la tarjeta ya que este tipo de prevención evita que seamos víctimas de una estafa o fraude, ha habido campañas acerca de esto mismo y se han fortalecido con el tiempo sobre todo en la publicidad de los bancos.

5.3. Mejora en la distribución de productos de higiene personal en áreas rurales

Mejorar la disponibilidad de productos de limpieza en todas las partes del país, para que siempre puedan obtenerse también seria interesante poder generar productos de bajo costo para este segmento de personas de bajo recursos para que puedan aprovechar de mejor manera su dinero.

6. REFERENCIAS BIBLIGRÁFICAS

Colussi, M. (30 de Julio de 2024). *Violencia en Guatemala: un problema que rebasa la salud mental*. Plaza Publica: https://www.plazapublica.com.gt/content/violencia-en-guatemala-un-problema-que-rebasa-la-salud-mental

Ixcol, C. (5 de julio de 2024). *La migración y las múltiples violencias en Guatemala*. Dialogos GT: https://dialogos.org.gt/la-migracion-y-las-multiples-violencias-en-guatemala/

Sapalú, L. (13 de mayo de 2023). *Estos son los cinco departamentos con mayores índices de migración*. La Hora GT: https://lahora.gt/nacionales/lucero_sapalu/2023/05/13/estos-son-los-cinco-departamentos-con-mayores-indices-de-migracion/

Villanueva., G. S. (2022). *LA VIOLENCIA LABORAL EN GUATEMALA EN EL MARCO DEL CONVENIO 190 DE LA OIT*. GROW genero y trabajo.

7. ANEXOS

7.1. Repositorio

andresdz777/proyecto-mineria-datos