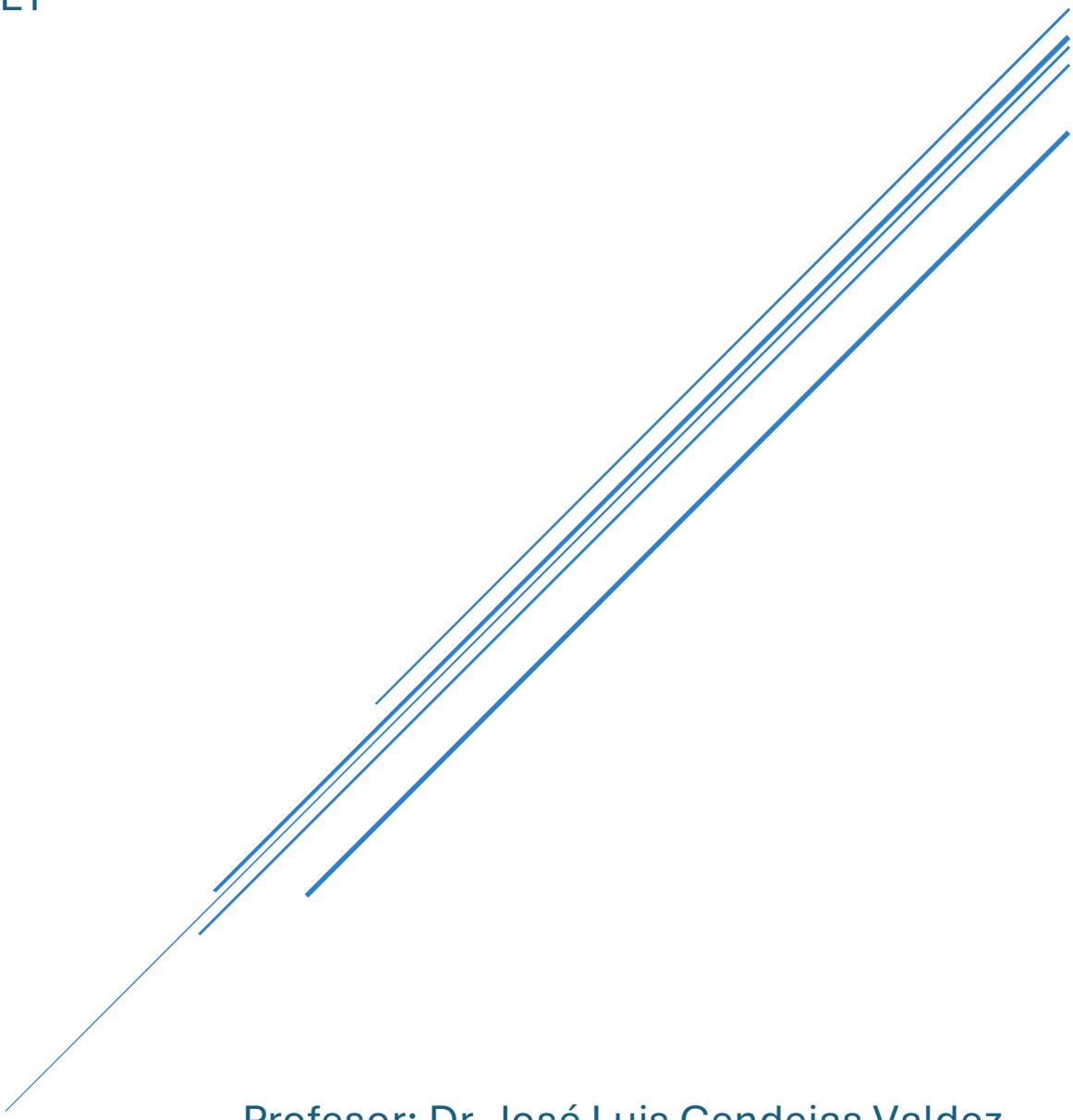
EXTRACCION DE CONOCIMIENTOS DE BASE DE DATOS

PROCESO DE MINERÍA DE DATOS – NETFLIX USERBASE DATASET



Profesor: Dr. José Luis Cendejas Valdez
Alumno: Luis Omar Avalos Ortiz

Etapa 1 y 2 del modelo CRISP-DM

1. Entendimiento del negocio:

En esta etapa, se enfoca en comprender los objetivos comerciales y las necesidades relacionadas con el conjunto de datos de usuarios de Netflix, tales como Identificar los patrones de suscripción, la retención de usuarios y su impacto en los ingresos.

2. Entendimiento de los Datos:

- En esta etapa, se revisarán las columnas del conjunto de datos (por ejemplo, detalles de suscripción, información de la cuenta).
- Identificar valores faltantes o inconsistencias.
- Analizar relaciones entre variables.

Pre – Procesamiento de los datos

En esta fase se preparo el Dataset de usuarios de Netflix con la cantidad de 2500 registros, primero limpiando los datos antes de su análisis, ya que los datos sin procesar pueden contener errores o valores faltantes, lo que puede afectar la calidad de los resultados del análisis y del aprendizaje automático.

UserID	SubscriptionType	MonthlyRevenue	JoinDate	LastPaymentDate	Country	Age	Gender	Device	PlanDuration
1	Basic	10	15-01-22	10-06-23	United States	28	Male	Smartphone	1 Month
2	Premium	15	05-09-21	22-06-23	Canada	35	Female	Tablet	1 Month
3	Standard	12	28-02-23	27-06-23	United Kingdom	42	Male	Smart TV	1 Month
4	Standard	12	10-07-22	26-06-23	Australia	51	Female	Laptop	1 Month
5	Basic	10	01-05-23	28-06-23	Germany	33	Male	Smartphone	1 Month
6	Premium	15	18-03-22	27-06-23	France	29	Female	Smart TV	1 Month
7	Standard	12	09-12-21	25-06-23	Brazil	46	Male	Tablet	1 Month
8	Basic	10	02-04-23	24-06-23	Mexico	39	Female	Laptop	1 Month
9	Standard	12	20-10-22	23-06-23	Spain	37	Male	Smartphone	1 Month
10	Premium	15	07-01-23	22-06-23	Italy	44	Female	Smart TV	1 Month
11	Basic	10	16-05-22	22-06-23	United States	31	Female	Smartphone	1 Month
12	Premium	15	23-03-23	28-06-23	Canada	45	Male	Tablet	1 Month
13	Standard	12	30-11-21	27-06-23	United Kingdom	48	Female	Laptop	1 Month
14	Basic	10	01-08-22	26-06-23	Australia	27	Male	Smartphone	1 Month
15	Standard	12	09-05-23	28-06-23	Germany	38	Female	Smart TV	1 Month
16	Premium	15	07-04-22	27-06-23	France	36	Male	Tablet	1 Month
17	Basic	10	24-01-22	25-06-23	Brazil	30	Female	Laptop	1 Month
18	Standard	12	18-10-21	24-06-23	Mexico	43	Male	Smartphone	1 Month
19	Premium	15	15-02-23	23-06-23	Spain	32	Female	Smart TV	1 Month
20	Basic	10	27-05-23	22-06-23	Italy	41	Male	Tablet	1 Month
21	Premium	15	10-06-23	22-06-23	United States	26	Female	Laptop	1 Month
22	Basic	10	22-07-22	28-06-23	Canada	34	Male	Smartphone	1 Month
23	Standard	12	05-12-21	27-06-23	United Kingdom	49	Female	Smart TV	1 Month
24	Standard	12	03-04-22	26-06-23	Australia	31	Male	Tablet	1 Month
25	Basic	10	14-03-23	28-06-23	Germany	40	Female	Laptop	1 Month
26	Premium	15	12-01-22	27-06-23	France	29	Male	Smartphone	1 Month
27	Basic	10	29-08-22	25-06-23	Brazil	47	Female	Smart TV	1 Month
28	Standard	12	27-09-21	24-06-23	Mexico	33	Male	Tablet	1 Month
29	Premium	15	19-12-22	23-06-23	Spain	36	Female	Laptop	1 Month

TipoSubscripcion	MesesRenovados	Pais	genero	dispositivo	Edad	DuracionPlan
1	1	10	2	3	3	1
2	6	3	1	4	10	1
3	3	9	2	2	17	1
3	3	1	1	1	26	1
1	1	5	2	3	8	1
2	6	4	1	2	4	1
3	3	2	2	4	21	1
1	1	7	1	1	14	1
3	3	8	2	3	12	1
2	6	6	1	2	19	1
1	1	10	1	3	6	1
2	6	3	2	4	20	1
3	3	9	1	1	23	1
1	1	1	2	3	2	1
3	3	5	1	2	13	1
2	6	4	2	4	11	1
1	1	2	1	1	5	1
3	3	7	2	3	18	1
2	6	8	1	2	7	1
1	1	6	2	4	16	1
2	6	10	1	1	1	1
1	1	3	2	3	9	1
3	3	9	1	2	24	1
3	3	1	2	4	6	1
1	1	5	1	1	15	1
2	6	4	2	3	4	1
1	1	2	1	2	22	1
3	3	7	2	4	8	1
2	6	8	1	1	11	1

Análisis de confiabilidad

Para medir el nivel de confiabilidad del Dataset, se aplicó una medida estadística que permitiría determinar qué tan consistentes y confiables serían los datos, para esto la medida a utilizar fue el alfa de Cronbach.

Para confirmar y consolidar esta información se utilizó la herramienta estadística de SPSS, la cual entrego el siguiente resultado.

Fiabilidad

Escala: ALL VARIABLES

Resumen de procesamiento de casos

		N	%
Casos	Válido	2500	100.0
	Excluido[a]	0	.0
	Total	2500	100.0

a. La eliminación por lista se basa en todas las variables del procedimiento.

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
.095	5

Esta herramienta permitió comprobar el nivel de confiabilidad del Dataset, arrojando un 0.95 de fiabilidad, que con la escala de Cronbach comprobamos que la consistencia de los datos era excelente.

Rangos del Alfa de Cronbach

Alfa de Cronbach	Consistencia Interna
$\alpha \geq 0,9$	Excelente
$0,8 \leq \alpha < 0,9$	Buena
$0,7 \leq \alpha < 0,8$	Aceptable

Estudio de correlaciones

Identificación de las correlaciones más altas y/o relevantes del Dataset a través de la herramienta estadística SPSS.

Correlaciones

Correlaciones

		genero	Edad	Pais	TipoSubscripcion	MesesRenovados	dispositivo	DuracionPlan
genero	Correlación de Pearson	1	-.040*	-.004	.014	-.006	.009	.b
	Sig. (bilateral)		.048	.833	.477	.759	.670	.
	N	2500	2500	2500	2500	2500	2500	2500
Edad	Correlación de Pearson	-.040*	1	.021	.009	-.021	-.014	.b
	Sig. (bilateral)	.048		.283	.639	.291	.470	.
	N	2500	2500	2500	2500	2500	2500	2500
Pais	Correlación de Pearson	-.004	.021	1	.200**	.025	-.018	.b
	Sig. (bilateral)	.833	.283		<.001	.213	.374	.
	N	2500	2500	2500	2500	2500	2500	2500
TipoSubscripcion	Correlación de Pearson	.014	.009	.200**	1	-.002	.015	.b
	Sig. (bilateral)	.477	.639	<.001		.926	.449	.
	N	2500	2500	2500	2500	2500	2500	2500
MesesRenovados	Correlación de Pearson	-.006	-.021	.025	-.002	1	-.002	.b
	Sig. (bilateral)	.759	.291	.213	.926		.924	.
	N	2500	2500	2500	2500	2500	2500	2500
dispositivo	Correlación de Pearson	.009	-.014	-.018	.015	-.002	1	.b
	Sig. (bilateral)	.670	.470	.374	.449	.924		.
	N	2500	2500	2500	2500	2500	2500	2500
DuracionPlan	Correlación de Pearson	.b	.b	.b	.b	.b	.b	.b
	Sig. (bilateral)	.	.	.	.	.	.	
	N	2500	2500	2500	2500	2500	2500	2500

*. La correlación es significativa en el nivel 0,05 (bilateral).

**. La correlación es significativa en el nivel 0,01 (bilateral).

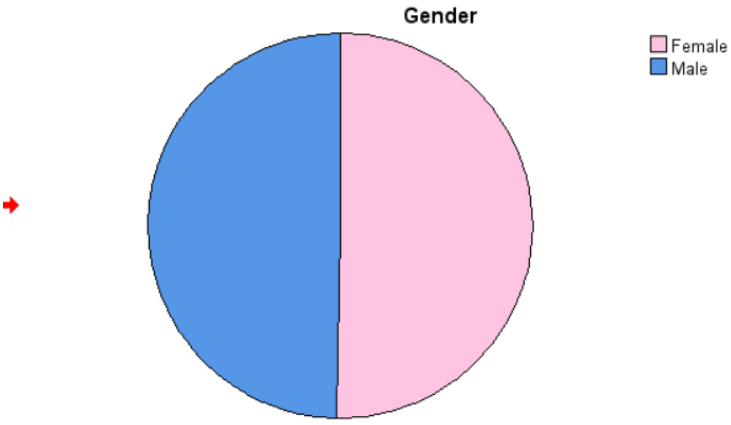
b. No se puede calcular porque, como mínimo, una de las variables es constante.

Predicciones

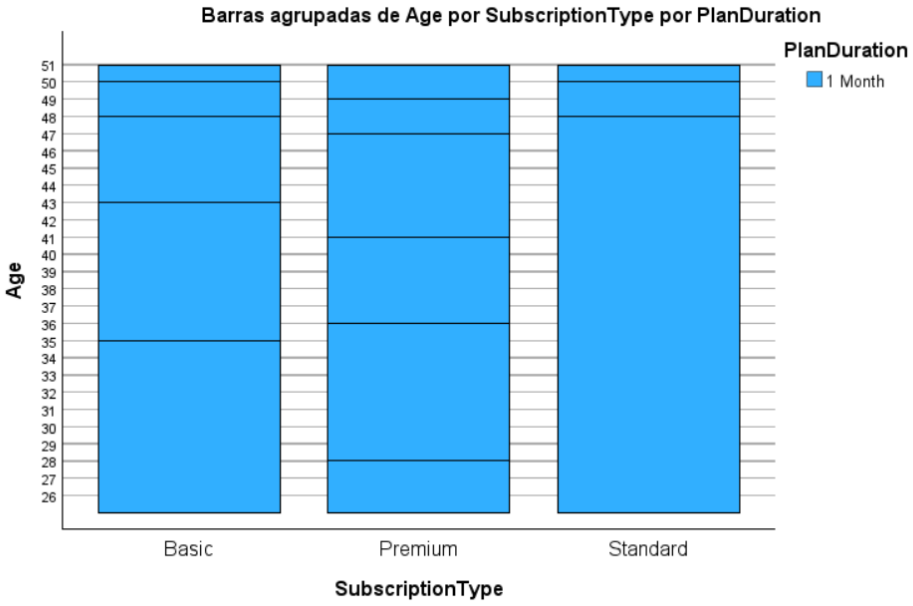
Género y Duración del Plan: Existe una correlación significativa entre el género y la duración del plan. Las personas del género femenino pueden estar más inclinadas a elegir planes de mayor duración.

Gender

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Female	1257	50.3	50.3	50.3
	Male	1243	49.7	49.7	100.0
	Total	2500	100.0	100.0	

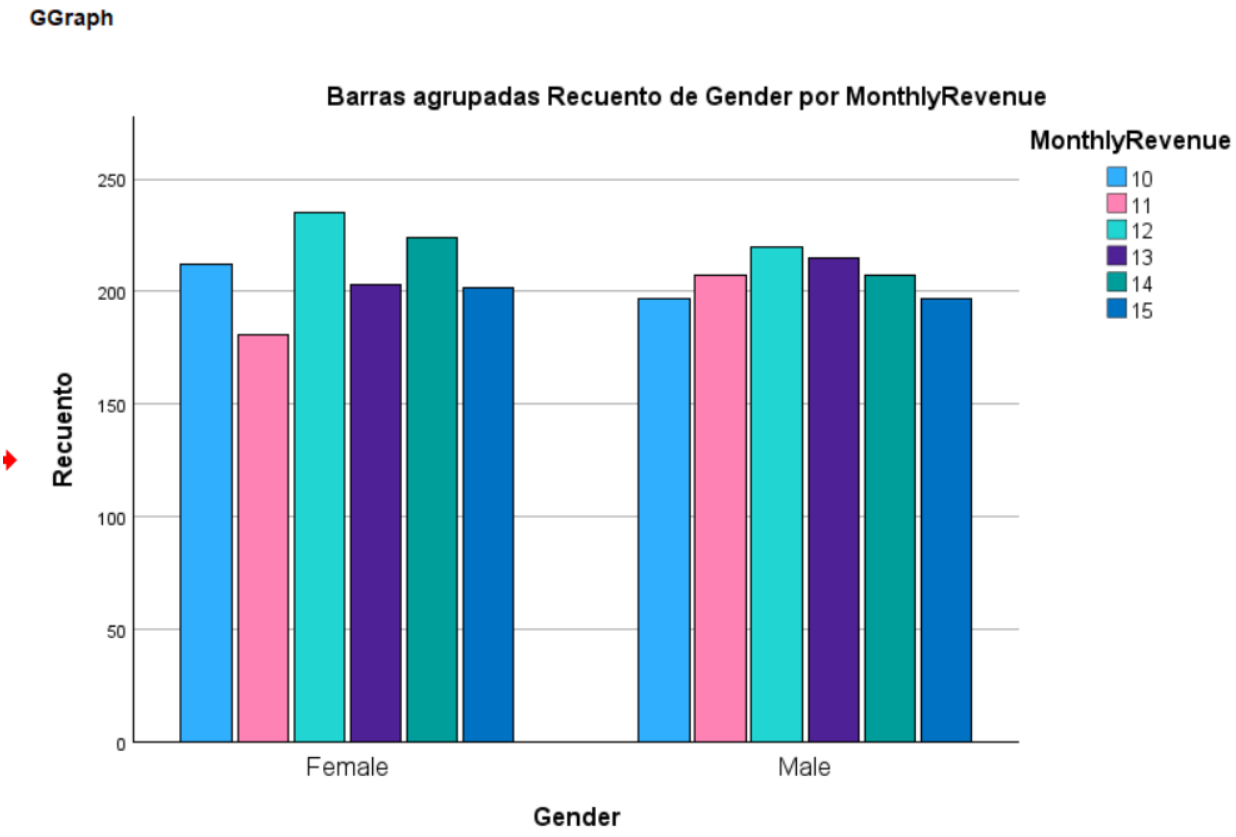


Edad y Tipo de Suscripción: La edad podría estar relacionada con el tipo de suscripción elegido. Los usuarios más jóvenes podrían preferir suscripciones standard, mientras que los mayores optan por planes Premium.



Meses Renovados y Tipo de Suscripción: La cantidad de meses renovados podría estar vinculada al tipo de suscripción. Por ejemplo, los usuarios con suscripciones anuales podrían renovar más meses que aquellos con suscripciones mensuales.

Género y Meses Renovados: El género podría influir en la cantidad de meses renovados. Las personas del género femenino podrían ser más propensos a renovar durante más tiempo.

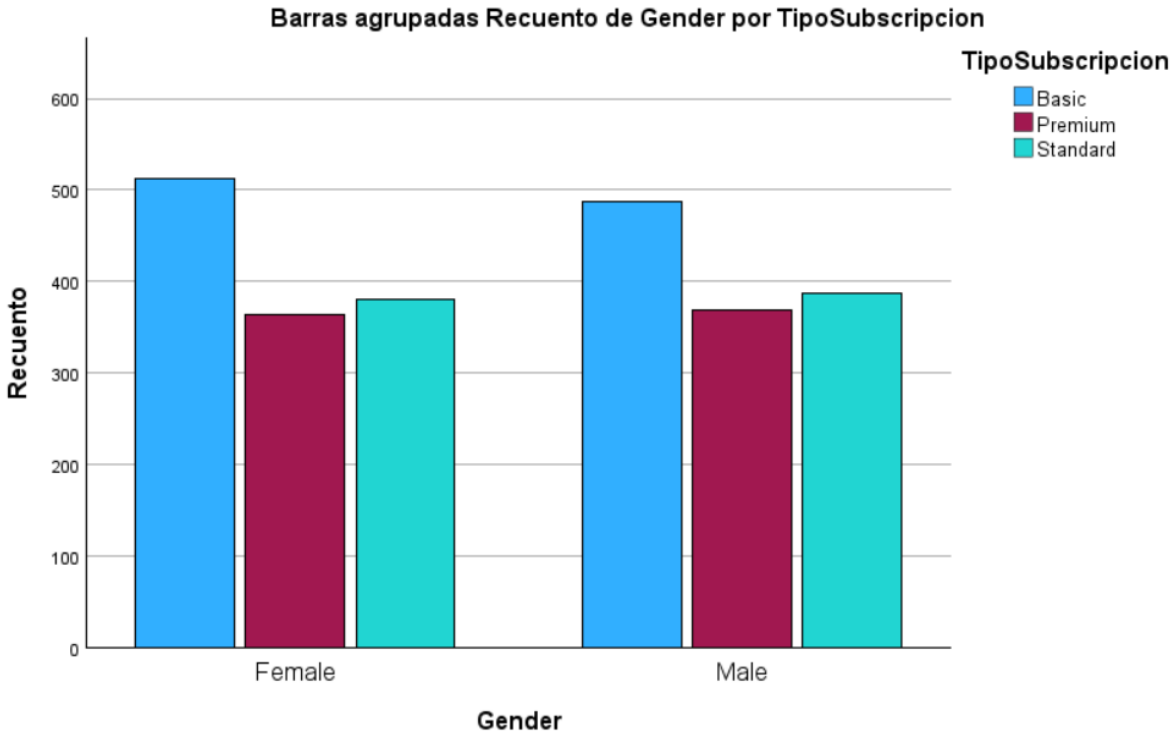


Edad y Meses Renovados: La edad también podría estar relacionada con la cantidad de meses renovados. Por ejemplo, los usuarios más jóvenes podrían renovar durante menos meses que los mayores.

Tipo de Suscripción y Duración del Plan: La elección del tipo de suscripción podría afectar la duración del plan. Por ejemplo, los usuarios con suscripciones anuales podrían optar por planes más largos.

Género y Tipo de Suscripción: Es posible que el género influya en la elección del tipo de suscripción. En este caso las personas del genero masculino se suscriben a menos cuentas básicas, que en el caso del género femenino.

GGraph



Edad y Duración del Plan: La edad podría estar relacionada con la duración del plan. Por ejemplo, los usuarios más jóvenes podrían preferir planes más cortos.

Meses Renovados y Duración del Plan: La cantidad de meses renovados podría estar vinculada a la duración del plan. Por ejemplo, los usuarios con planes más largos podrían renovar durante más meses.

Clustering K – means

Clustering realizado a través de la herramienta estadística SPSS.

Clúster rápido

Centros de clústeres iniciales

	Clúster	
	1	2
TipoSubscripcion	2	2
MesesRenovados	6	1
Pais	10	1
genero	1	2
dispositivo	1	3
Edad	1	26
DuracionPlan	1	1

Historial de iteraciones[a]

	Cambiar en centros de clústeres	
Iteración	1	2
1	8.100	8.259
2	.311	.254
3	.000	.000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para cualquier centro es .000. La iteración actual es 3. La distancia mínimo entre los centros iniciales es 27.129.

Centros de clústeres finales

	Clúster	
	1	2
TipoSubscripcion	2	2
MesesRenovados	4	4
Pais	6	6
genero	2	1
dispositivo	3	2
Edad	7	20
DuracionPlan	1	1