ISTEA

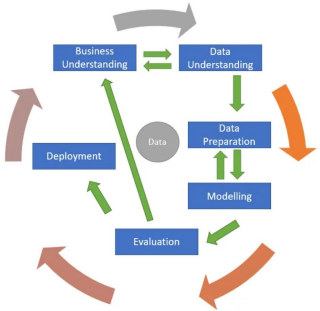
CRIPS-DM

Cross-Industry Standard Process for Data Mining

Data driven

Actores

CRIPS-DM distingue dos niveles



Modelo de Referencia

Guía de usuario

Modelo de Referencia

Diagrama de procesos

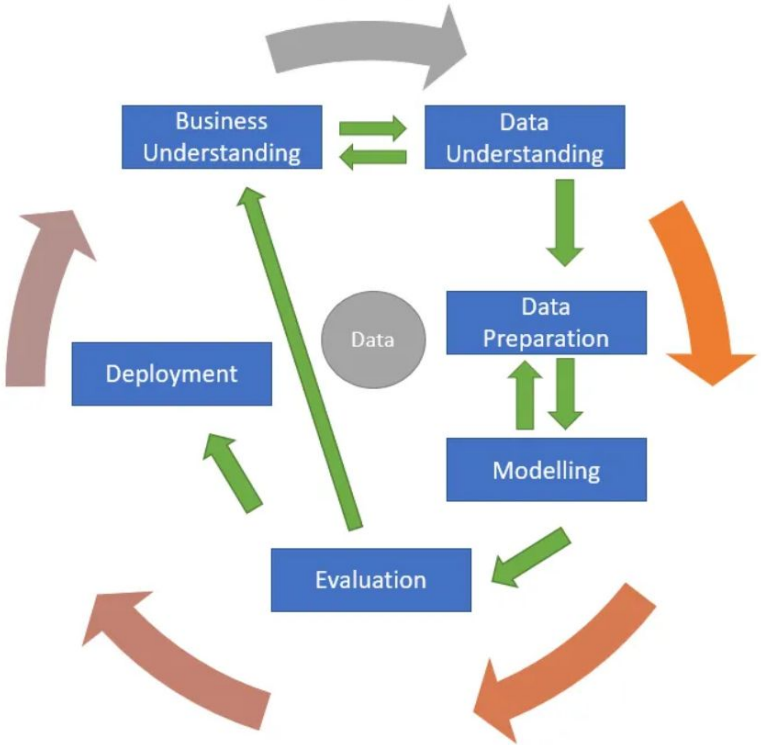
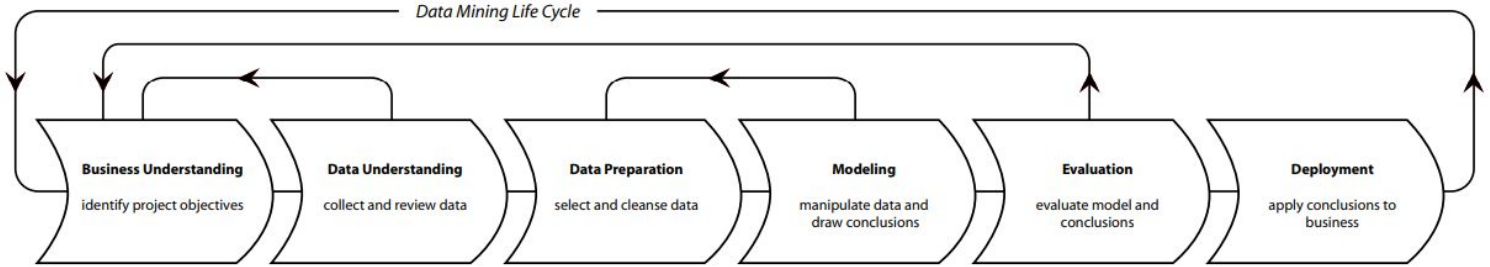


Diagrama de procesos CRIPS-DM

Metodología iterativa para gestionar el ciclo de vida de proyectos de minería de datos orientados a data driven

Proceso y tareas



Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience Documentation
(Log and Report Process)

Ejemplo

Dominio del negocio

- **Sector de la empresa**: aseguradora
- **Departamento**: fraude
- **Alcance**: seguros para:
  - Bienes muebles
  - Bienes inmuebles
  - De vida
  - Automotor

Comprensión de negocio

Las compañías aseguradoras son susceptibles a fraudes porque es muy difícil controlar todas las variables involucradas. Detectar fraudes en las aseguradoras puede ser difícil por varias razones:

Complejidad de los casos
Falta de información completa
Naturaleza subjetiva
Evolución de los métodos de fraude
Volumen de datos
Colaboración limitada
Recursos limitados
Privacidad y regulaciones

Comprensión de negocio

Objetivos

Identificar de manera precisa y eficiente las reclamaciones fraudulentas o engañosas presentadas por los asegurados.
Reducir las pérdidas económicas causadas por pagos indebidos de reclamaciones fraudulentas
Automatizar el proceso de revisión de reclamaciones, permitiendo que los investigadores se enfoquen en casos más prometedores en lugar de revisar todas las reclamaciones manualmente.
Mejorar la experiencia del cliente al procesar más rápidamente las reclamaciones legítimas.
Aprender de nuevos patrones y técnicas de fraude, lo que mejora su capacidad para detectar fraudes en evolución.

Comprensión de negocio

Contexto comercial

La aseguradora contrata a su equipo de Científicos de Datos para trabajar como consultores en el departamento de Fraude. Actualmente, las pólizas que se emiten se controlan y cualquier reclamo presentado se examina y evalúa, de manera manual, para determinar la legitimidad y la aprobación final para el pago por parte de la compañía de seguros.

Es función del equipo de fraude determinar qué reclamos presentados deben aprobarse y cuáles deben rechazarse.

Comprensión de negocio

Problema comercial

La tarea de su equipo es responder la siguiente pregunta:

¿Existen patrones particulares en los grupos de reclamos presentados que puedan ser indicativos de fraude?

Comprensión de negocio

Contexto analítico

El departamento de finanzas les ha proporcionado datos sobre todas las reclamaciones recientes realizadas por **1000 clientes**.

Los datos no están etiquetados; es decir, no hay una variable que nos diga cuáles de estas afirmaciones son fraudulentas o no.

Comprensión de los datos

Suponga que los datos que se encuentran en el siguiente archivo se corresponden a los datos que proporcionó el departamento de finanzas de la aseguradora.

Datos

Comprensión de los datos

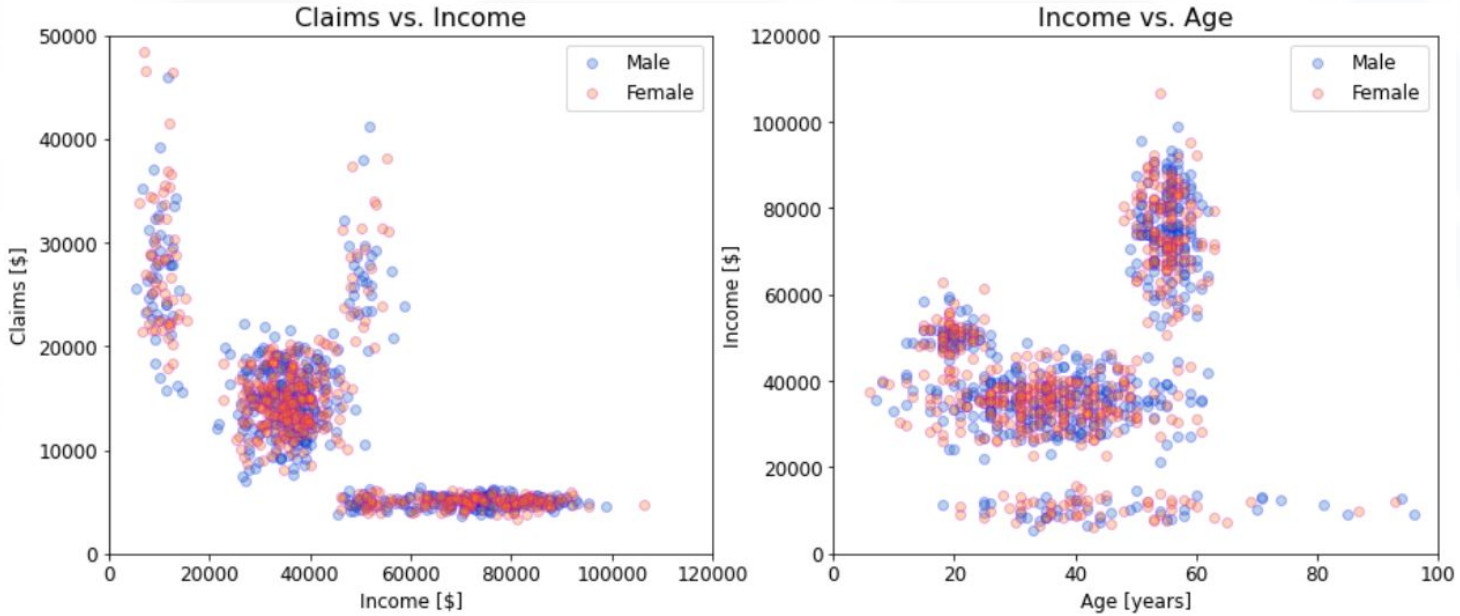
Muestra de los datos

case_id	income	age	claims
59982	35750.95	42	9518.95
87249	24078.27	19	19354.23
50406	39241.52	37	13056.04
59391	33248.31	26	19238.37
96622	38649.96	54	14427.42

- **case_id**: número de caso en la aseguradora. Se refiere al número único que identifica el expediente del caso.
- **income**: Ingresos mensuales de los clientes expresados en moneda local.
- **age**: edad del cliente en años
- **claims**: monto del reclamo por siniestro expresado en moneda local.

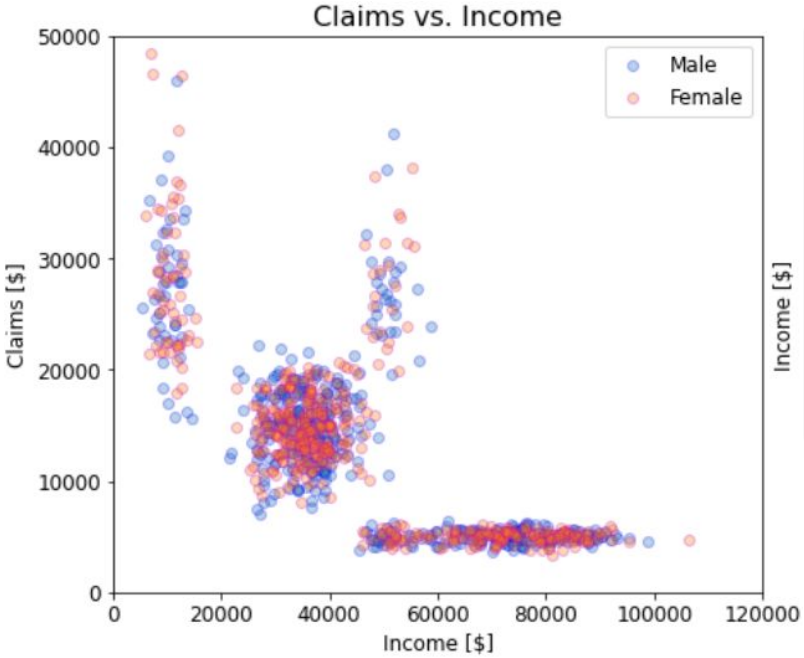
Comprensión de los datos

Análisis exploratorio inicial



Comprensión de los datos

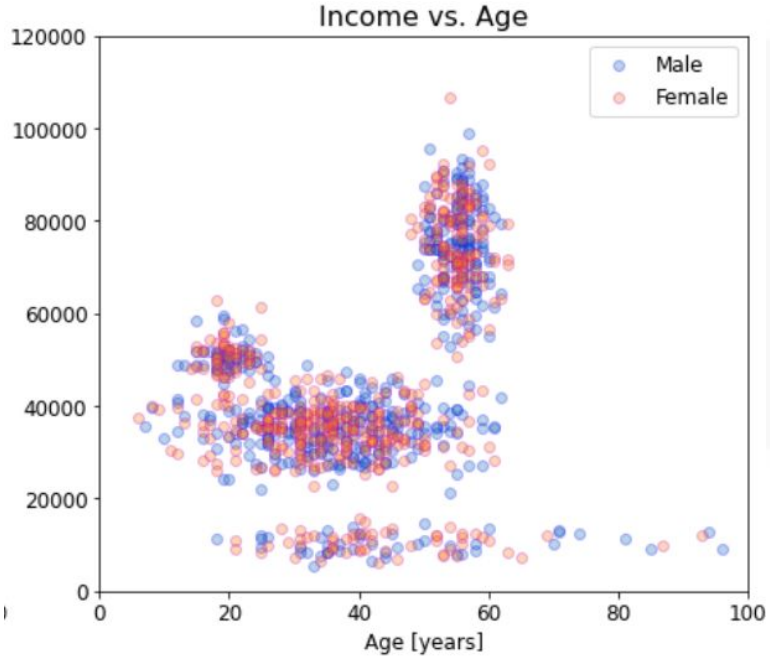
Análisis exploratorio inicial



- No parece haber grandes diferencias en las distribuciones basadas en el género.
- Grupo grande de reclamos alrededor del rango de ingresos de 30.000-45.000 (Valores típicos de ingreso)
- Una franja de reclamaciones con ingresos de entre 50.000- 100.000 que valen 5.000 aproximadamente. No está claro exactamente qué son, pero podrían ser cosas cotidianas con las que las personas más pudientes pueden lidiar (por ejemplo, reclamos por accidentes automovilísticos).
- Franja de reclamos por al menos 20.000 entre las personas que ganan solo 10.000, lo cual es inusual y bien puede consistir en **reclamos fraudulentos**.

Comprensión de los datos

Análisis exploratorio inicial

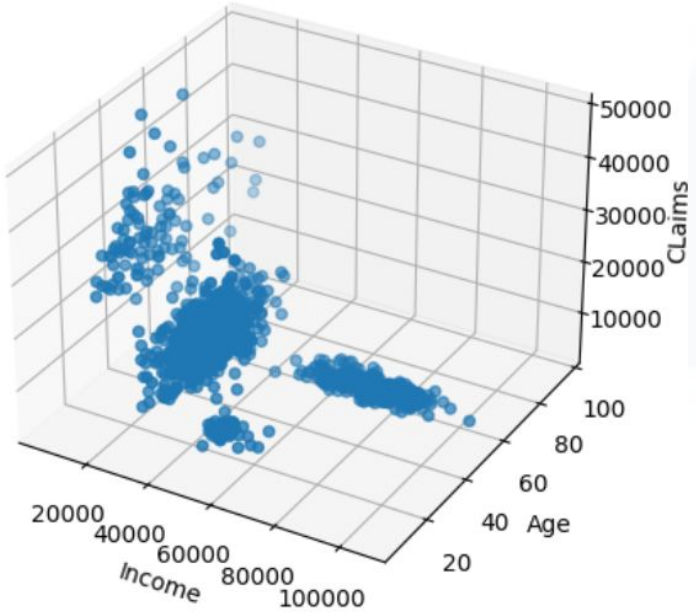


- Hay una franja de personas que ganan 10.000 en todas las edades (salario mínimo).
- Hay un gran grupo de personas que ganan entre 30.000-40.000 en todas las edades (salario medio)
- Hay muchas personas de ingresos más altos (60.000-100.0000) justo antes de los 60 años. A la edad de 59 es cuando las personas en los EE.UU. pueden comenzar a sacar ahorros de sus cuentas de jubilación, por lo que esto puede tener algo que ver con este patrón.

Comprensión de los datos

Análisis exploratorio inicial

Parece que los datos se concentran en 4 grupos diferenciables.



Preparación de los datos

Escalado de los datos:

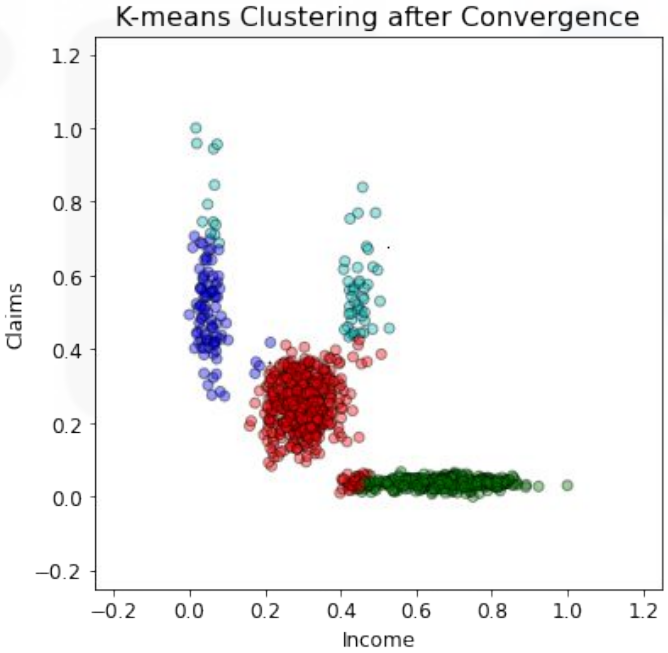
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

	age	income	claims
0	0.400000	0.301264	0.137252
1	0.144444	0.185900	0.355004
2	0.344444	0.335762	0.215563
3	0.222222	0.276530	0.352439
4	0.533333	0.329915	0.245925

Modelado

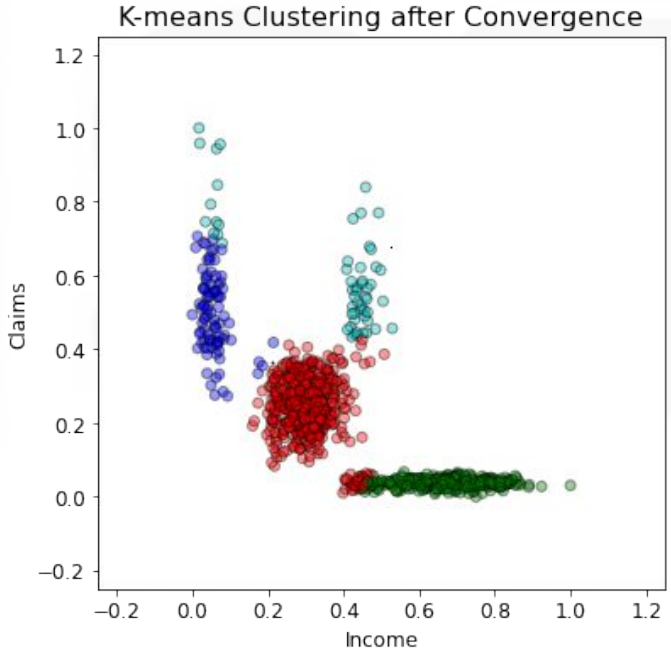
Algoritmo de agrupamiento k-mean:

Se entrena un algoritmo de k-means con k = 4



Modelado

Análisis preliminar:

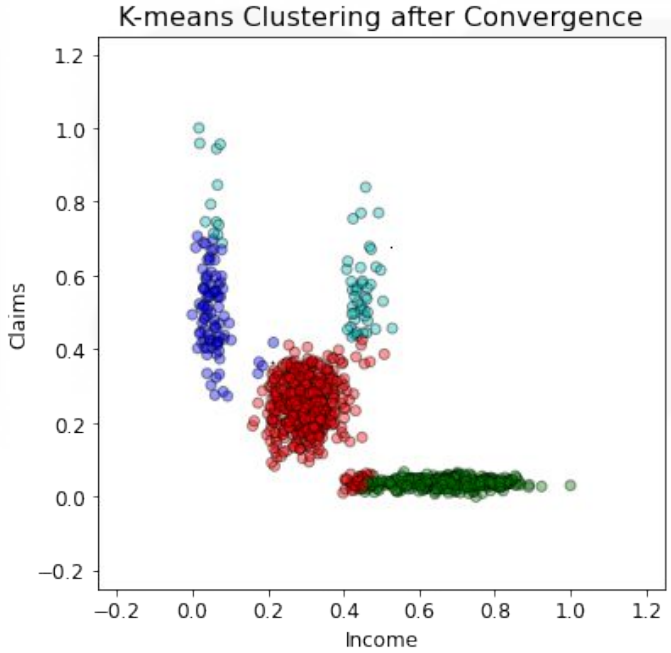


Grupo #1: **ingresos altos y reclamos bajos**, que probablemente sean reclamos ordinarios hechos por familias acomodadas. Es muy probable que estos no sean fraudulentos y que la aseguradora los acepte.

Grupo #2: ingresos moderados con valores de reclamación moderados. Estos son bastante abundantes y podrían ser elementos cotidianos como reclamos de automóviles. Lo más probable es que se deban aceptar.

Modelado

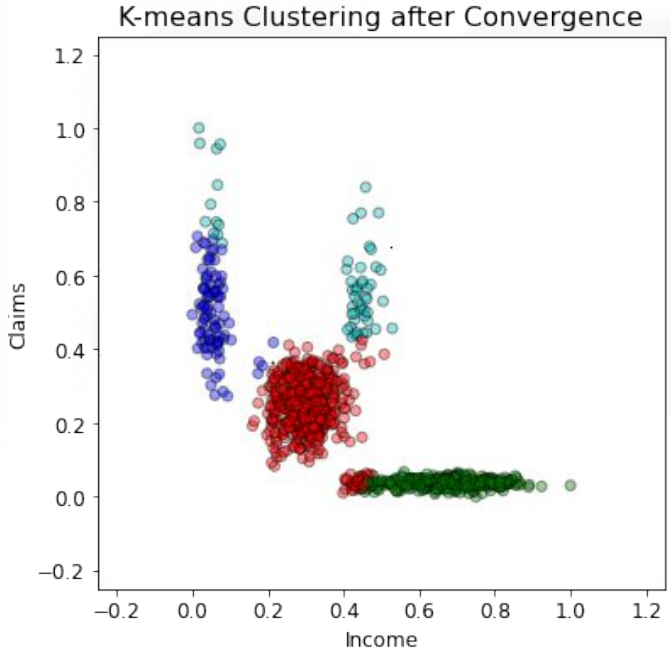
Análisis preliminar:

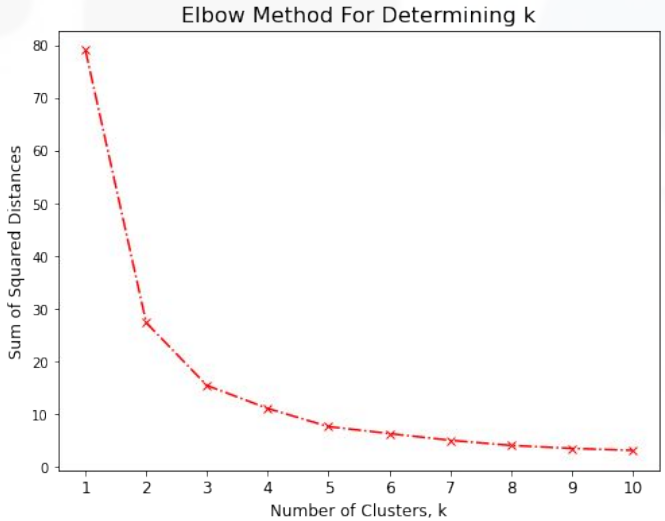


Grupo #3: **ingresos moderados y las altas reclamaciones.** Esto podría ser plausible si es algo que las personas de ingresos medios necesitan pero que no siempre pueden pagar, pero hay que recomendar investigar esto más a fondo.

Grupo #4: ingresos bajos pero reclamaciones muy elevadas. Estos claramente no son asequibles y podrían ser intentos de obtener efectivo gratis. Hay que recomendar rechazar.

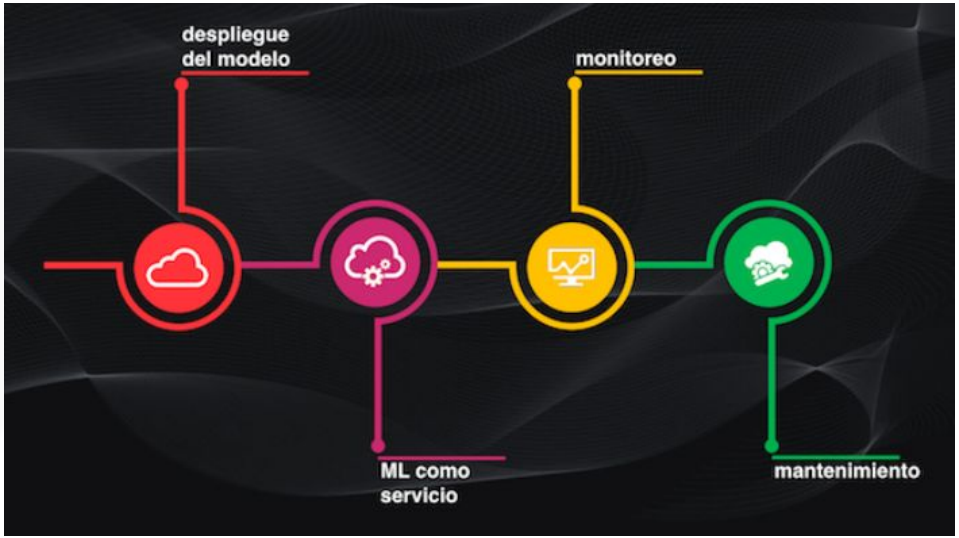
Evaluación


K-means Clustering after Convergence
Claims
Income
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Elbow Method For Determining k
Sum of Squared Distances
Number of Clusters, k
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10

Despliegue

- Plan de puesta en producción
- Plan de monitoreo
- Plan de mantenimiento



Guía de Usuario

Términos modernos

- Data acquisition
- Data wrangling
- EDA
- Preprocesing
- Feature engineering
- …

En la próxima sesión veremos…

Introducción a la Minería de Datos y Conceptos Fundamentales:

- Introducción a la Minería de Datos: definición, objetivos y aplicaciones.
- Proceso de Minería de Datos: desde la selección de datos hasta la evaluación de modelos.
- Conceptos de conjunto de datos, atributo, instancia y clases.
- Tipos de datos: numéricos, categóricos y mixtos.
- Preprocesamiento de datos: limpieza, transformación y reducción de dimensiones.

Preguntas