



MACHINE LEARNING





Logro Unidad 1

Al finalizar la unidad, el alumno es capaz de aplicar adecuadamente técnicas de pre procesamiento de datos para posibilitar la implementación de una solución de Machine Learning para un problema del mundo real.



Contenido 1

- Definición, importancia y aplicaciones de Machine Learning
- Técnicas de aprendizaje
- Proceso de extracción de conocimiento (KDD) y su relación con Machine Learning
- Ciclo de vida de un proyecto de Machine Learning



Contenido 1

- Definición, importancia y aplicaciones de Machine Learning
- Técnicas de aprendizaje
- Proceso de extracción de conocimiento (KDD) y su relación con Machine Learning
- Ciclo de vida de un proyecto de Machine Learning



Data Mart

Almacén de datos a un área específica de la organización.





Data Warehouse

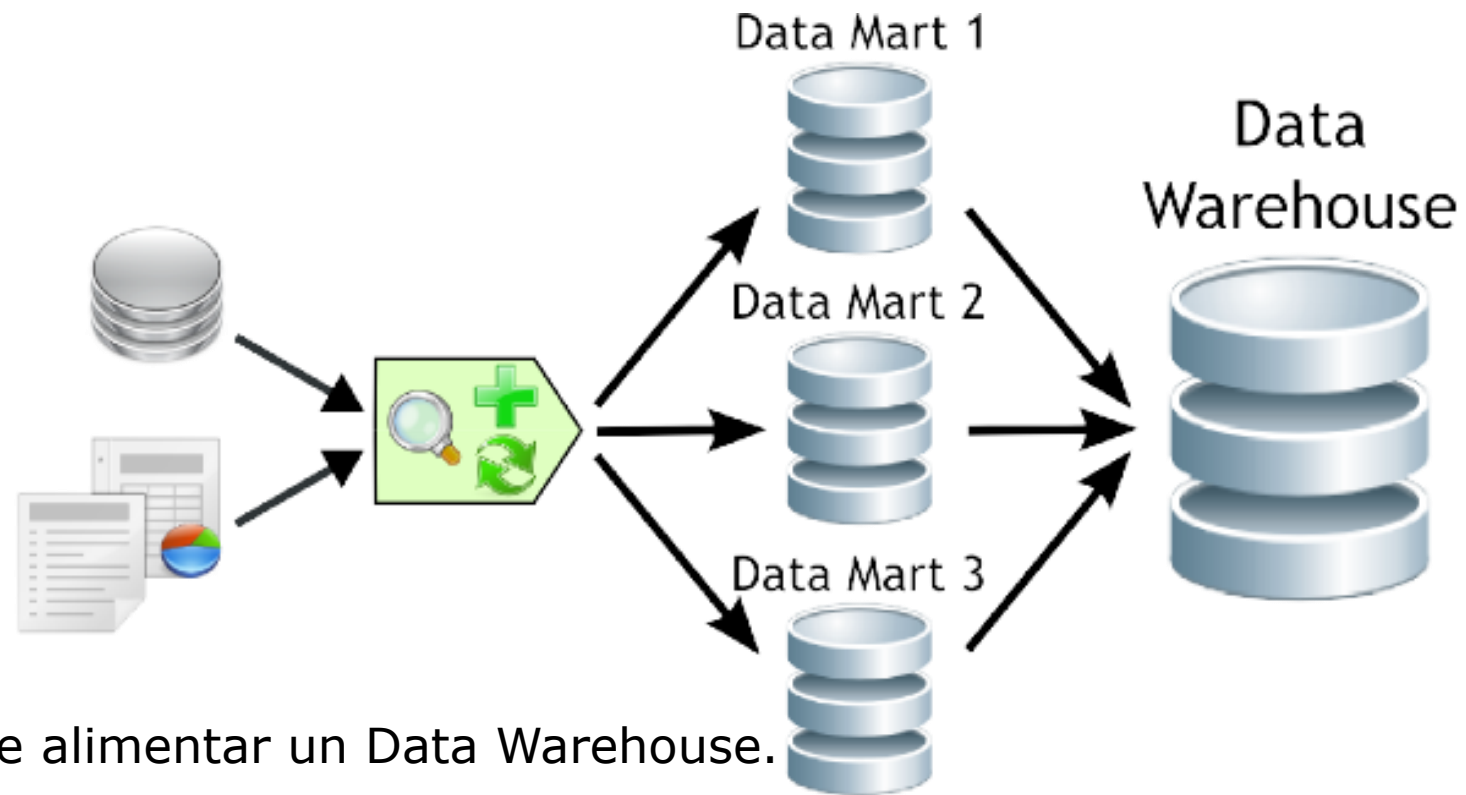
Es el repositorio central de la organización.



Los datos se extraen de diferentes sistemas operacionales o fuentes externas.



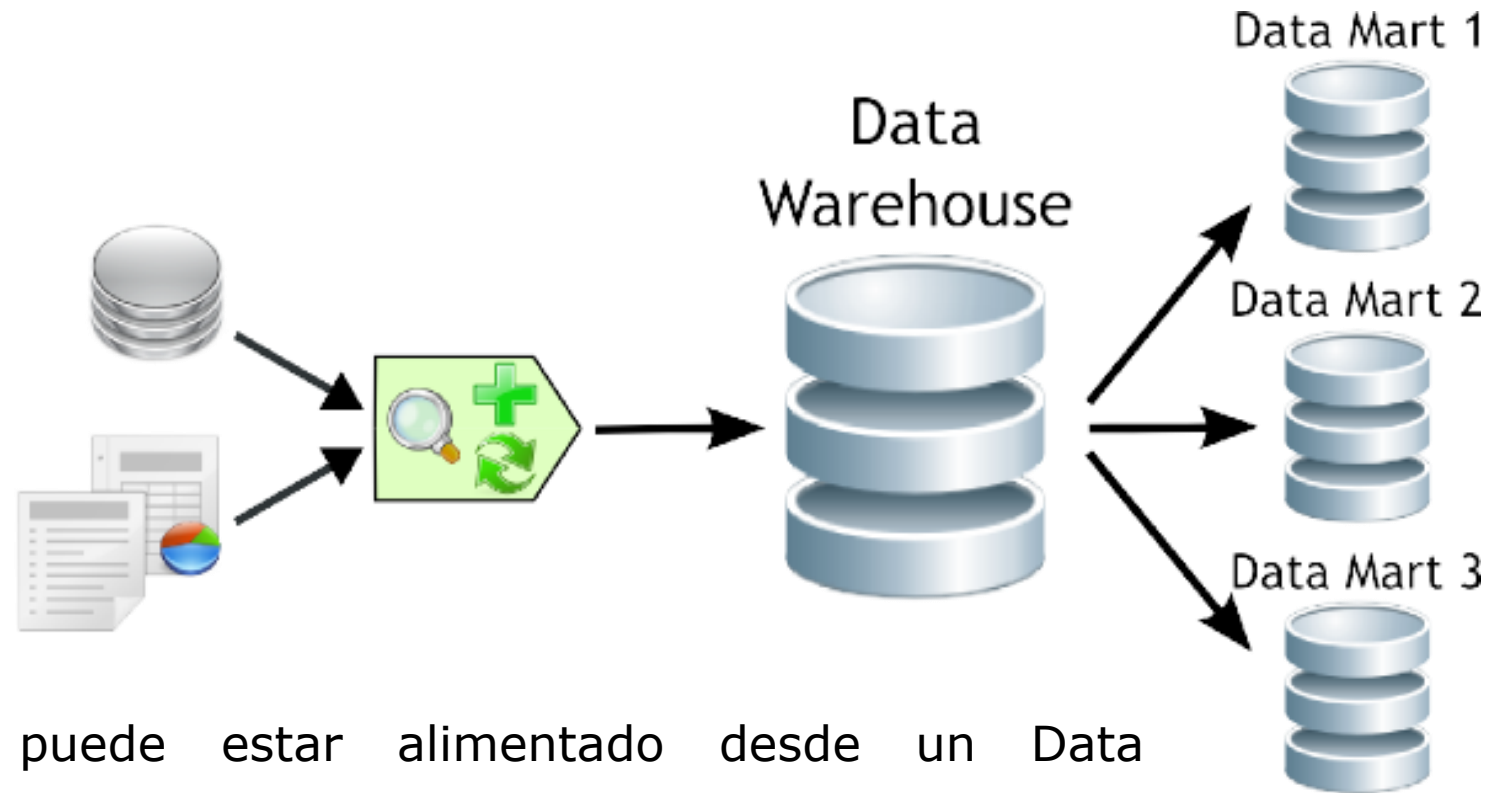
Data Mart vs Data Warehouse



El Data Mart puede alimentar un Data Warehouse.



Data Mart vs Data Warehouse

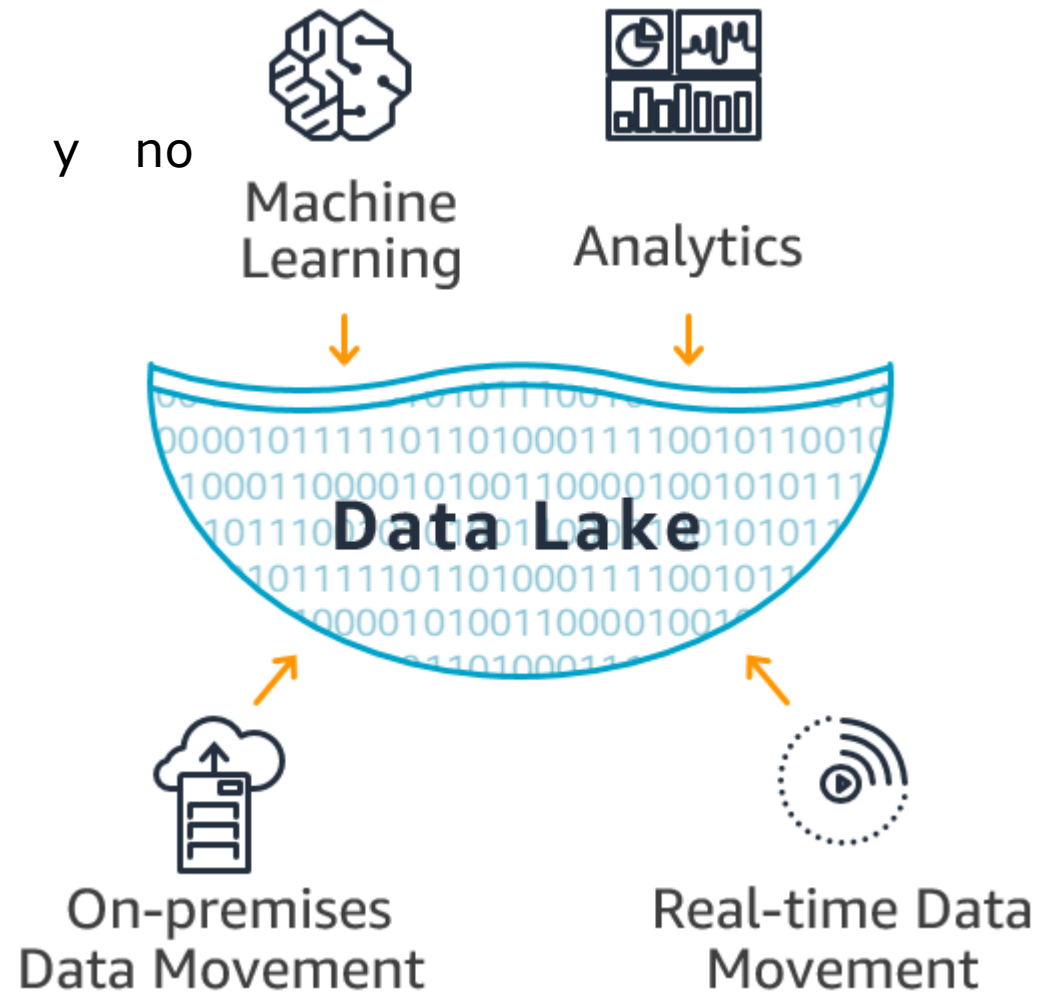


El Data Mart puede estar alimentado desde un Data Warehouse.



Data Lake

Almacena la información estructurada y no estructurada de manera centralizada.





Data Warehouse vs Data Lake

Características	Data Warehouse	Data Lake
Data	Sistemas transaccionales y bases de datos operativas	Base de datos relacional y no relacional desde dispositivos IoT, aplicaciones móviles, redes sociales, etc
Costo	Consultas rápidas con almacenamiento a mayor costo	Consultas rápidas con almacenamiento a menor costo
Calidad	Datos procesados	Datos sin procesar
Usuarios	Analistas de negocios	Analistas de negocios, Científicos de datos, Ingenieros de datos



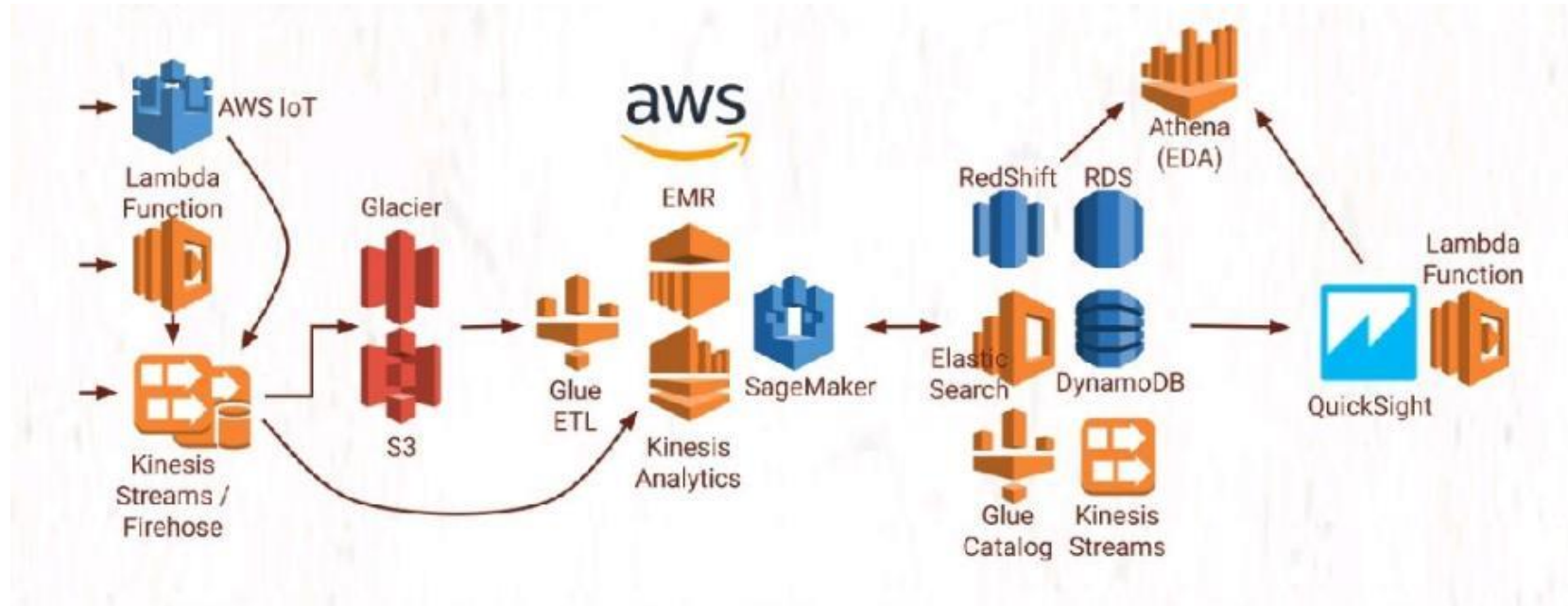
Proveedores



HUAWEI CLOUD

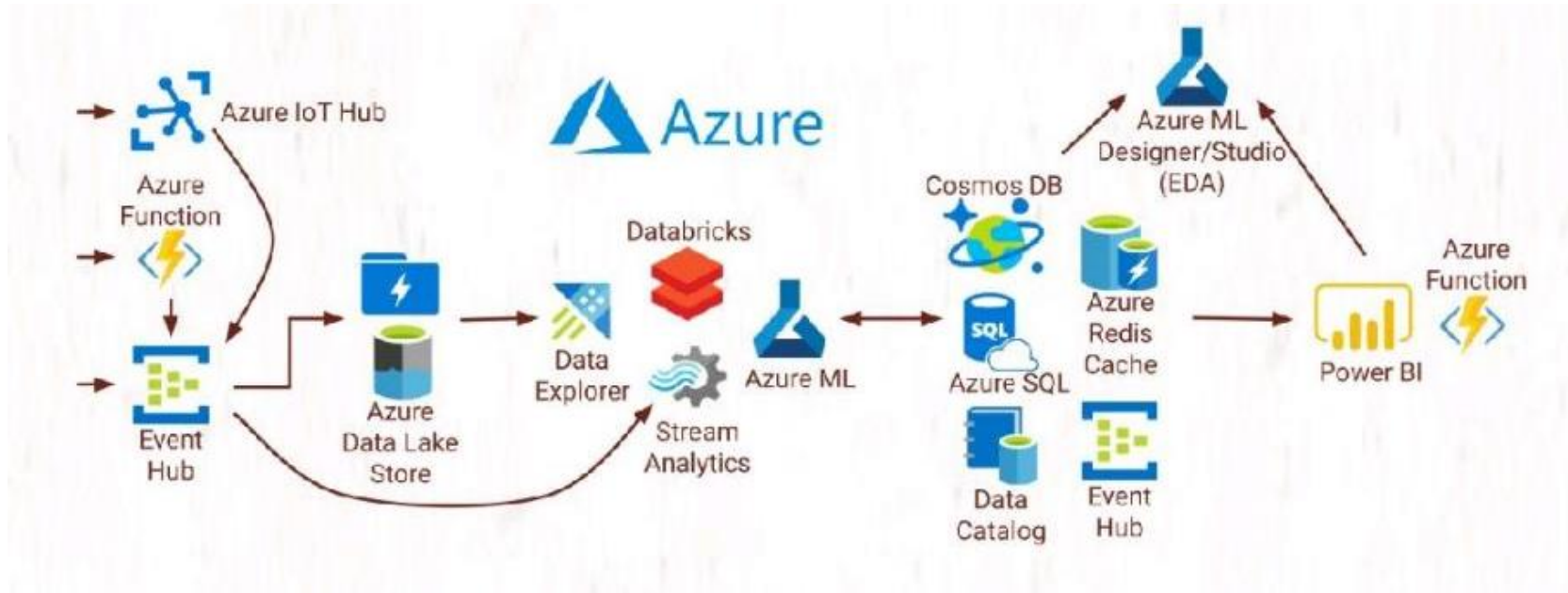


Arquitectura



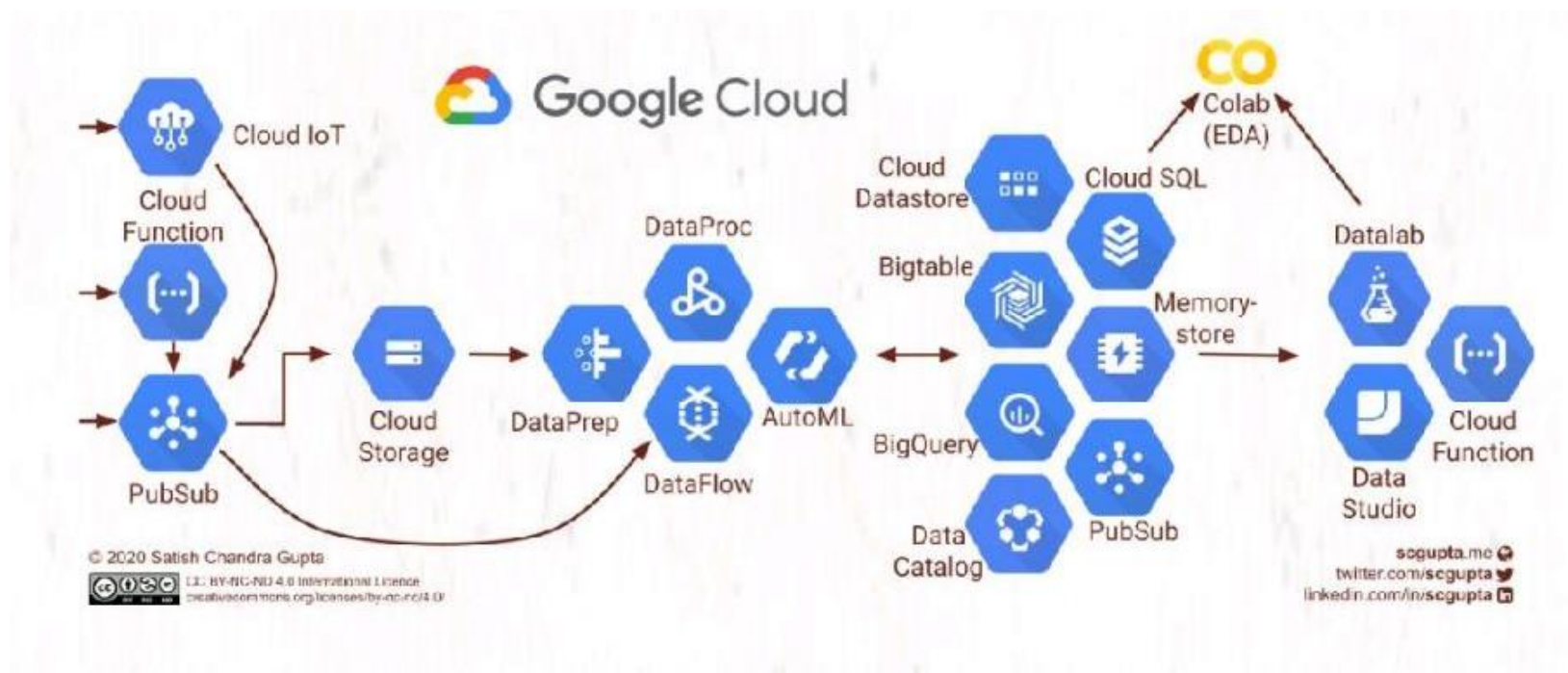


Arquitectura





Arquitectura





Healthcare Perú

Caso de aplicación

Health Care Perú es una empresa del rubro salud.

- La información clínica sobre los exámenes médicos y otros datos relevantes se encuentran en el local donde me atendí.
- Si deseo atenderme en otro local diferente, me deben registrar nuevamente o solicitar la historia clínica al otro local.

Como responsable del área de Data Analytics,

- 1.¿Qué sugiere realizar para mejorar la atención de los pacientes?
- 2.¿Por qué tendría que aceptar su propuesta?
- 3.¿Cómo lo va a llevar a cabo?
- 4.¿Qué necesita de parte de Health Care Perú?



Preguntas de negocio

Telecom

¿Cuál es la probabilidad de que un cliente abandone la compañía?

Retail

¿Quiénes son mis clientes potenciales?

Seguros

¿Qué clientes debo enviar a que realicen un examen médico para emitir un seguro?

Banca

¿Quiénes son los clientes que no pagarán los préstamos?

Streaming

¿Qué le recomiendo a los usuarios que miran películas de acción?



Formular 2 preguntas de negocio

Telecom

AAA

Retail

AAA

Seguros

AAA

Banca

AAA

Streaming

AAA



El Negocio y Data Analytics

El Personal del Negocio

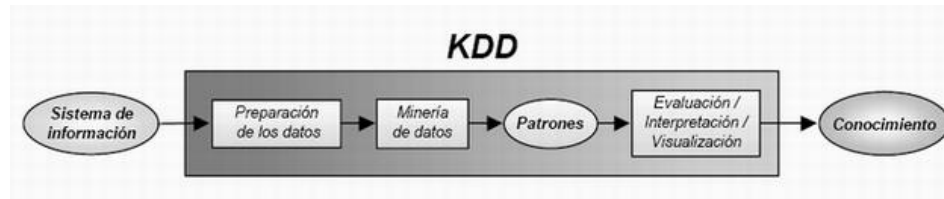
Debe de expresar de manera clara lo que desea

El Personal de Data Analytics

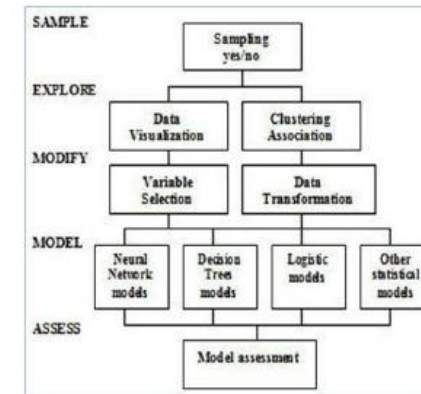
Debe de entender lo que están pidiendo



Metodologías



SEMMA



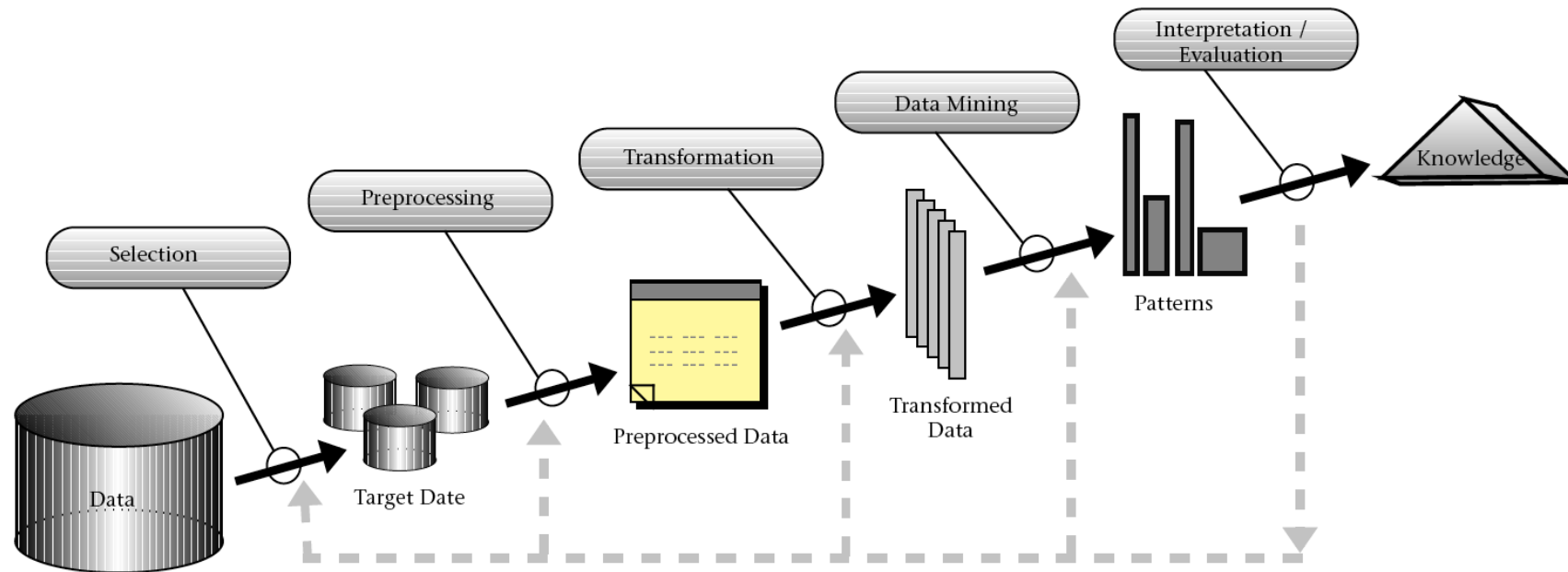
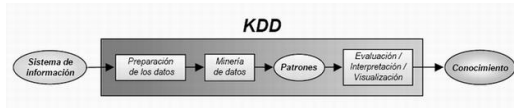
CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto

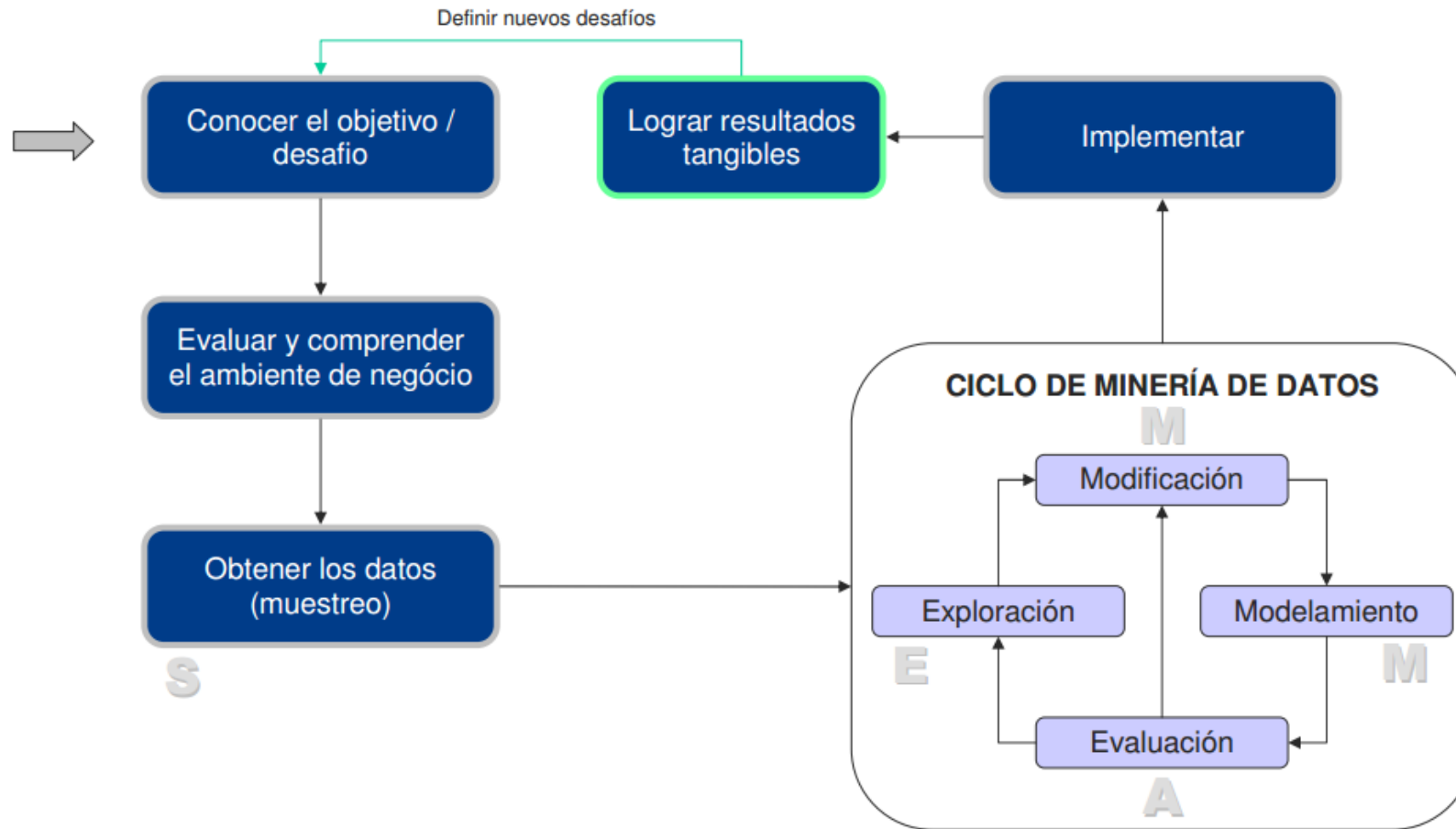
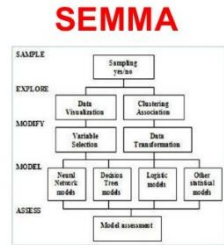


Metodologías





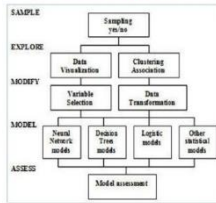
Metodologías





Metodologías

SEMMA



Sample (muestreo)

■ Ejemplo conceptual



Población total

Registros:

Miliones de registros

Variable:

Monto de Compras

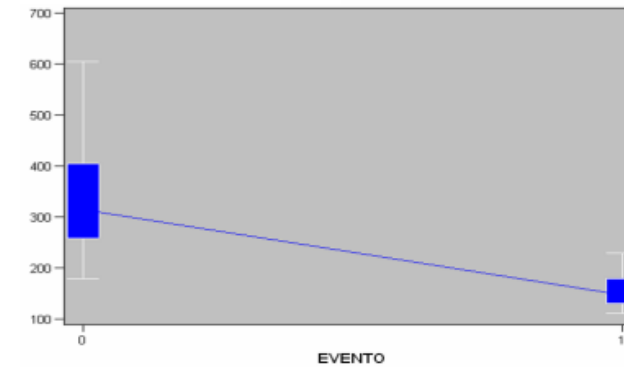
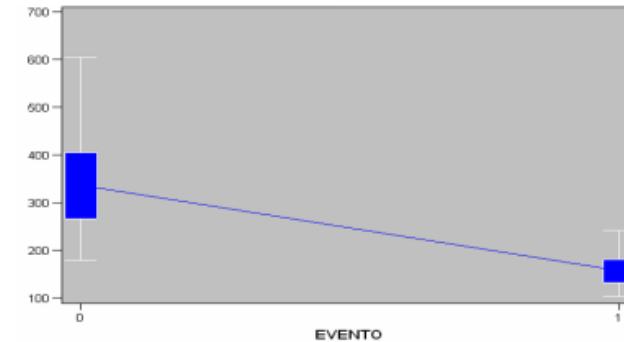
Muestreo

Registros:

5% de los registros

Variable:

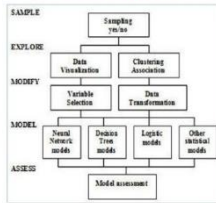
Monto de Compras





Metodologías

SEMMA



xplore (exploración)

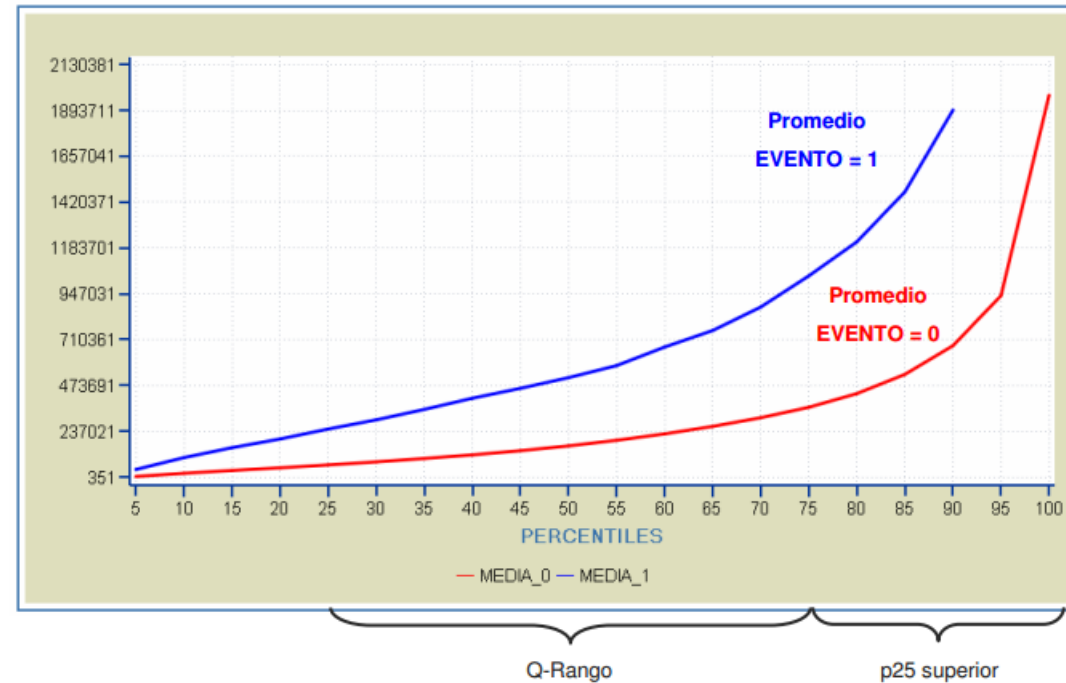
Variables CONTÍNUAS

Line Plot

MONTO_COM_UAV - PROMEDIO

MONTO TOTAL DE LAS COMPRAS EN UAV

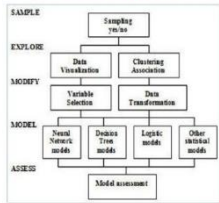
Ejemplo





Metodologías

SEMMA



E xplore (exploración)

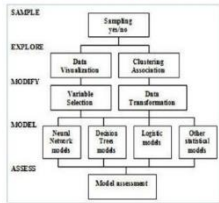
ANÁLISIS DE CORRELACIÓN ($\rho_{x,y}$) – Evitar Multicolinearidad

- ✓ Si $|\text{MAX}(\rho_{x,y})| \geq 0,85$ Seleccionar 'X' o 'Y' para seguir en el test
- ✓ Si $0,30 \leq |\text{MAX}(\rho_{x,y})| < 0,85$ Combinar variables
- ✓ Si $|\text{MAX}(\rho_{x,y})| < 0,30$ Siguen el desarrollo del modelo



Metodologías

SEMMA



Modify (modificación)

■ Generación de variables nuevas

- Resumir información (Componentes Principales)
- Variables de Tendencia
- Tratamiento de Outliers
- Tratamiento de Missings (¿Missing=Cero?)
- Generación de nuevas variables. No abusar!

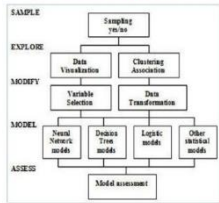
■ Discretización de variables

- Tratar outliers
- Maximizar la correlación con la variable respuesta (Con Árboles por ejemplo)
- Conocer correlación entre covariables de mismo tipo (Discretas vs Continuas)
- Percibir efectos no lineales
- Más fácil interpretar e explicar el modelo
- Aumenta la estabilidad del modelo en el tiempo
- Encontrar equilibrio entre las categorías



Metodologías

SEMMA

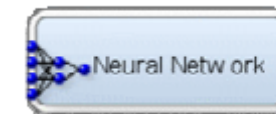
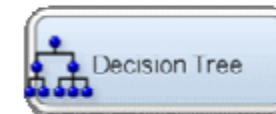


Model (modelado)

Modelo de propensión

Una vez cumplida las etapas anteriores de exploración, discretización, análisis de correlación y selección de las variables se sigue con el modelamiento de los datos.

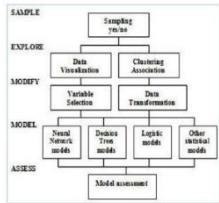
Probar distintas técnicas con distintos parámetros y compararlos





Metodologías

SEMMA



Assess (evaluación)

Indicadores de Calidad del Modelo:

▪ Criterio KS

Este criterio se basa en la comparación entre las distribuciones de probabilidad acumulado de los clientes clasificados como “evento” y “no evento”. Buscamos, entonces, la mayor diferencia observada entre estos dos grupos. Esta distancia (valor del criterio KS) puede fluctuar entre 0 y 1 y cuanto más próximo de uno mejor es el ajuste del modelo.

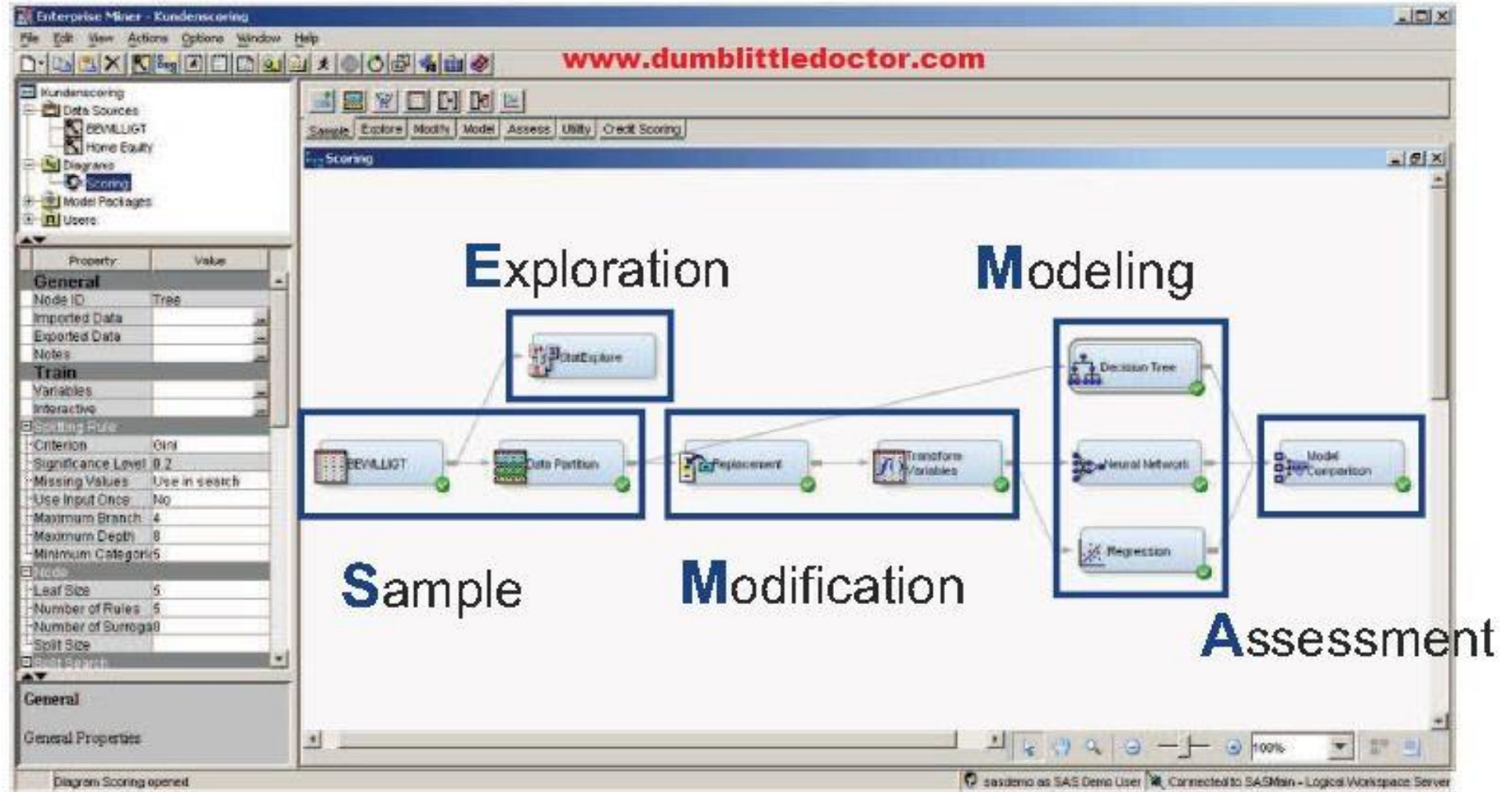
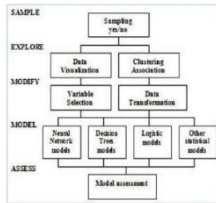
▪ Criterio ROC

Es una curva de la tasa de **verdadero-positivos** (sensibilidad) versus la tasa de **falso-positivo** (1 – especificidad). El área bajo la curva es el ROC. Puede fluctuar entre 0.5 y 1 y cuanto más próximo de 1 mejor es el ajuste del modelo.



Metodologías

SEMMA





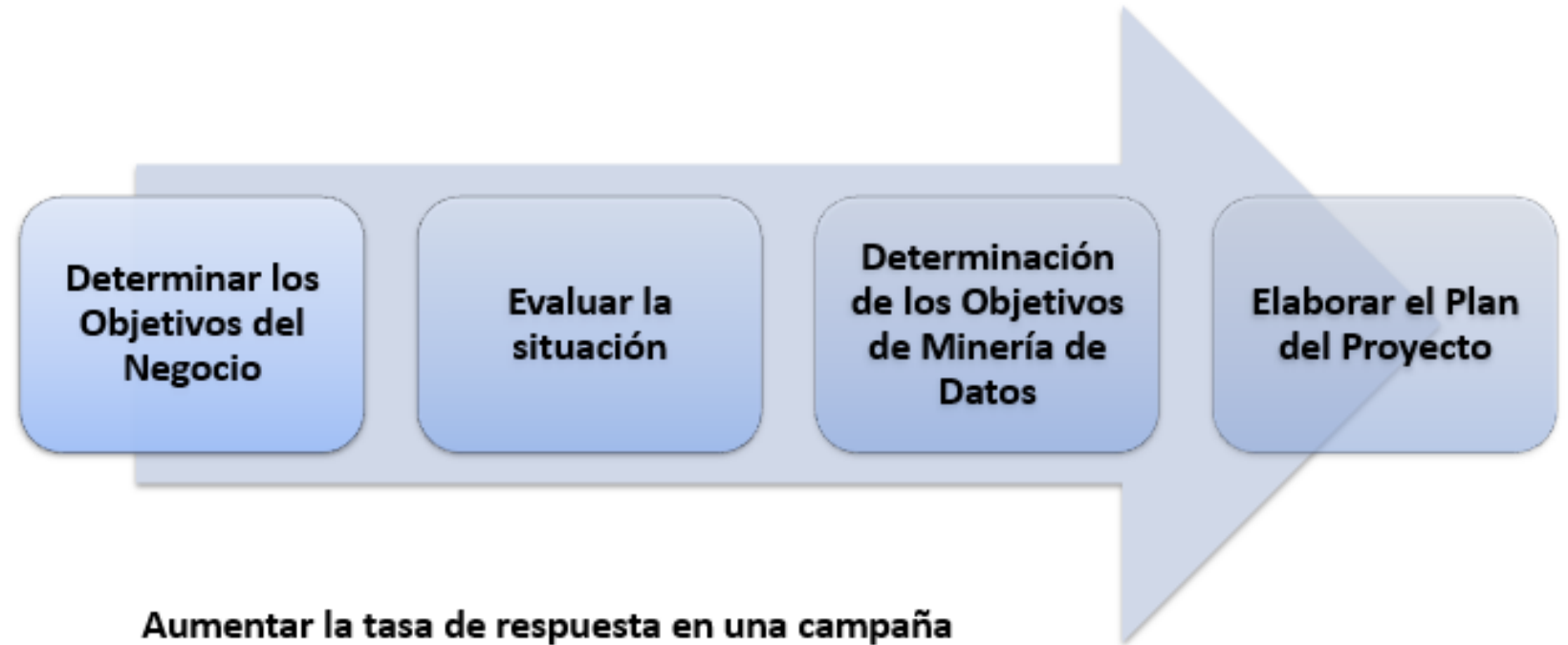
Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto



Data Cleaning

- Generación de datos de calidad.
- Datos primarios pueden llevar a conclusiones erróneas en el análisis.
- Mejora Considerable en el proceso de Análisis de Datos.

Data Collecting



- Se obtiene datos de diferentes fuentes.

Data Cleaning



- Resuelve conflictos entre datos.
- Elimina Outliers

Data Transformation



- Transformación y consolidación de los datos

Data Reduction



- Selección de características.
- Muestra del total.



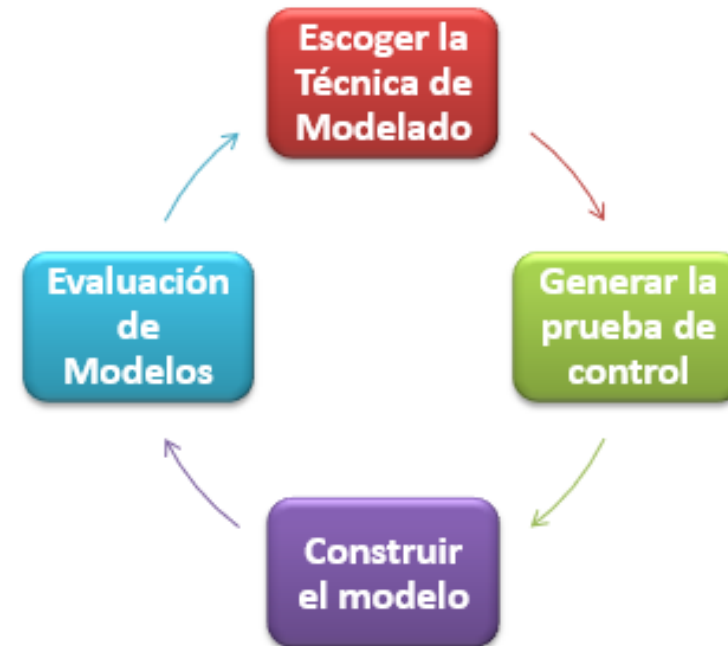
Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





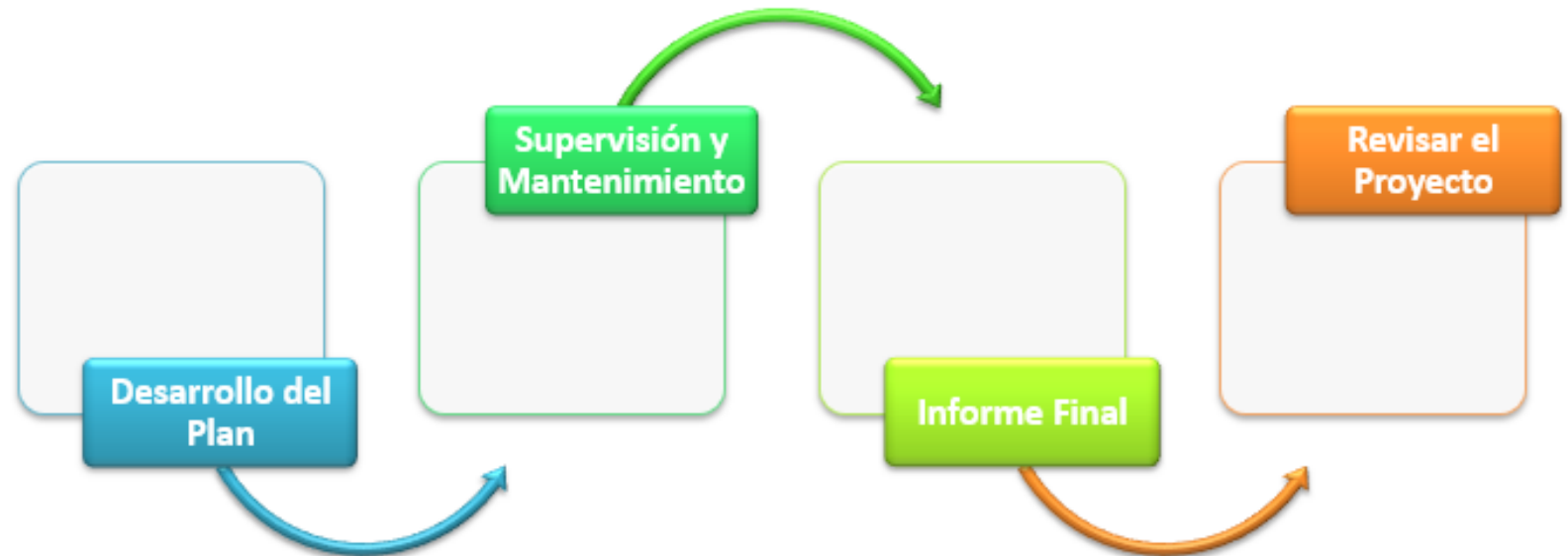
Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto





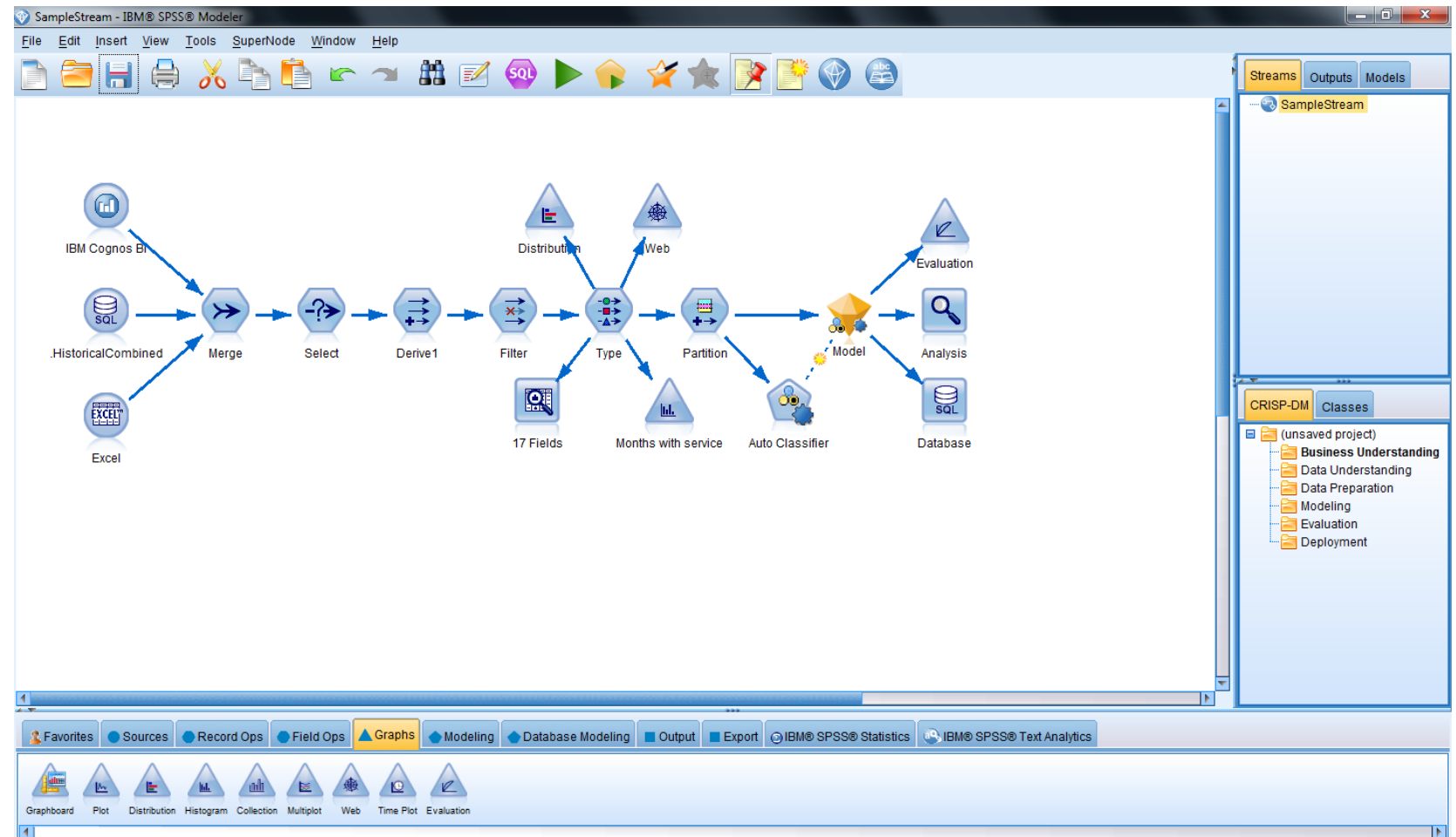
Metodologías



CRISP - DM

Cross Industry Standard Process for Data Mining

Desarrollo, Implementación y mantenimiento de CRISP - DM. Como un proceso de Outsourcing que simplifica la obtención de conocimiento sin la adquisición de software o personal experto



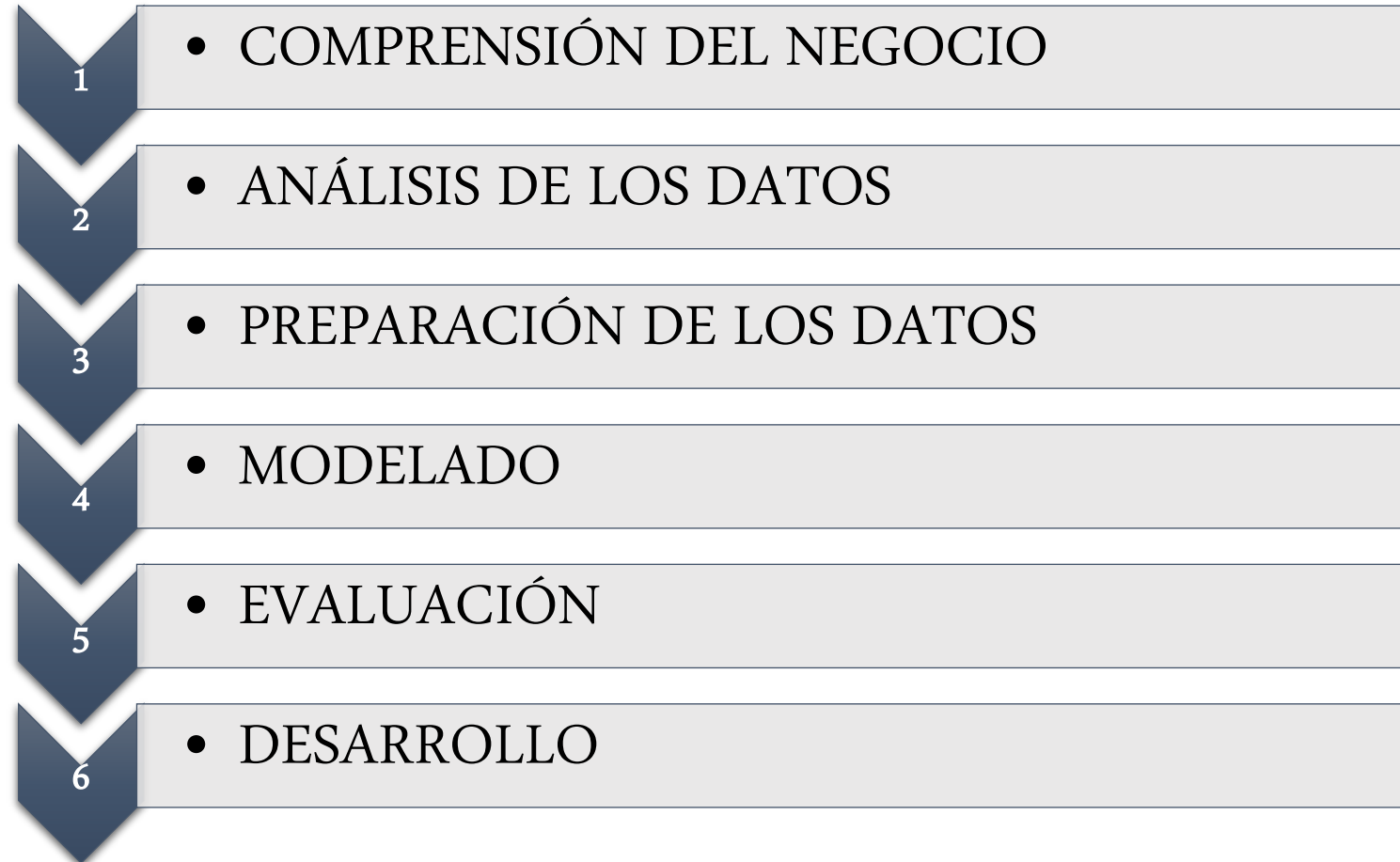
PROYECTO DE MACHINE LEARNING

MODELO DE PROPENSIÓN A LA COMPRA CON EL USO DE TARJETAS DE CRÉDITO



METODOLOGÍA CRISP - DM

(Cross Industry Standard Process For Data Mining)



CAMPAÑA KFC PARA CLIENTES DE PERUVIAN BANK

- Objetivo del negocio:
Motivar al cliente a utilizar su tarjeta de crédito para que pueda realizar sus consumos en la cadena de comida rápida KFC.
- Base de clientes:
7,000 clientes

Por compras en **KFC** con tu
TARJETA solicita descuentos de:

10%
De tus
consumos
del mes

Por compras mínimas de S/40.00

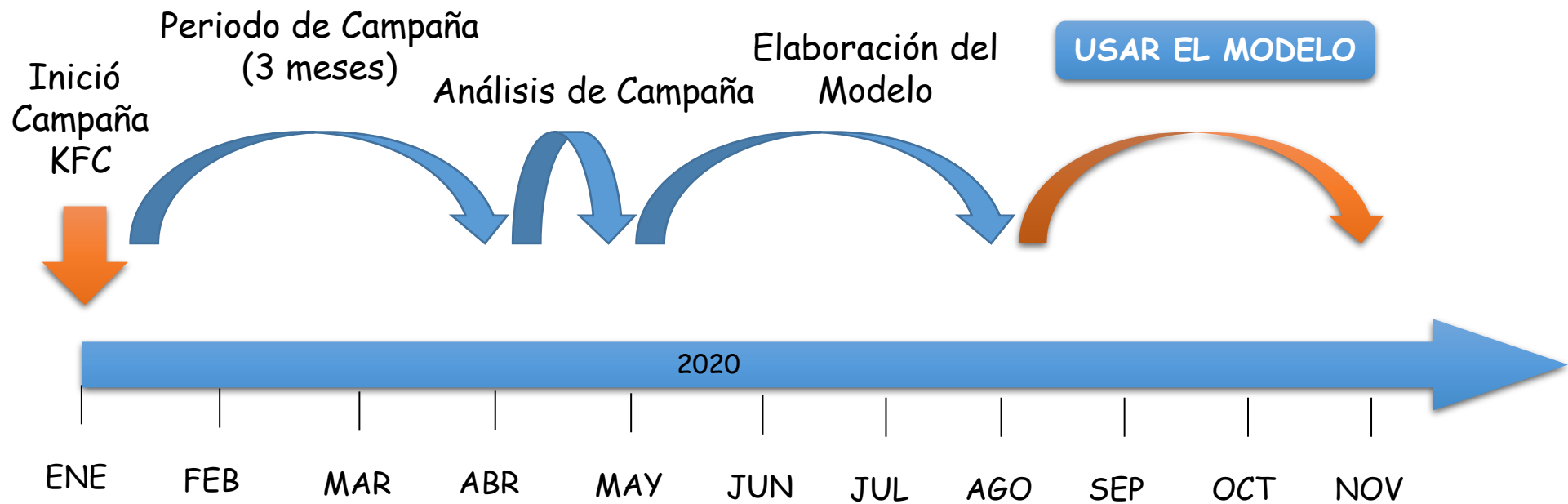


Acercaté a cualquiera de los locales de KFC y empieza a disfrutar de este beneficio exclusivo con tu Tarjeta de Crédito.



CAMPAÑA KFC
SIN USO DE UN MODELO DE DM

CAMPAÑA KFC
CON USO DE UN MODELO DE DM



Se plantea la elaboración de un Modelo de Machine Learning

COMPRENSIÓN
DEL NEGOCIO

ANÁLISIS DE
LOS DATOS

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

OBJETIVO

- Elaborar un modelo de Machine Learning para identificar a los clientes con mayor propensión a usar la **Tarjeta de crédito** para consumos en **KFC**.

PERÍODO

- El período con el que se elaborará el modelo de Machine Learning será con la información de la última campaña KFC, el cual abarca de Enero a Marzo .

EQUIPO

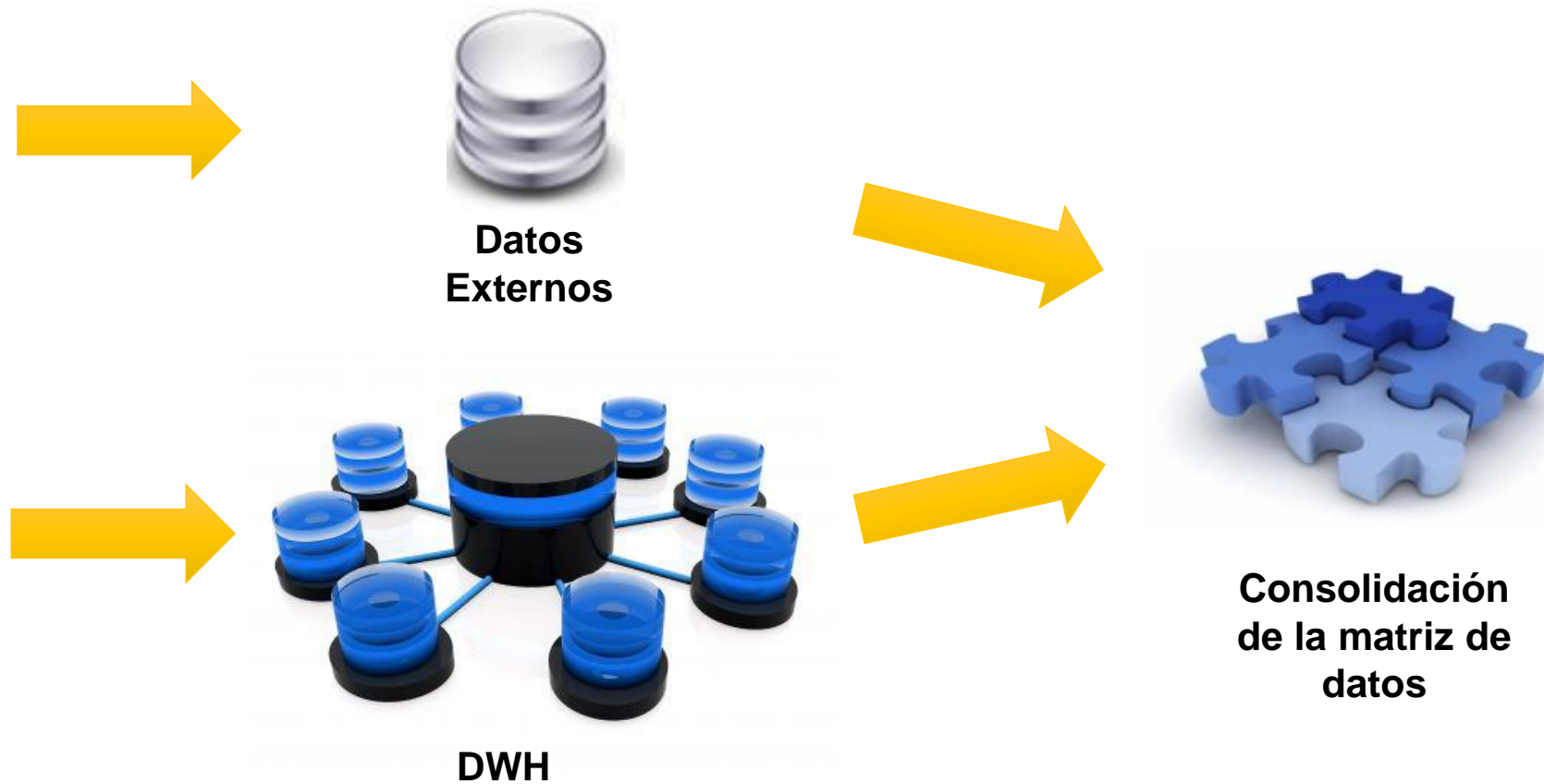
- 1 Jefe de Proyecto
- 1 Analista de Datawarehouse.
- 1 Analista de Machine Learning.

LICENCIAS

- 1 Licencia de Modeler.
- 1 Licencia de SQL SERVER.



Recolección inicial de datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Descripción de los datos

PROPENSIÓN

Actualizar

C:\Users\Toshiba\Desktop\Trabajo de Data Mining\bankloan.sav

Leer valores | Borrar valores | Borrar todos los valores

Campo	Tipo	Valores	Perdidos	Comprobar	Dirección
Edad	Rango	[20,56]		Ninguna	Entrada
NivelEducativo	Conjunto ordenado	1,2,3,4,5		Ninguna	Entrada
Empleo	Rango	[0,31]		Ninguna	Entrada
Residencia	Rango	[0,34]		Ninguna	Entrada
Ingresos	Rango	[14,446]		Ninguna	Entrada
# Deuda_Ingreso	Rango	[0.4,41.3]		Ninguna	Entrada
Region	Conjunto	1,2,3,4		Ninguna	Entrada
Casado	Conjunto	0,1		Ninguna	Entrada
Género	Conjunto	0,1		Ninguna	Entrada
Respuesta	Conjunto	0,1		Ninguna	Salida

☒ Ver campos actuales ☐ Ver configuración de campos no utilizados

Datos | Filtro | **Tipos** | Anotaciones

Aceptar | Cancelar | Aplicar | Restablecer

COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

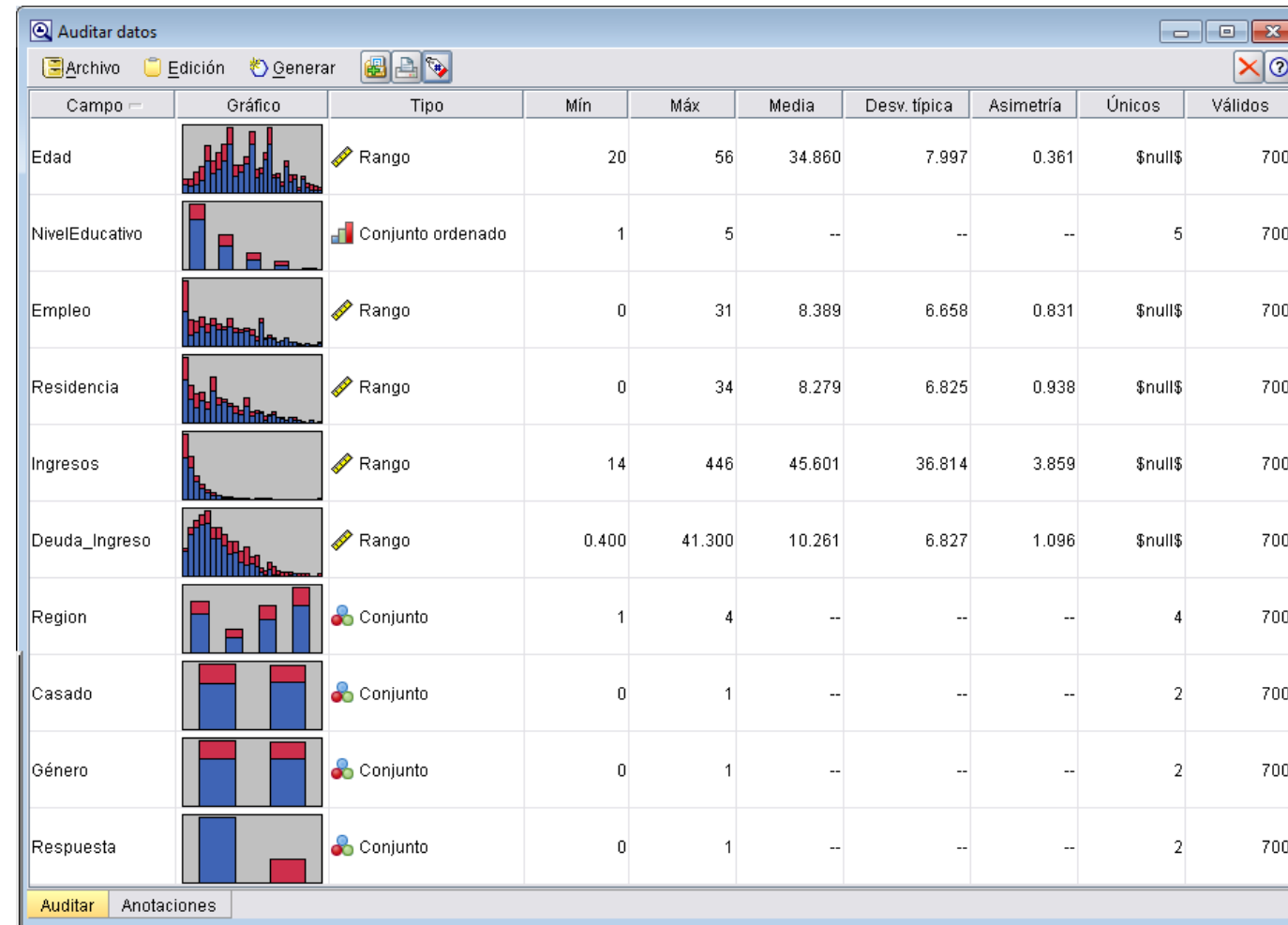
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

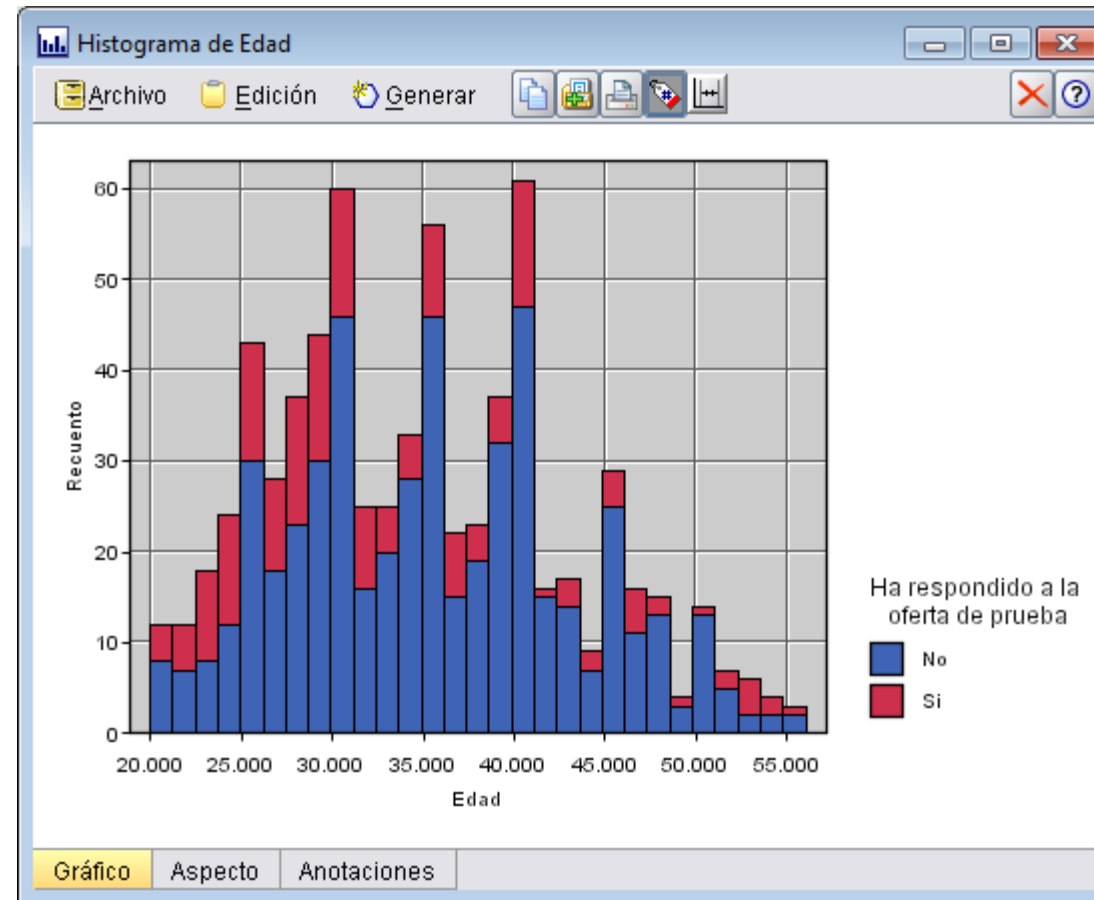
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

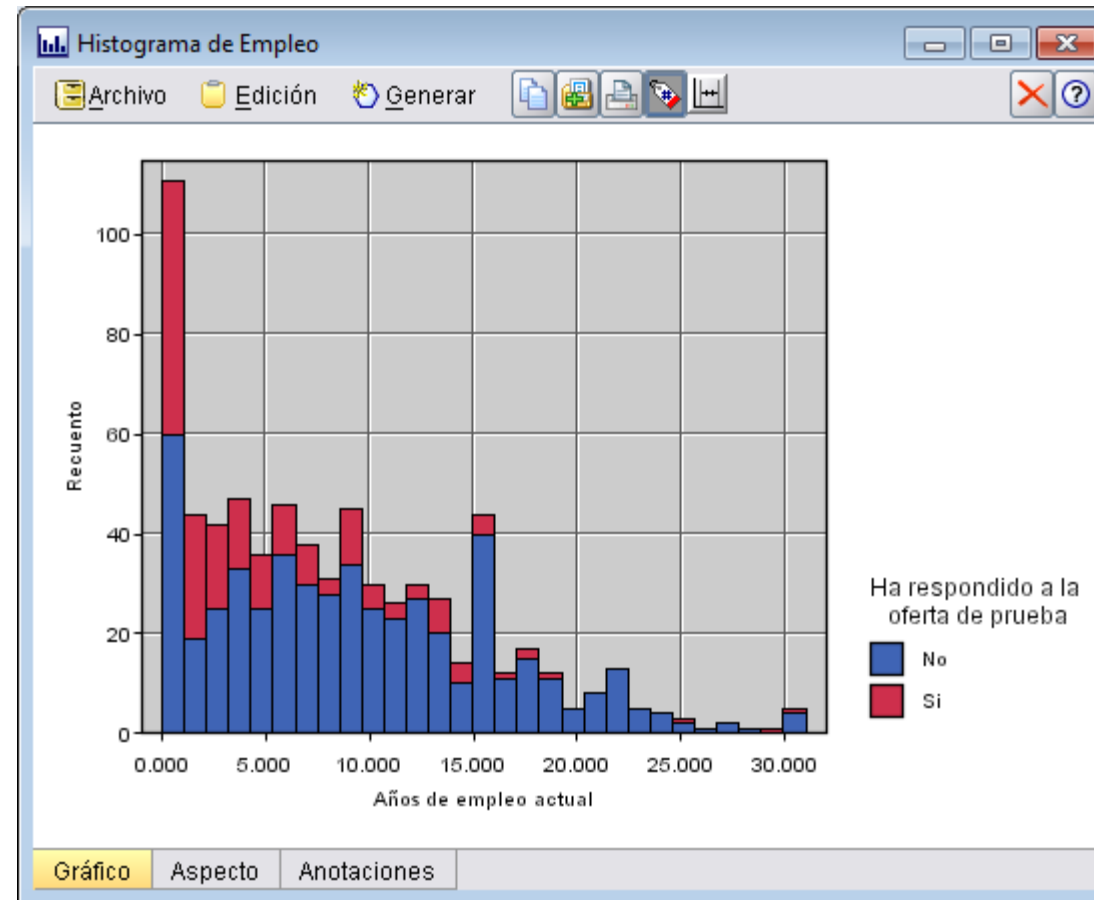
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

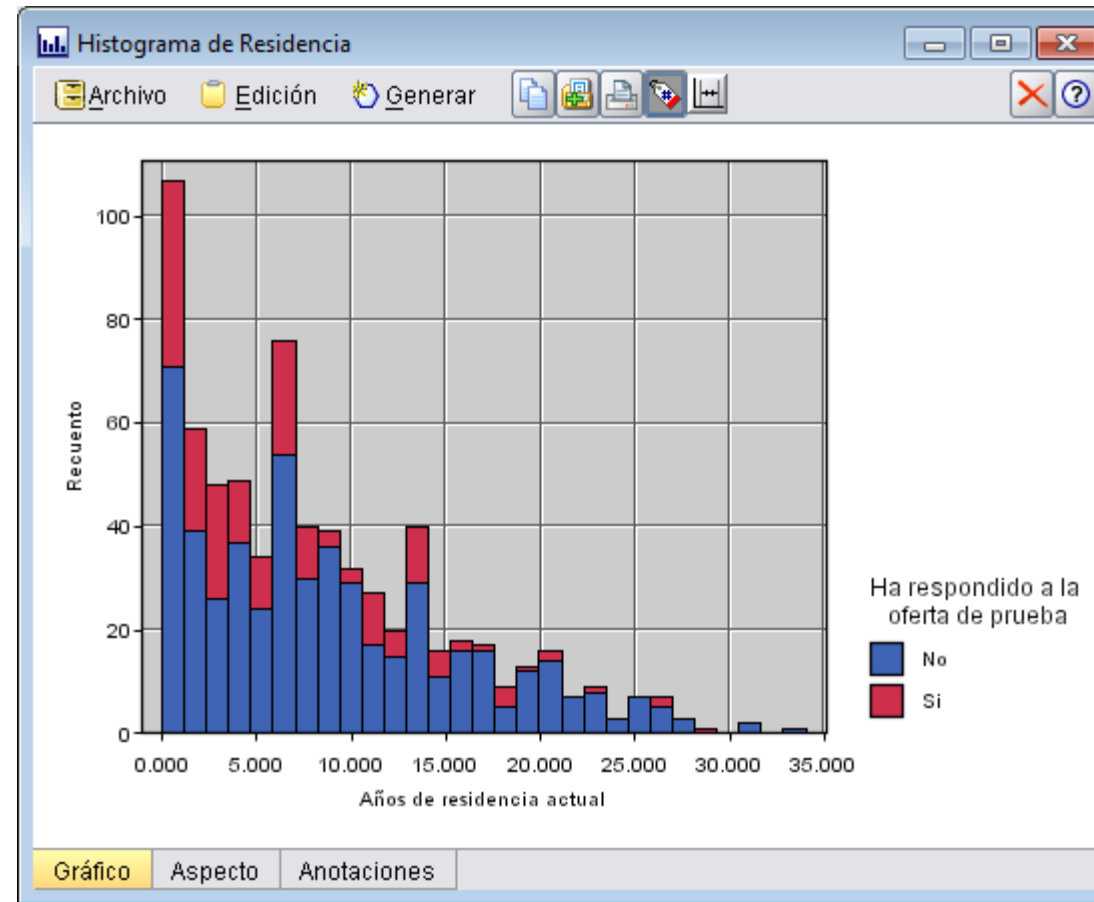
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

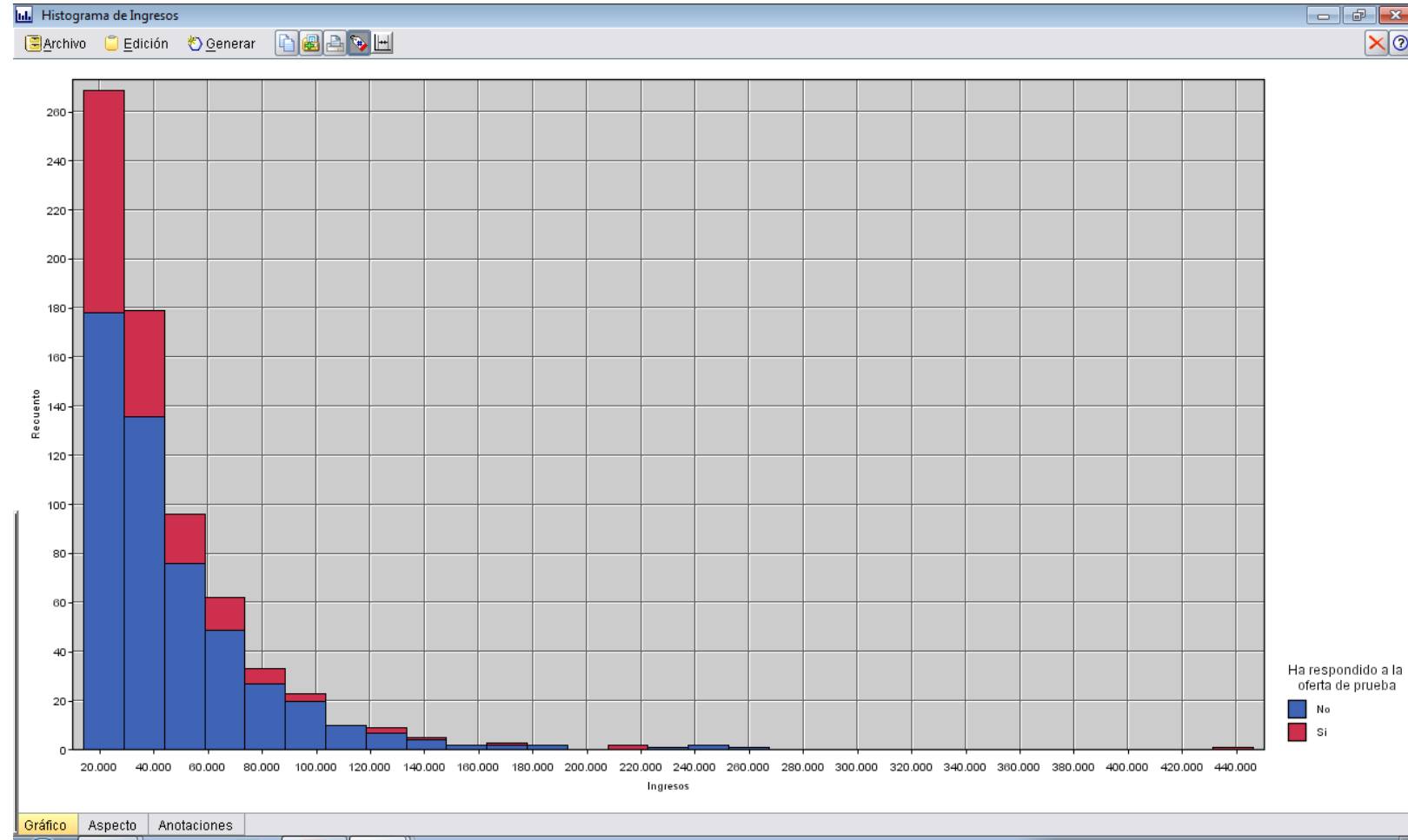
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

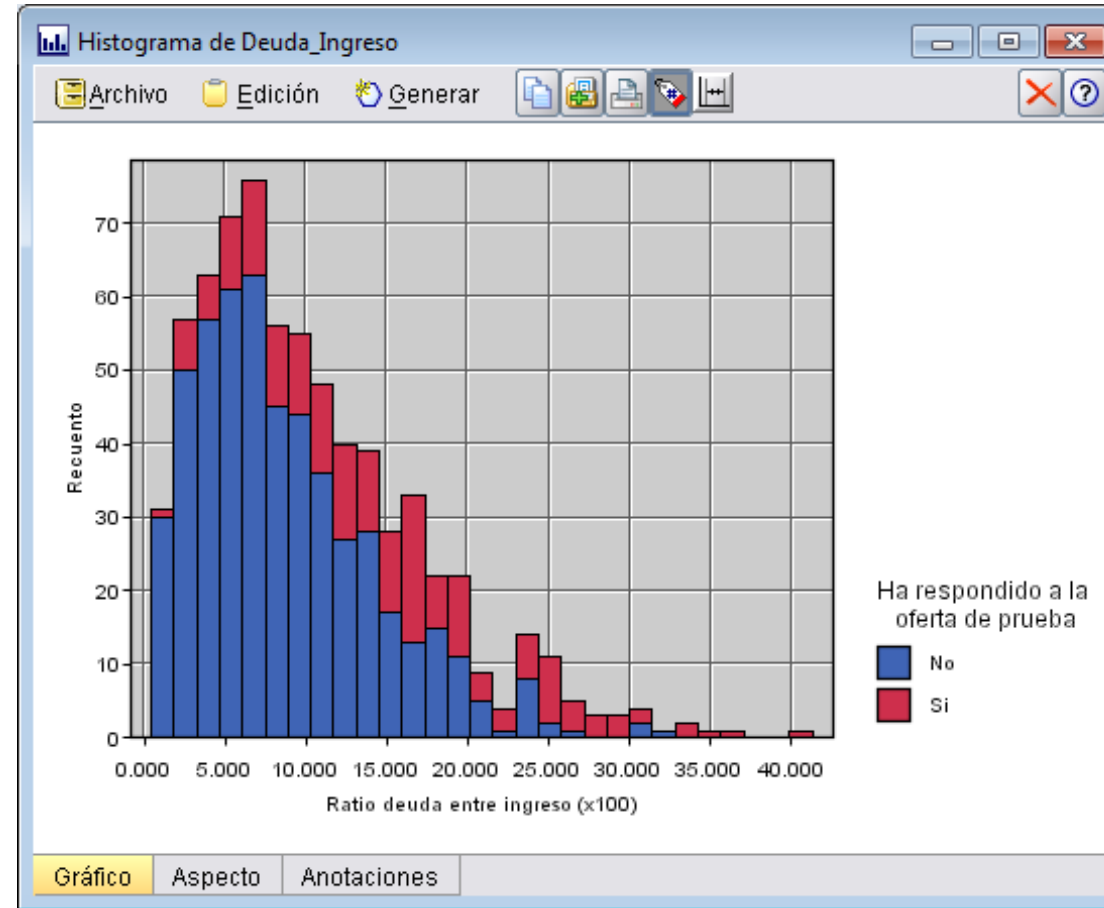
PREPARACIÓN
DE LOS DATOS

MODELADO

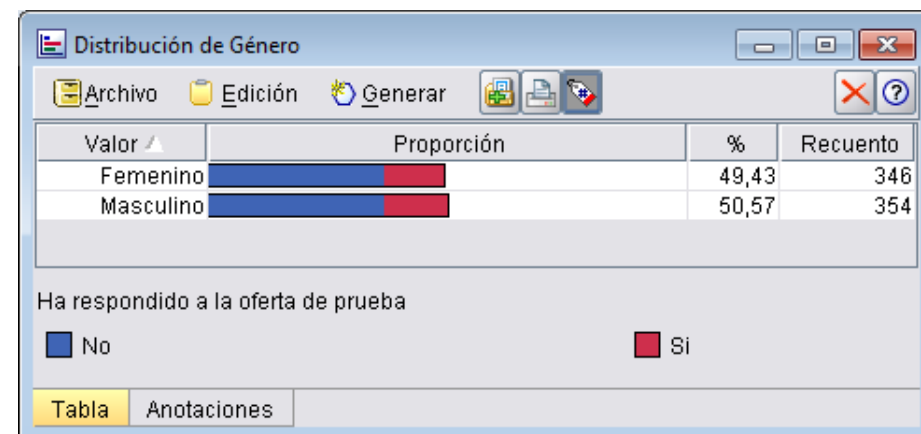
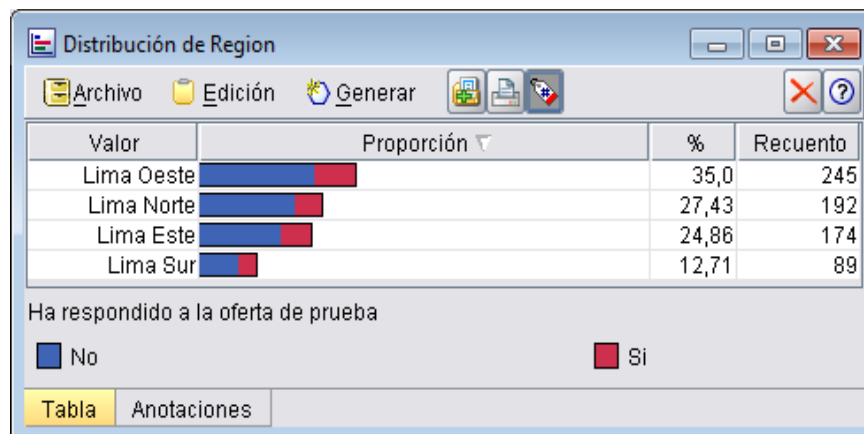
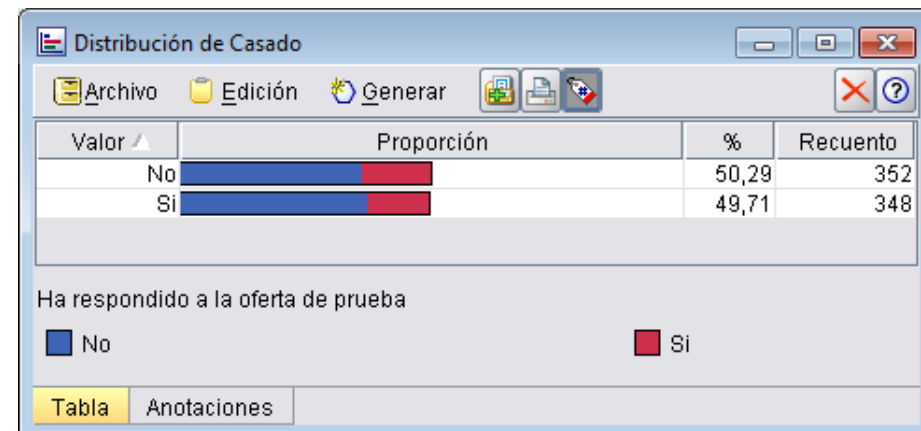
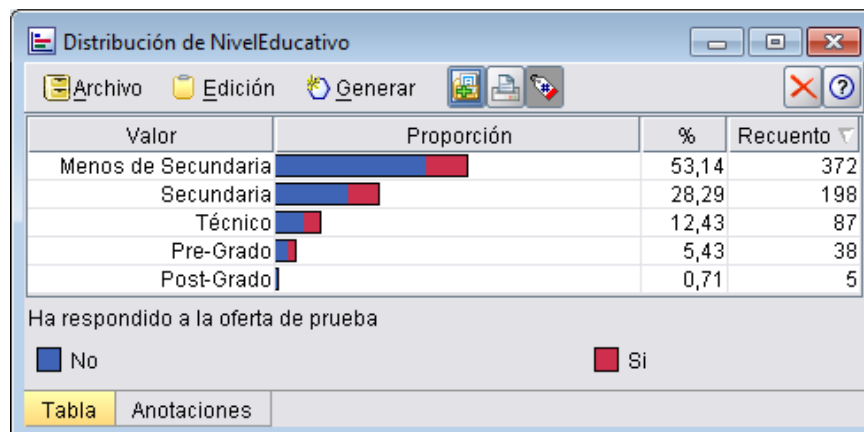
EVALUACIÓN

DESARROLLO

Exploración de los datos



Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

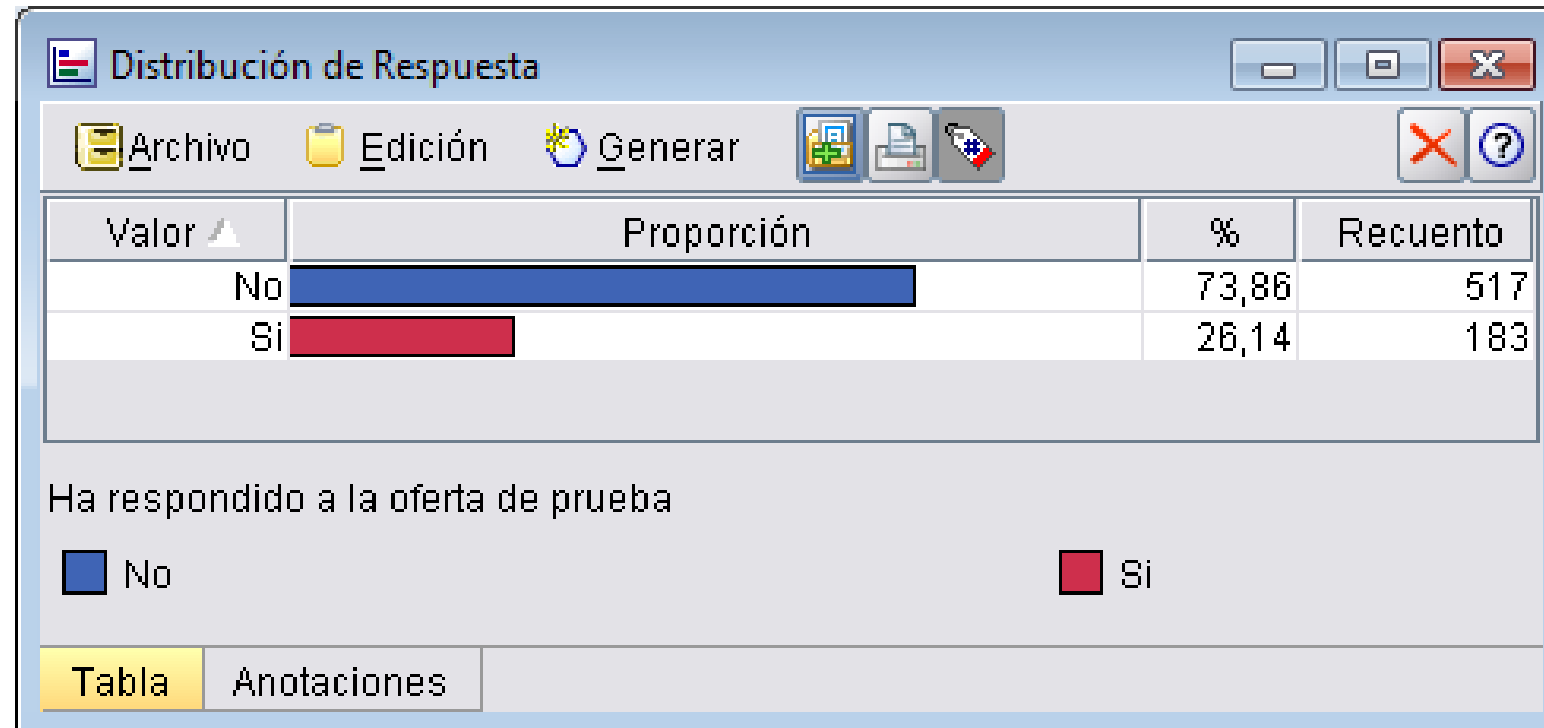
PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Exploración de los datos



COMPRENSIÓN DEL
NEGOCIO

**ANÁLISIS
DE LOS
DATOS**

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO

Calidad de los datos

Calidad de datos						
Archivo Edición Generar						
Campo	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blanco ▾	Valor vacío
Edad	100	700	0	0	0	0
NivelEducativo	100	700	0	0	0	0
Empleo	100	700	0	0	0	0
Residencia	100	700	0	0	0	0
Ingresos	100	700	0	0	0	0
Deuda_Ingreso	100	700	0	0	0	0
Region	100	700	0	0	0	0
Casado	100	700	0	0	0	0
Género	100	700	0	0	0	0
Respuesta	100	700	0	0	0	0
Calidad	Anotaciones					

COMPRESIÓN DEL
NEGOCIO

ANÁLISIS DE
LOS DATOS

**PREP. DE
LOS
DATOS**

MODELADO

EVALUACIÓN

DESARROLLO

Tipo

?

Leer valores Borrar valores Borrar todos los valores

Campo	Tipo	Valores	Perdidos	Comprobar	Dirección
Edad	Rango	[20,56]		Ninguna	Entrada
NivelEducativo	Conjunto ordenado	1,2,3,4,5		Ninguna	Entrada
Empleo	Rango	[0,31]		Ninguna	Entrada
Residencia	Rango	[0,34]		Ninguna	Entrada
Ingresos	Rango	[14,446]		Ninguna	Entrada
# Deuda_Ingreso	Rango	[0.4,41.3]		Ninguna	Entrada
Region	Conjunto	1,2,3,4		Ninguna	Entrada
Casado	Conjunto	0,1		Ninguna	Entrada
Género	Conjunto	0,1		Ninguna	Entrada
Respuesta	Conjunto	0,1		Ninguna	Salida

☒ Ver campos actuales ☐ Ver configuración de campos no utilizados

Tipos Formato Anotaciones

Aceptar Cancelar Aplicar Restablecer

COMPRENSIÓN DEL
NEGOCIO

ANÁLISIS DE
LOS DATOS

PREPARACIÓN
DE LOS DATOS



MODELADO

EVALUACIÓN

DESARROLLO

Partición de los datos

Partición

 Generar 

Campo de partición:

Particiones: ☒ Entrenamiento y comprobación ☐ Entrenamiento, comprobación y validación

Tamaño de partición de entrenamiento: Etiqueta: Valor =

Tamaño de partición de comprobación: Etiqueta: Valor =

Tamaño de partición de validación: Etiqueta: Valor =

Tamaño total: 100%

Valores: ☒ Utilizar valores definidos por el sistema ("1", "2" y "3")
☐ Añadir etiquetas a los valores definidos por el sistema
☐ Utilizar etiquetas como valores

☐ Establecer semilla aleatoria Semilla:

Configuración **Anotaciones**

COMPRESIÓN DEL
NEGOCIO

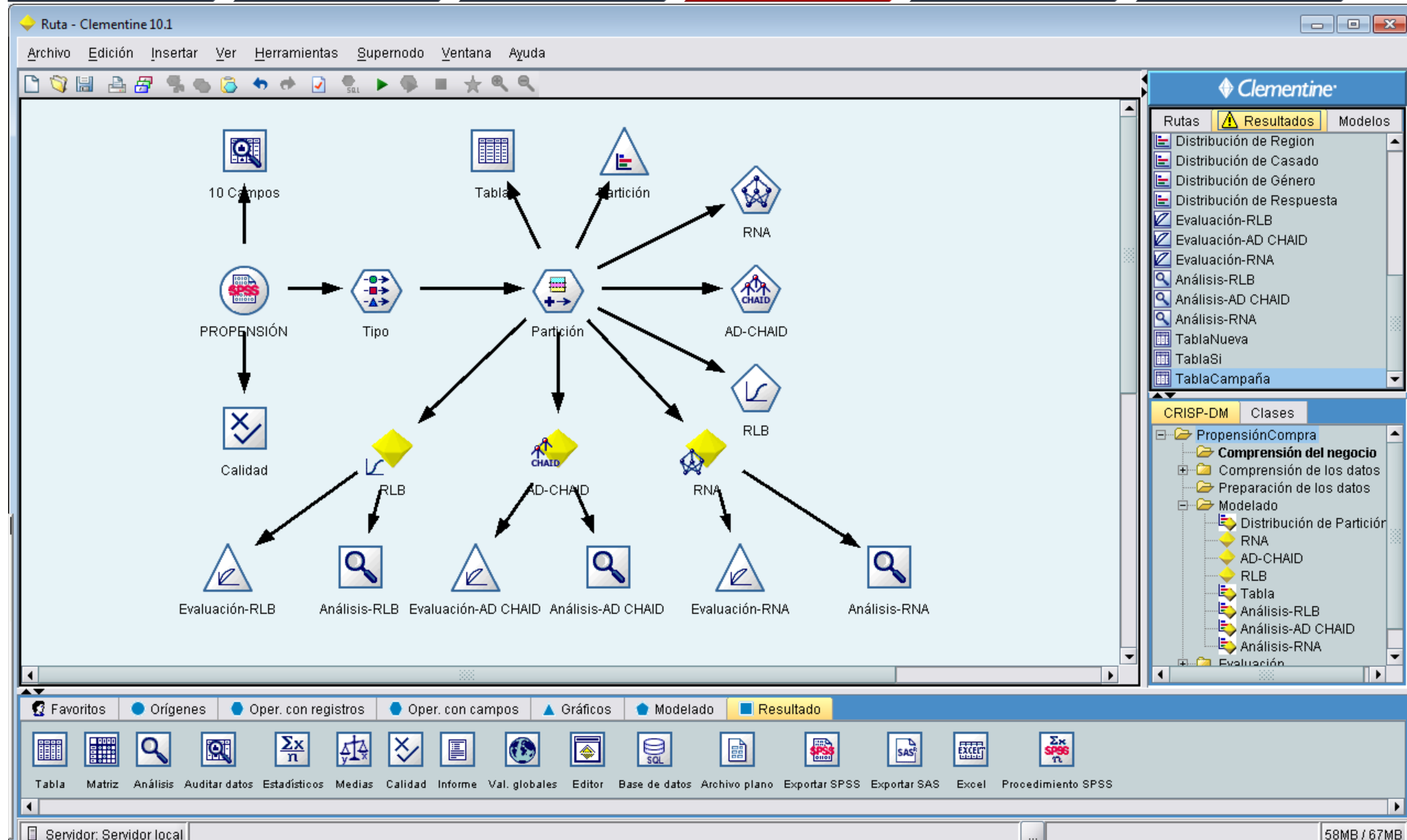
ANÁLISIS DE
LOS DATOS

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO





RLB

Entrenamiento	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	327	39	89.3%
SI	68	60	46.9%
Porcentaje Global Correcto			78.3%
Porcentaje Global Incorrecto			21.7%

Validación	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	143	8	94.7%
SI	30	25	45.5%
Porcentaje Global Correcto			81.6%
Porcentaje Global Incorrecto			18.4%

AD-CHAID

Entrenamiento	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	337	28	92.3%
SI	82	58	41.4%
Porcentaje Global Correcto			78.2%
Porcentaje Global Incorrecto			21.8%

Validación	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	139	13	91.4%
SI	24	21	46.7%
Porcentaje Global Correcto			81.2%
Porcentaje Global Incorrecto			18.8%

RNA

Entrenamiento	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	321	29	91.7%
SI	72	48	40.0%
Porcentaje Global Correcto			78.5%
Porcentaje Global Incorrecto			21.5%

Validación	Pronosticado		
Observado	NO	SI	Porcentaje Correcto
NO	158	9	94.6%
SI	40	23	36.5%
Porcentaje Global Correcto			78.7%
Porcentaje Global Incorrecto			21.3%



Logistic



CHAID



Neural Net

MODELOS (RESPUESTA=SI)	RLB	AD	RNA
Entrenamiento	46.9%	41.4%	40.0%
Validación	45.5%	46.7%	36.5%
% Entrenamiento Global Correcto	78.3%	78.2%	78.5%
% Validación Global Correcto	81.6%	81.2%	78.7%
% Entrenamiento Global Incorrecto	21.7%	21.8%	21.5%
% Validación Global Incorrecto	18.4%	18.8%	21.3%

COMPRENSIÓN DEL
NEGOCIO

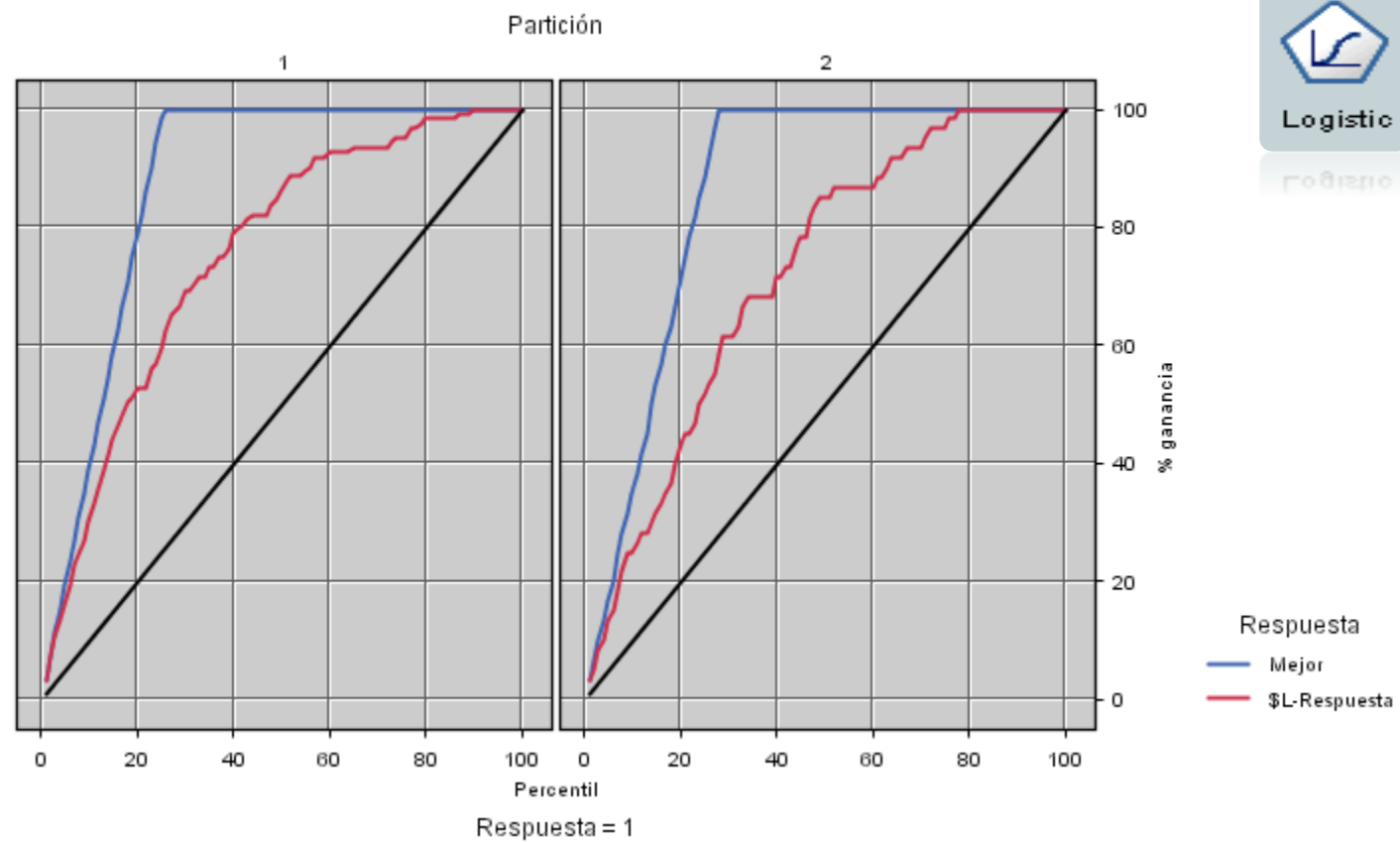
ANÁLISIS DE
LOS DATOS

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO



COMPRESIÓN DEL
NEGOCIO

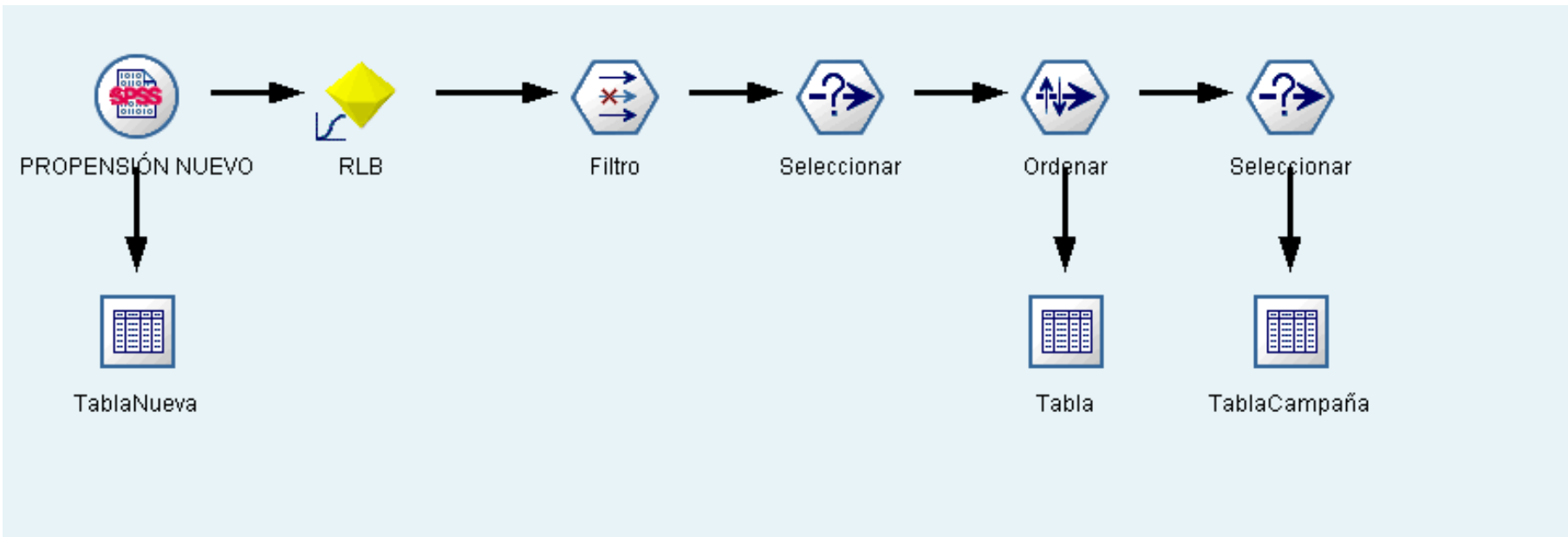
ANÁLISIS DE
LOS DATOS

PREPARACIÓN
DE LOS DATOS

MODELADO

EVALUACIÓN

DESARROLLO



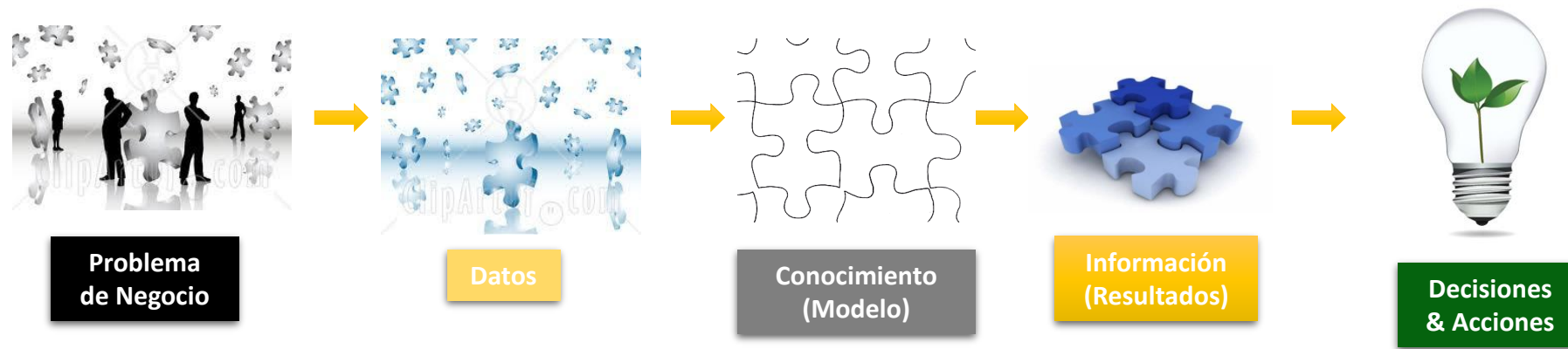


Contenido 1

- Definición, importancia y aplicaciones de Machine Learning
- Técnicas de aprendizaje
- Proceso de extracción de conocimiento (KDD) y su relación con Machine Learning
- Ciclo de vida de un proyecto de Machine Learning



Ciclo de vida de un Proyecto de Machine Learning





Caso de aplicación



Bank Perú es una empresa del rubro financiero.

- Se sabe que un gran número de clientes migran a otros bancos por medio de la compra deuda o simplemente gestiona la baja de su tarjeta de crédito.
- El nuevo gerente del área de producto cuenta con un presupuesto para implementar una **campana de retención** que va dirigido a 50,000 clientes de un total de 500,000.

Como responsable del área de Data Analytics:

- 1.¿Cuál es el problema de negocio?
- 2.¿Qué variables debería de usar?
- 3.¿Cómo se debería elegir a los clientes con los que se entrará en contacto para maximizar la eficacia de la campaña de retención?



Caso de aplicación



Variable negocio
Monto de Facturación

Score Churn	Alto (Más de S/5K)	Medio (Entre S/2K-S/5K)	Bajo (Menos de S/2K)
Alto (0.70 a más)	P1	P1	P3
Medio (0.50-0.70)	P1	P2	P3
Bajo (Menos de 0.50)	P2	P3	P3

P1 => Call Center
P2 => Email Marketing
P3 => SMS



Caso de aplicación



Variable negocio
Deuda a comprar o Línea de Crédito Ofrecida

Score CD	Alto (Más de S/50K)	Medio (Entre S/20K-S/50K)	Bajo (Menos de S/20K)
Alto (0.70 a más)	P1	P1	P3
Medio (0.50-0.70)	P1	P2	P3
Bajo (Menos de 0.50)	P2	P3	P3

P1 => Call Center

P2 => Email Marketing

P3 => SMS



TELECOM



Variable negocio
Tipo de plan pospago

Caso de aplicación

Score Captación	Alto (Más de S/150)	Medio (Entre S/70-S/150)	Bajo (Menos de S/70)
Alto (0.70 a más)	P1	P1	P3
Medio (0.50-0.70)	P1	P2	P3
Bajo (Menos de 0.50)	P2	P3	P3

P1 => Call Center

P2 => Email Marketing

P3 => SMS



CIERRE



¿Cuál es la
diferencia entre
DM, DWH y DL?

¿Por qué es
importante
entender las
preguntas de
negocio?

Explicar el
proceso de
Machine
Learning



CONSULTAS

pcsirife@upc.edu.pe