



## DESAFIO 2

### Árbol de decisiones

Apellidos	Nombres	Carné
Navas de Alvarenga	Sandra Noemy	ND221932

# Datawarehouse y Minería de Datos DMD941 G01T

## Introducción

El presente desafío tiene como finalidad aplicar técnicas de **Minería de Datos** mediante la construcción de **árboles de decisión** en **RapidMiner**, con el objetivo de identificar patrones que permitan determinar si una persona puede ser considerada **cliente potencial**.

Para el ejercicio se utilizaron tres archivos base (Whisky, Estado\_Civil y Primer\_Compra) que representan diferentes características de clientes, tales como edad, estado civil, tipo de compra y condiciones económicas.

El propósito es **integrar, limpiar y analizar** estos datos para descubrir cómo influyen variables como **tipo de pago** y **edad** en la probabilidad de que una persona sea cliente potencial.

## Recursos Utilizados

RapidMiner Studio

Archivos Excel (.xls / .xlsx)

GitHub

## Resultados

### 1. Estructura de las bases de datos

Para el desarrollo del desafío se proporcionaron tres archivos en formato Excel (.xls y .xlsx), los cuales representan distintas dimensiones de información sobre los posibles clientes. Cada archivo contiene atributos que fueron posteriormente integrados, renombrados y transformados para realizar el análisis en RapidMiner.

#### a. Archivo: Whisky

Contiene información relacionada con las características de distintos productos de whisky, principalmente su precio y categoría.

Columna	Tipo de dato	Descripción
id_Whisky	Integer	Identificador único de cada whisky.
Precio	Real	Valor monetario del producto.
Malta	Integer	Tipo o porcentaje de malta utilizada.
Categoría	Polynomial	Clasificación del producto (por ejemplo: "Lujo").
Añejamiento	Real	Tiempo de añejamiento del producto en años.
Calidad	Integer	Nivel de calidad asignado al producto.

#### b. Archivo: Estado\_Civil

Contiene datos demográficos y económicos de los clientes, representando variables personales.

## Datawarehouse y Minería de Datos DMD941 G01T

Columna	Tipo de dato	Descripción
dinero	Real	Nivel de ingresos o capacidad económica.
casa	Integer	Indica si la persona posee vivienda (1 = Sí, 0 = No).
estado	Polynomial	Estado civil del individuo (soltero, casado, viudo, etc.).
sexo	Integer	Género de la persona (0 = Femenino, 1 = Masculino).
auto	Integer	Indica si posee vehículo (1 = Sí, 0 = No).
edad	Integer	Edad del individuo.

### c. Archivo: Primer\_Compra

Incluye información sobre el tipo de producto adquirido por primera vez, así como la edad y sueldo del comprador.

Columna	Tipo de dato	Descripción
Primer_Compra	Polynomial	Tipo de producto adquirido (carro, casa, ropa, etc.).
Estado Civil	Polynomial	Estado civil al momento de la compra.
Edad	Integer	Edad del comprador.
Sueldo	Real	Ingreso mensual en el momento de la compra.

## 2. Proceso RapidMiner

El proyecto se implementó como un flujo completo dentro de RapidMiner Studio, cumpliendo las tres etapas básicas de un proceso de minería de datos: **preparación, transformación y análisis**.

### 1. Carga y preparación de los datos

Se importaron los tres archivos base mediante el operador Read Excel. Cada dataset fue verificado para definir correctamente el tipo de dato de cada atributo (integer, real, polynomial).

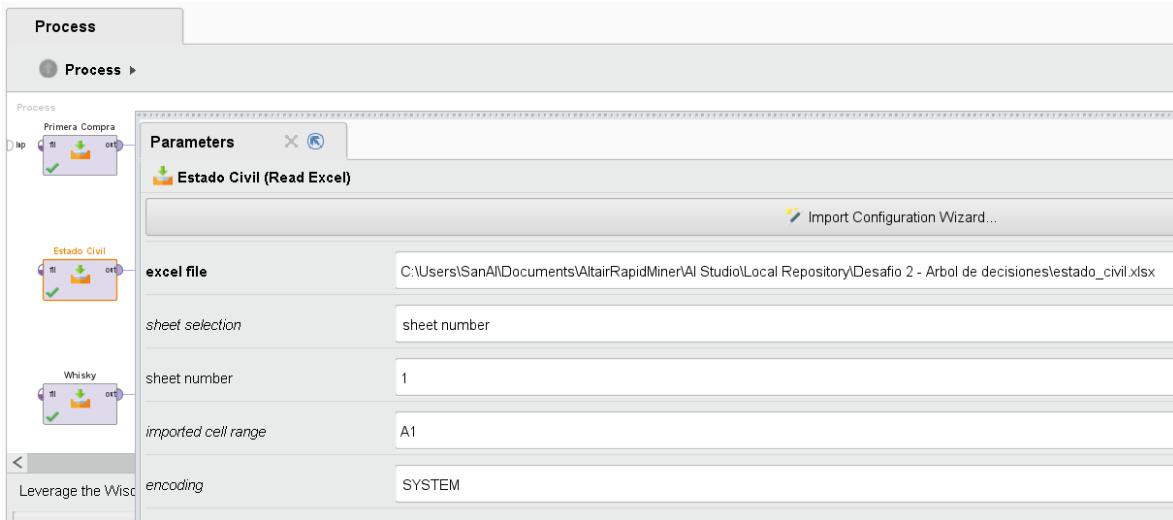
#### A) Read Excel para Primer\_Compra.xlsx

- Parámetro **file** → selecciona Primer\_Compra.xlsx.
- En Import Configuration Wizard:
  - **First row as names** = true
  - **Sheet number** = 1 (o la hoja correcta)
  - En **Attributes** verifica y asigna:
    - Primer\_Compra → **polynomial**
    - Estado Civil → **polynomial**

## Datawarehouse y Minería de Datos DMD941 G01T

- Edad → **integer**
  - Sueldo → **real**
  - **Salida:** puerto exa.
- B) Read Excel para Estado\_Civil.xlsx**
- Parámetro **file** → Estado\_Civil.xlsx.
  - Import Configuration Wizard:
    - **First row as names** = true
    - **Sheet number** = 1
    - Atributos:
      - dinero → **real**
      - casa → **integer**
      - estado → **polynomial**
      - sexo → **polynomial** (si es 0/1 puedes dejar integer, pero polynomial es más claro)
      - auto → **integer**
      - edad → **integer**
  - **Salida:** exa.
- C) Read Excel para Whisky.xlsx**
- Parámetro **file** → Whisky.xlsx.
  - Wizard:
    - **First row as names** = true
    - **Sheet number** = 1
    - Atributos:
      - id\_Whisky → **integer**
      - Precio → **real**
      - Malta → **integer** o **real** según datos
      - Categoría → **polynomial**
      - Añejamiento → **real**
      - Calidad → **integer**
  - **Salida:** exa.

# Datawarehouse y Minería de Datos DMD941 G01T



1.2 Posteriormente, se aplicaron transformaciones en los nombres con operadores de Rename para unificar nombres de columnas y generar atributos comunes.

## A) Rename para Primer\_Compra.xlsx

- Primer\_Compra → **Producto**

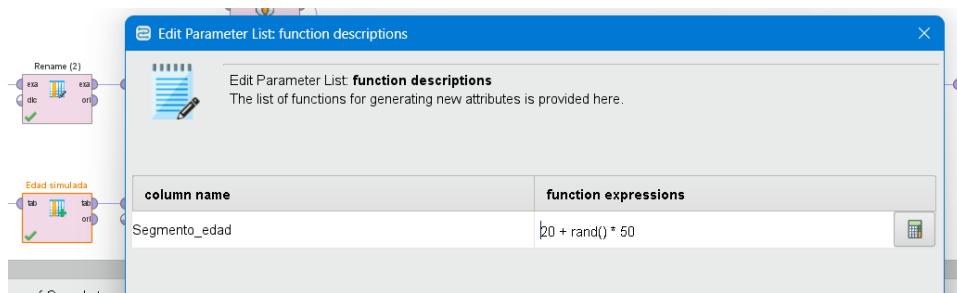
## B) Rename para Estado\_Civil.xlsx

- dinero → **Sueldo**
- estado → **Estado\_Civil**
- edad → **Edad**

## C) Rename para Whisky.xlsx

- Precio → **Sueldo**
- Categoria → **Producto**
- Segmento\_edad → **Edad**

1.3 Dado que no contábamos con un campo de edades para la tabla de Whisky.xlsx, generamos una edad aleatoria con un "Generador de atributos" y la formula "**20 + rand() \* 50**" para crear grupos de edades entre 21 a 70 años.



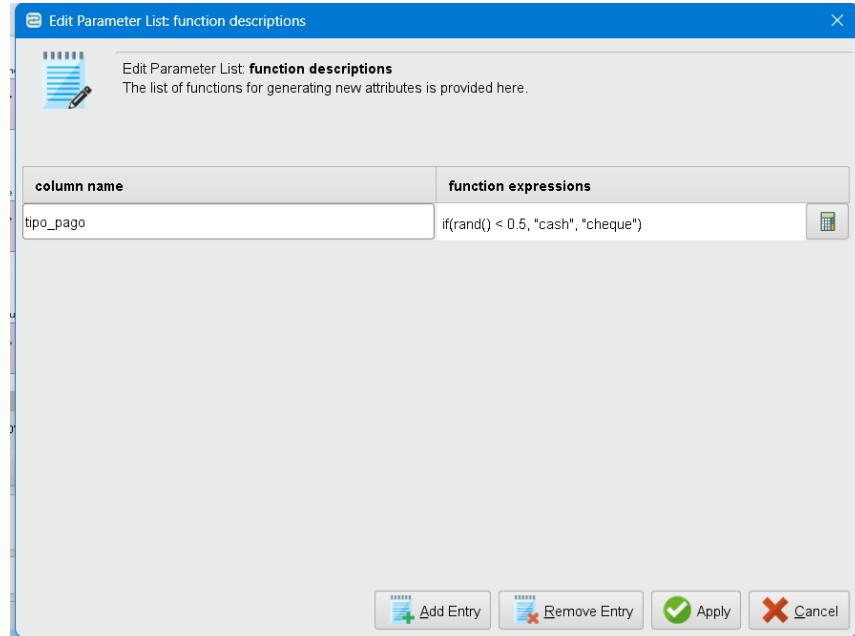
# Datawarehouse y Minería de Datos DMD941 G01T

## 2. Creación del atributo “tipo\_pago”

Dado que los archivos originales no incluían información sobre el método de pago, se generó un nuevo atributo denominado tipo\_pago, utilizando la siguiente fórmula:

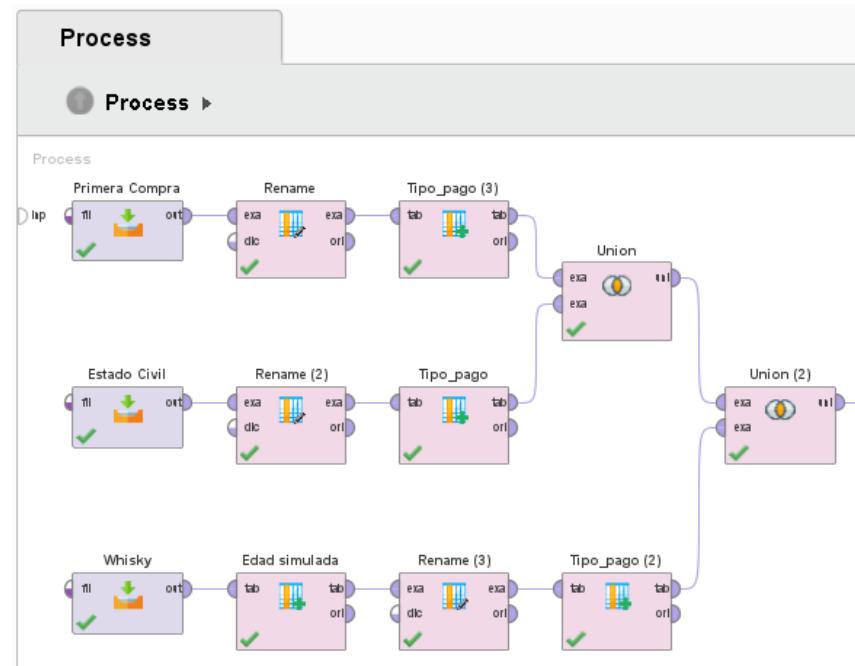
tipo\_pago = if(rand() < 0.5, "cash", "cheque")

Esto permite simular los dos métodos de pago solicitados para el análisis: **efectivo (cash) y cheque**.



## 3. Integración de datos

Los tres conjuntos se unieron mediante el operador Union, generando un único dataset consolidado con las variables de interés.



# Datawarehouse y Minería de Datos DMD941 G01T

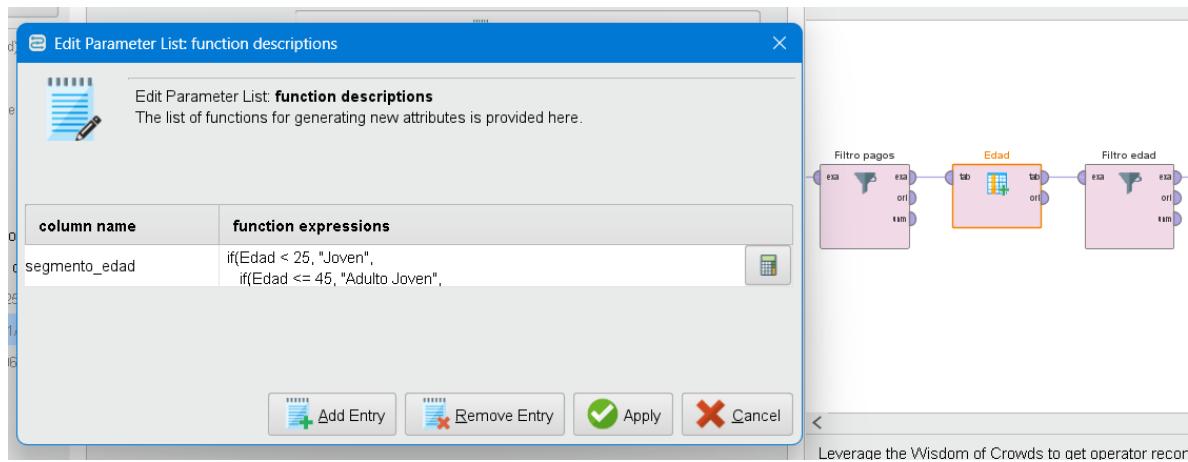
## 4. Filtrado de registros

Se aplicó un filtro para conservar únicamente los registros con tipo de pago válido (efectivo o cheque):

```
tipo_pago == "cash" || tipo_pago == "cheque"
```

Además, se creó una segmentación de edad con el operador Generate Attributes para clasificar a los clientes según su grupo etario:

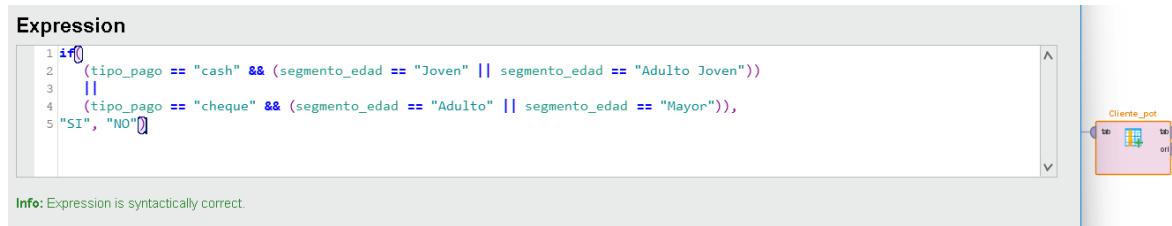
```
segmento_edad = if(Edad < 25, "Joven",
                    if(Edad <= 45, "Adulto Joven",
                       if(Edad <= 60, "Adulto", "Mayor")))
```



## 5. Generación del atributo “cliente\_potencial”

Se diseñó una regla basada únicamente en el tipo de pago y la edad, de acuerdo con la hipótesis del ejercicio:

```
cliente_potencial =  
  
if(  
  
(tipo_pago == "cash" && (Edad <= 45))  
  
||  
  
(tipo_pago == "cheque" && (Edad > 45)),  
  
"SI", "NO")
```



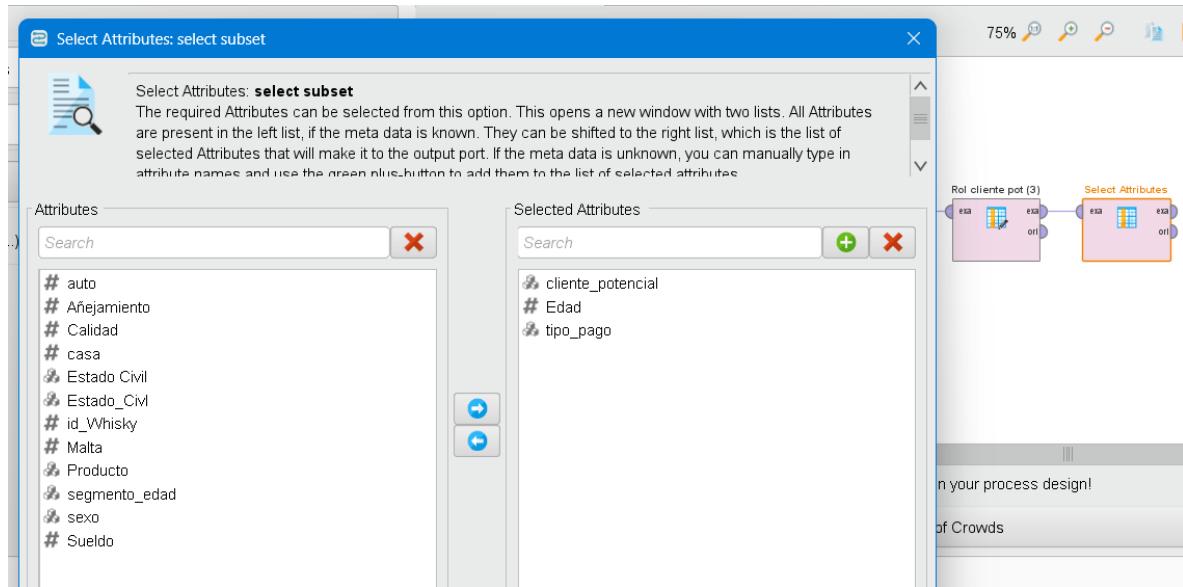
# Datawarehouse y Minería de Datos DMD941 G01T

## 6. Selección de atributos relevantes

Para el modelo se utilizaron únicamente tres atributos:

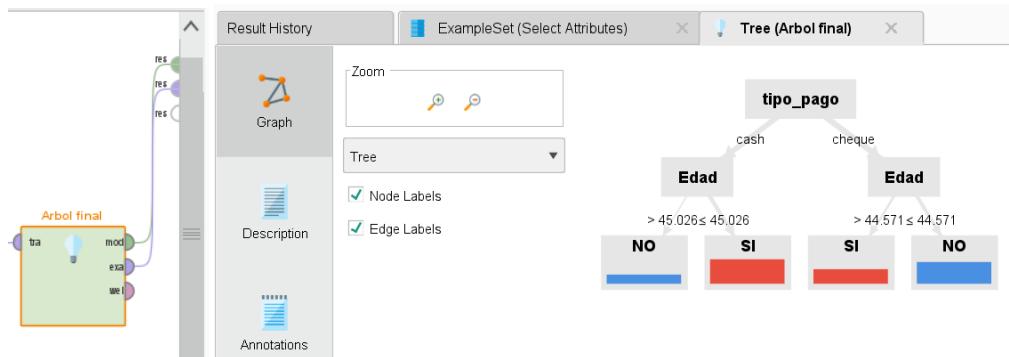
- tipo\_pago
- Edad
- cliente\_potencial

Esto se realizó con el operador Select Attributes, seguido de Set Role para definir cliente\_potencial como la variable **label**.



## 7. Generación del Árbol de Decisión

Finalmente, se utilizó el operador Decision Tree para construir el modelo predictivo.



# Datawarehouse y Minería de Datos DMD941 G01T

## 8. Resultados:

### Tree

tipo\_pago = cash

| Edad > 45.026: NO {NO=9, SI=0}

| Edad ≤ 45.026: SI {NO=0, SI=26}

tipo\_pago = cheque

| Edad > 44.571: SI {NO=0, SI=15}

| Edad ≤ 44.571: NO {NO=23, SI=0}

### Interpretación de resultados:

- Los clientes que pagan en efectivo (cash) y tienen **menos de 45 años** presentan mayor probabilidad de ser **clientes potenciales**.  
💡 Asociado a compras más impulsivas o de menor costo.
- Los clientes que pagan con cheque y tienen **más de 45 años** son también **clientes potenciales**, indicando un perfil de **compradores planificados** o de productos de mayor valor.

The screenshot shows the RapidMiner interface with the following details:

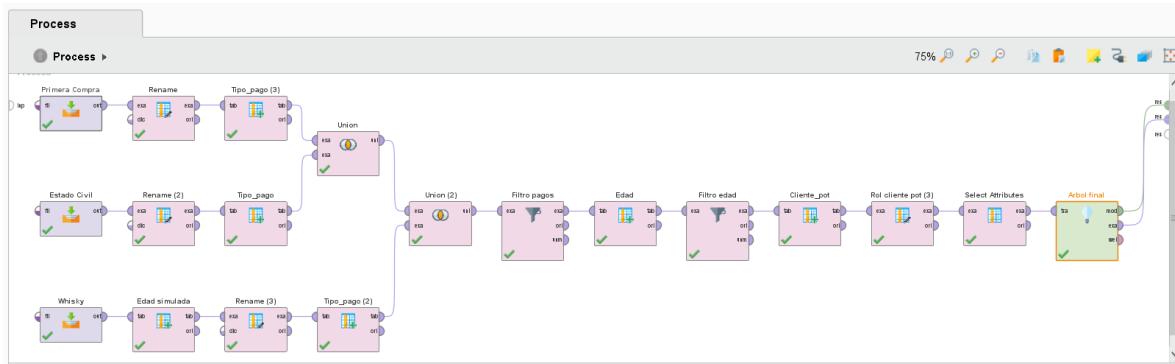
- Result History:** Shows the execution history of the process.
- ExampleSet (Select Attributes):** The current step selected.
- Tree (Arbol final):** The final model step.
- Data:** A table showing the filtered data with columns: Row No., cliente\_pote..., Edad, and tipo\_pago.
- Statistics:** A section for statistical analysis.
- Visualizations:** A section for creating visual representations.
- Annotations:** A section for adding annotations.
- Buttons:** Open in Turbo Prep, Auto Model, Interactive Analysis.

Row No.	cliente_pote...	Edad	tipo_pago
1	NO	22	cheque
2	NO	21	cheque
3	NO	21	cheque
4	SI	21	cash
5	SI	25	cash
6	NO	22	cheque
7	SI	21	cash
8	NO	23	cheque
9	SI	25	cash
10	SI	23	cash
11	SI	23	cash
12	SI	21	cash
13	NO	22	cheque
14	NO	23	cheque
15	SI	25	cash

Al correr la simulación, este también nos da una tabla con los elementos filtrados y el listado de clientes potenciales

# Datawarehouse y Minería de Datos DMD941 G01T

## 9. Proceso completo



## Conclusión

El análisis realizado en RapidMiner permitió identificar con claridad la relación entre **tipo de pago** y **edad** como variables decisivas en el perfil de un cliente potencial.

El modelo generado cumple con los requerimientos establecidos: se filtraron correctamente los tipos de pago, se aplicó la segmentación por edad y se generó un árbol de decisión claro y coherente con el comportamiento esperado de los consumidores.

El ejercicio demuestra cómo, mediante un proceso de minería de datos estructurado, es posible transformar información dispersa en conocimiento útil para la **toma de decisiones comerciales** y la **segmentación de clientes**.