PROYECTO MÓDULO 4 - CREACIÓN DE UN ÁRBOL DE DECISIÓN CON RAPIDMINER

Esteban Ramírez Pérez

Introducción

- Se presenta un trabajo implementando el algoritmo de árbol de decisión para determinar el estado civil de una persona, esto aplicado a un dataset disponible en el enlace (1).
- Mediante la herramienta de Altair AI Studio (Rapid Miner), se realizaron algunas manipulaciones al dataset: corrección de errores ortográficos en algunas entradas, así como asignación de tipos de datos binomiales a las variables “casa” y “guapo” y el rol de “Label”/etiqueta a la variable “estado_civil”.
- Finalmente, se presentan la estructura del árbol de decisión, así como el esquema de diseño del programa implementado en el software antes mencionado

(1) https://docs.google.com/spreadsheets/d/1kAya7iKggajUdAglvOYbsQkbuXiVfocOMK6MHg4W78A/edit?usp=sharing

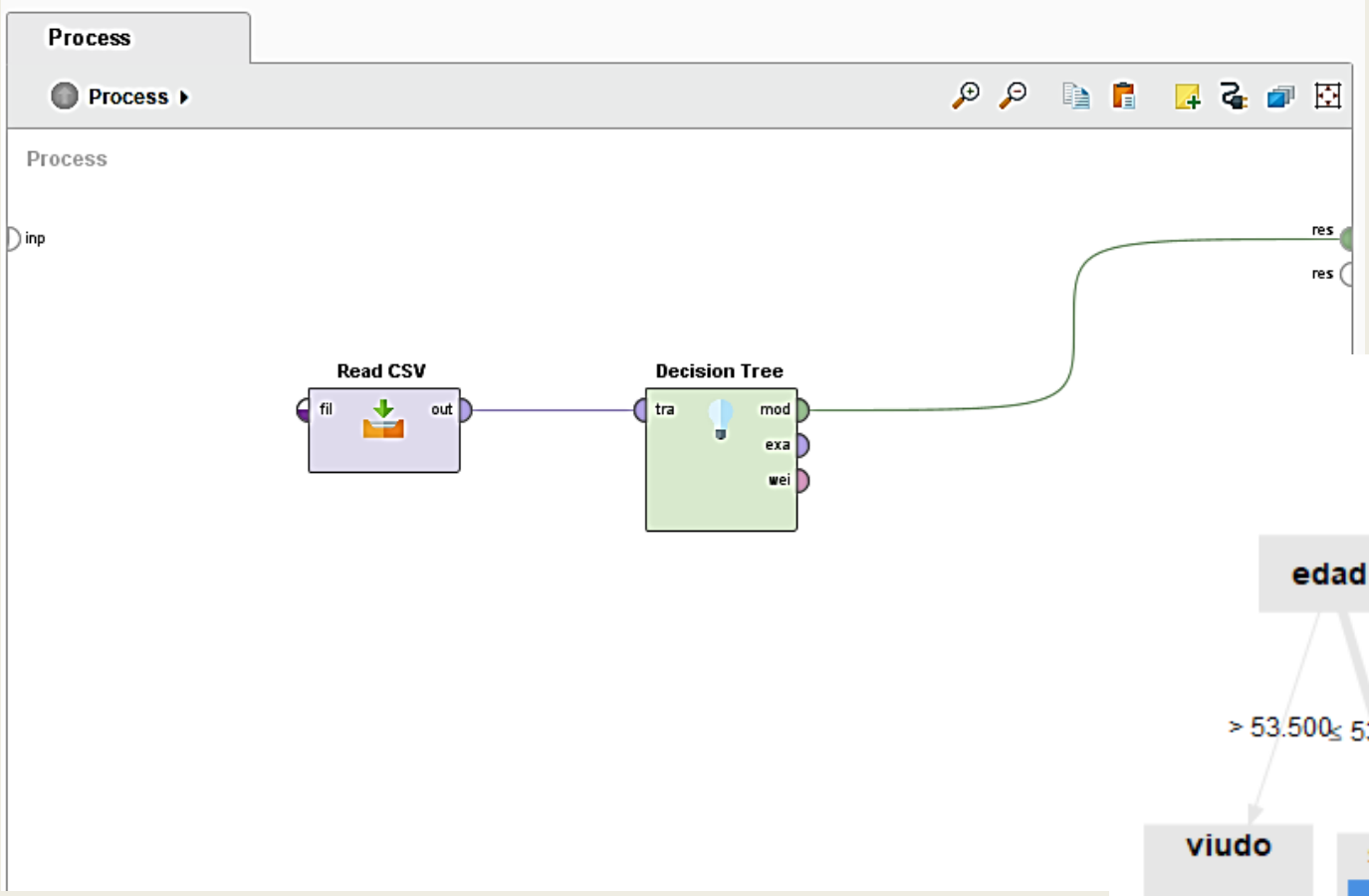
Visión general de los datos

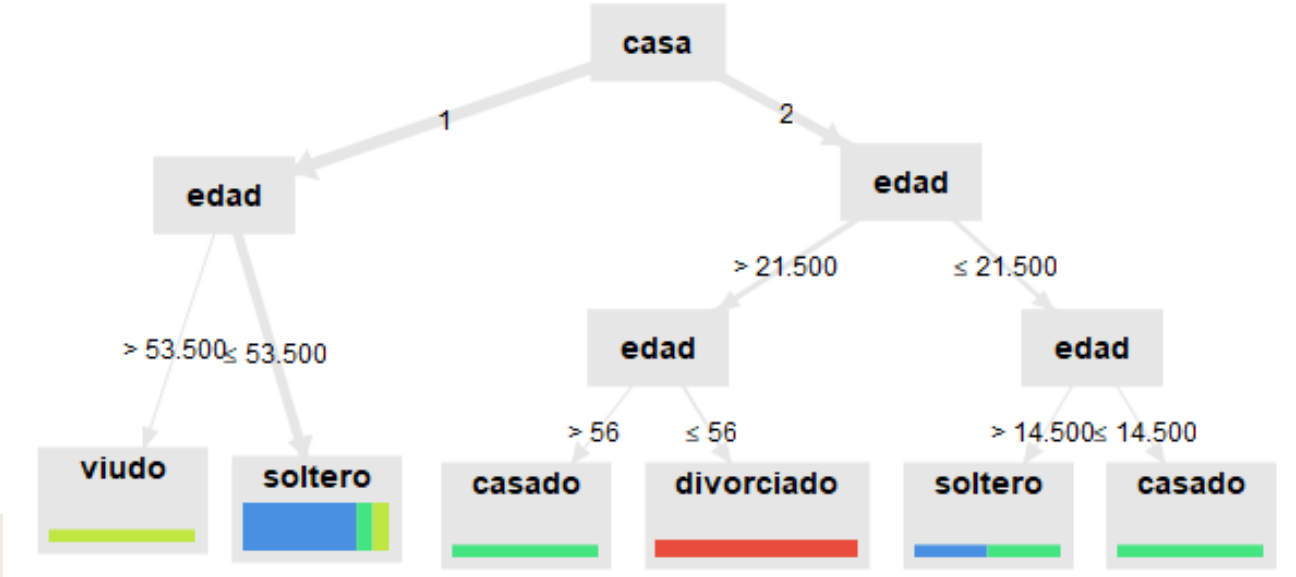
Se incluyen las siguientes columnas dentro del dataset:

- casa (variable binomial)
- guapo (variable binomial)
- edad (variable tipo entero)
- estado_civil (Variable con rol de label)
- dinero (variable tipo entero)

casa	guapo	edad	estado_civil	dinero
1	1	25	soltero	1
2	1	15	soltero	2
2	1	14	casado	3
2	1	18	casado	2
1	1	35	soltero	1
1	2	29	viudo	2
1	2	40	casado	3
1	2	65	viudo	2
1	2	42	soltero	1
1	2	12	soltero	2

Estructura de árbol de decisión y resultado del algoritmo.





[1] I. Sanchez, « Árbol De Decisión Con Excel.csv En Rapidminer,» Youtube, 2017. [En línea]. Disponible en: https://www.youtube.com/watch?v=KDvQdECxPtg.

Análisis de Resultados y Conclusiones

- Se presenta el árbol de decisión que permite analizar múltiples variables, para determinar el estado civil de una persona.
- En primera instancia, se generan dos grandes *categorías*, *los que cuentan con casa y los que no*. Se observa la gente mayor o igual a 53 años, tiene una mayor tendencia de contar con casa; dentro de este grupo, se establece que los mayores a 53 años se encuentran viudos.
- La segunda *categoría* mayor (los que no tienen casa), se divide en otros dos grupos, el primero, donde los mayores a 21 años y a su vez mayores de 56 se encuentran casados, mientras que los de edad igual o menor que 56 años, se han divorciado. Sin embargo, en el segundo grupo abarca a los que tienen 21 años o menos, los cuales están divididos en los mayores de 14 años (solteros) y en menores a 14 años y medio como "casados"; esto último siendo un claro error en la implementación de este algoritmo a este conjunto de datos en estas condiciones.
  - Es importante recordar que se está trabajando con un conjunto de datos de muestra. Por lo que esta aplicación es solo un ensayo de una aplicación del algoritmo de árbol de decisión en un conjunto de datos utilizando de forma limitada el algoritmo.
  - Esto demuestra la importancia del entrenamiento previo de los datos para observar su desempeño en un escenario real para contar con un análisis más objetivo con la finalidad de evitar sesgos en los resultados.