Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño

ccarrenovi@Gmail.com

Abril,2021

Modulo 12 Arboles de Decisión

- Introducción
- Classification And Regression Trees
- Algoritmo de Arboles de Decisión
- Ventajas y Desventajas
- Paquetes necesarios en R

Introducción

- Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción.
- Los árboles de decisión son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.

Classification And Regression Trees

- ***CART: Classification And Regression Trees***. Esta es una técnica de aprendizaje supervisado. Tenemos una variable objetivo (dependiente) y nuestra meta es ***obtener una función que nos permita predecir***, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.
- CART es una técnica con la que se pueden obtener árboles de clasificación y de regresión. Usamos ***clasificación cuando nuestra variable objetivo es discreta***, mientras que ***usamos regresión cuando es continua***.

Algoritmo de Arboles de Decisión

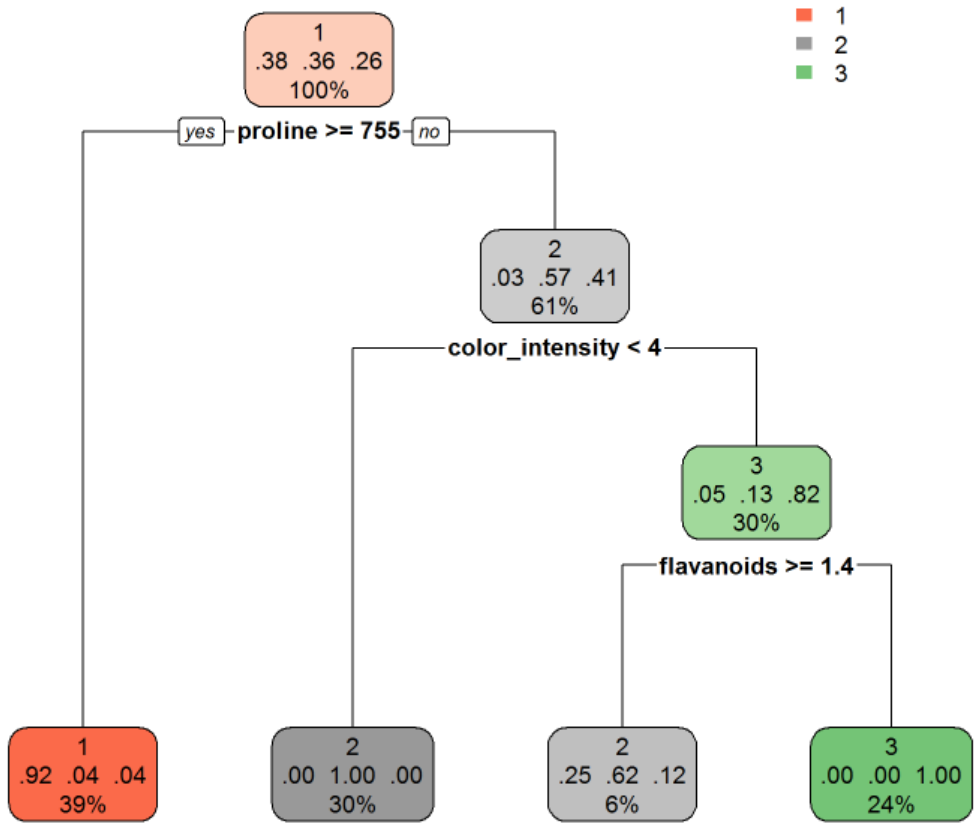
- Supongamos que nuestra variable objetivo tiene ***dos niveles***, ***deudor y no deudor***. Encontramos que la variable que mejor separa nuestros datos es ***ingreso mensual***, y la regla resultante es que ingreso ***mensual > X*** pesos. Esto quiere decir que los datos para los que esta regla es verdadera, tienen más probabilidad de pertenecer a un grupo, que al otro. En este ejemplo, digamos que si la regla es verdadera, un caso tiene más probabilidad de formar parte del grupo no deudor.

…

- Una vez hecho esto, ***los datos son separados (particionados)*** en grupos a partir de la regla obtenida. Después, para cada uno de los grupos resultantes, se repite el mismo proceso. Se busca ***la variable que mejor separa los datos en grupos***, se obtiene una regla, y se separan los datos. Hacemos esto de manera ***recursiva*** hasta que nos es imposible obtener una mejor separación. Cuando esto ocurre, el algoritmo se detiene. Cuando un grupo no puede ser partido mejor, se le llama nodo terminal u hoja.

Ejemplo: Árbol de Decisión

- Cada uno de los rectángulos representa un **nodo** de nuestro árbol, con su **regla de clasificación.**
- Cada nodo está coloreado de acuerdo a la categoría mayoritaria entre los datos que agrupa. Esta es la categoría que ha predicho el modelo para ese grupo.
- Dentro del rectángulo de cada nodo se nos muestra qué proporción de casos pertenecen a cada categoría y la proporción del total de datos que han sido agrupados allí. Por ejemplo, el rectángulo en el extremo inferior izquierdo de la gráfica tiene 92% de casos en el tipo 1, y 4% en los tipos 2 y 3, que representan 39% de todos los datos.



Ventajas y Desventajas

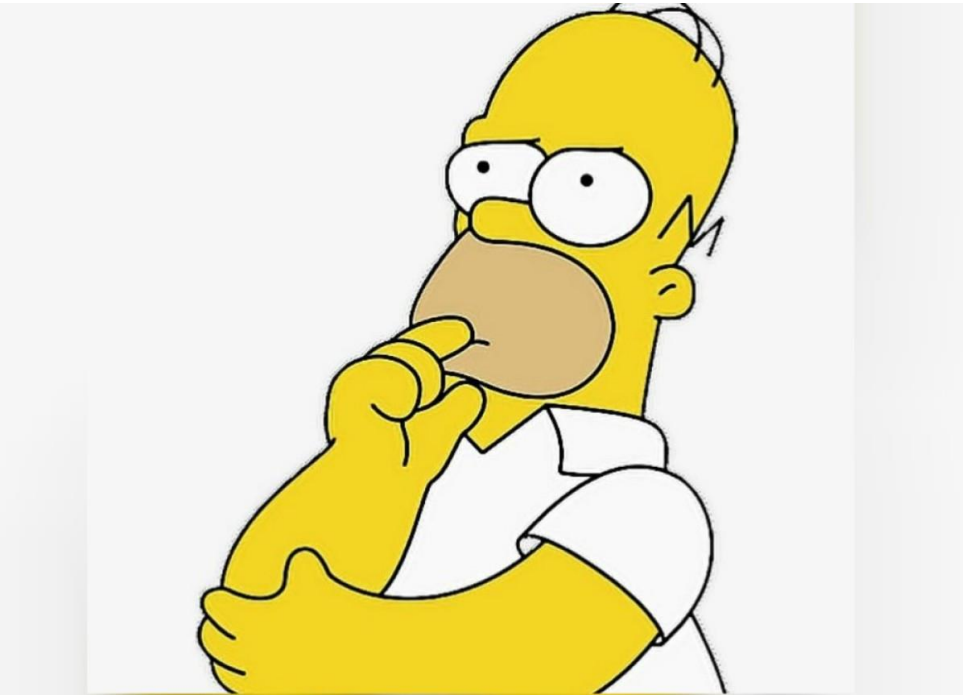
- ***Las principales ventajas*** de este método son su interpretabilidad, pues nos da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Este es un algoritmo que no es demandante en poder de cómputo comparado con procedimientos más sofisticados y, a pesar de ello, que tiende a dar buenos resultados de predicción para muchos tipos de datos.
- ***Sus principales desventajas*** son que este en tipo de clasificación "débil", pues sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo. Además es fácil ***sobre ajustar los modelos***, esto es, hacerlos excelentes para clasificar datos que conocemos, pero deficientes para datos desconocidos.

Paquetes Necesarios en R

- **tidyverse**: para llamar a la familia de paquetes tidyverse, que nos ayudaran al procesamiento de nuestros datos.
- **rpart**: el paquete con la implementación de árboles de clasificación que utilizaremos.
- **rpart.plot**: para graficar los resultados de rpart.
- **caret**: un paquete con utilidades para clasificación y regresión. Lo usaremos por su función para crear matrices de confusión

Preguntas

- Alguna pregunta?



Demo

- Realizar el ejemplo de arboles de decisión.