UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 5: Árboles de decisión

Integrantes:	Nicolás López
	Alberto Rodríguez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor:	Max Chacón
Ayudante:	Javier Arredondo

31 de Enero de 2021

Tabla de contenidos

1. Introducción

En el estudio de una población, ya sea un grupo de personas, una base de datos, entre otros, se puede ir dividiendo el espacio muestral en sub-regiones los datos que se van obteniendo al aplicar una serie de reglas o decisiones, esto itera hasta tener sub-regiones menores que integran datos de la misma clase. Este proceso se denomina árboles de decisión y es fundamental para los algoritmos automatizados que se enfocan en entrenar y conocer de forma adecuada los datos con los que desarrollan sus actividades.

En el presente laboratorio se busca trabajar una base de datos ya conocida de hongos de la familia Agaricus y Lepiota. Este data set fue extraído de la guía de campo de la sociedad Audubon. Todos los atributos del data set corresponden a variables categóricas.

El objetivo general de esta experiencia es primero procesar este conjunto de datos mencionado para luego aplicar los mencionados árboles de decisión. En específico, se debe usar el algoritmo de creación de árboles C5.0 y el lenguaje R. Para esto se tienen los siguientes objetivos específicos a cumplir:

1. Profundizar en la materia de árboles de decisión
2. Interiorizar el uso del algoritmo C5.0. En específico, la implementación en R correspondiente al paquete C50.
3. Comparar resultados con experiencias anteriores.
4. Identificar las diferencias entre el análisis mediante reglas de asociación y árboles de decisión.

A lo largo del documento, se entrega un marco teórico para introducir la materia al lector y luego se discute un pre-procesamiento realizado al set de datos, además de describir la metodología usada para obtener el árbol de decisión. Después se analiza los resultados obtenidos, realizando comparaciones con la experiencia anterior y con trabajos externos. Por último, se entrega la conclusión a partir de los resultados que se obtienen.

2. Marco Teórico

2.1. Arboles de decisión

Los arboles de decisión llaman la atención por la forma de su estructura y como se distribuyen los datos. Son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.(Vega, 2018)

Pueden existir diferentes tipos de problemas:

1. Regresión: son generalmente aquellos en los que intentamos predecir los valores de una variable continua a partir de una o más variables predictoras categóricas.

2. Clasificación: son aquellos en los que intentamos predecir los valores de una variable dependiente categórica a partir de una o más variables predictoras continuas.

Luego, se entrega la información de como está compuesto un árbol de decisión:

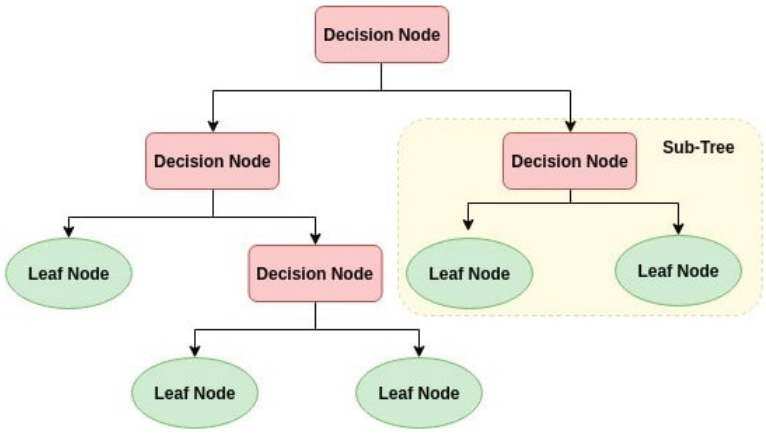

Figura 1: Árbol de decisión

- Raíz: representa a toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.

- Rama de división: cuando un nodo se divide en subnodos adicionales, se llama nodo de decisión.

- Nodo hoja: los nodos sin hijos (sin división adicional) se llaman Hoja o nodo terminal.
- Subárbol: es una subsección del árbol de decisión tomada dependiendo de las necesidades del estudio.

2.2. Medidas de selección de atributos

Esta sección va de la mano con lo que es el pre-procesamiento que se realiza a una base de datos antes de ser trabajada. Es útil para utilizar los datos que realmente aportan al estudio y también para "limpiar.[el] datasets con el cual se está estudiando.

La medida de selección de atributos es una heurística para seleccionar el criterio de división que divide los datos de la mejor manera posible. También se conoce como reglas de división porque nos ayuda a determinar puntos de interrupción para tuplas en un nodo dado.(Data, cido)

2.2.1. Ganancia de información

La ganancia de información es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con sus clasificación objetivo.

2.2.2. Gini

Gini dice que si seleccionamos dos elementos de una población al azar, entonces deben ser de la misma clase y la probabilidad de esto es 1 si la población es pura.

3. Obtención del árbol

3.1. Pre-procesamiento

En el pre-procesamiento se busca dejar los datos preparados para utilizarlos en el estudio, ya sea borrando variables o cambiando su tipo de dato. Para este laboratorio en particular, se busca preparar los datos para realizar un árbol de decisión. Primero se cambian los nombres de los datos provenientes del dataset de hongos, ya que, son poco característicos ya que solo utilizan las iniciales, por esto, la primera modificación a la base de datos es otorgar nombres que identifiquen de mejor forma los atributos.

Luego de tener los atributos ordenados, se analiza que variables deben ser eliminadas de la base de datos. Se elimina el atributo "*veil_type*" ya que solo puede tomar el valor "partial", otros atributos son "*gill_attachement*" en el cual el 97,42 % de las observaciones son "free", "*veil_color*" donde el 97,54 % pertencen a "white" y "*ring_number*" donde el 92,17 % pertenecen a "one". Como gran parte de las observaciones tienen atributos definidos por un solo valor, no aportan información al momento de clasificar comestibilidad, por lo tanto se eliminan del set de datos.

3.2. Creación del árbol mediante los paquete C50 y caret de R

Primero se dividen los datos en una muestra de entrenamiento y en otra de prueba, para esto se utiliza el comando *createDataPartition* del paquete *caret* (Kuhn, 2020). A este método se le entrega como parámetro el conjunto de datos con el atributo a clasificar, en este caso sería la clase "edibility" la cual la que indica si el hongo es comestible o venoso, esto junto al porcentaje de datos que conformaran la muestra de entrenamiento, para la cual se escoge el valor de 70 %. Esto en R se ve en la Figura 2.

```
training.index = createDataPartition(mushrooms$edibility, p=0.7)$Resample1
training.set <- mushrooms[training.index, ]
test.set <- mushrooms[-training.index, ]
```

Figura 2: Obtención de muestra de entrenamiento y prueba

Entonces se tiene *training.set* que son 5687 datos de entrenamiento y *test.set* que

son 2436 datos de prueba. Con la muestra de datos de entrenamiento se puede crear el árbol de decisión con el paquete *C50* y las reglas que se generan a partir de este (Kuhn, 2019). Implementado en R se ve en la Figura 3, donde se crea el árbol (*tree*) y las reglas (*tree.rules*).

```
tree = C5.0(edibility ~ ., training.set)
tree.rules <- C5.0(x = training.set[, -19], y = training.set$edibility, rules = T)
```

Figura 3: Creación de árbol y reglas

Ya con el árbol generado, se puede utilizar la función *predict* con la muestra de datos de prueba *test.set*, esta función lo que hace aplicarle las reglas lógicas (obtenidas por el árbol) a este conjunto de datos para decidir por cada observación si corresponde a la clase comestible o venenosa. Este con el objetivo de saber la calidad de las reglas obtenidas. La Figura 4 muestra la obtención de las variables *tree.pred.class* que entrega la clase predicha para cada dato de la muestra y *tree.pred.prob* entrega la probabilidad de pertenecer a comestible o venenoso de cada observación.

```
tree.pred.class <- predict(tree, test.set[,-19], type = "class")
tree.pred.prob <- predict(tree, test.set[,-19], type = "prob")
```

Figura 4: Aplicación de árbol y reglas al set de prueba

Finalmente cabe mencionar que se intentó con un valor menor de datos para la muestra de entrenamiento, se intento con 50 % y 30 % de los datos totales en este, pero no hubo grandes cambios en la obtención del árbol, es decir los árboles obtenidos eran similares al igual que la matriz de confusión (que se explica en la sección siguiente). No se intentó con valores menos a 30 % del total y mayor al 70 % del total de los datos, por el riesgo de que pueda existir un desajuste o un sobreajuste en la obtención del árbol y las reglas pierdan calidad.

4. Análisis de los resultados y comparación

4.1. Análisis de Árbol obtenido

El árbol obtenido se muestra en la Figura 5.

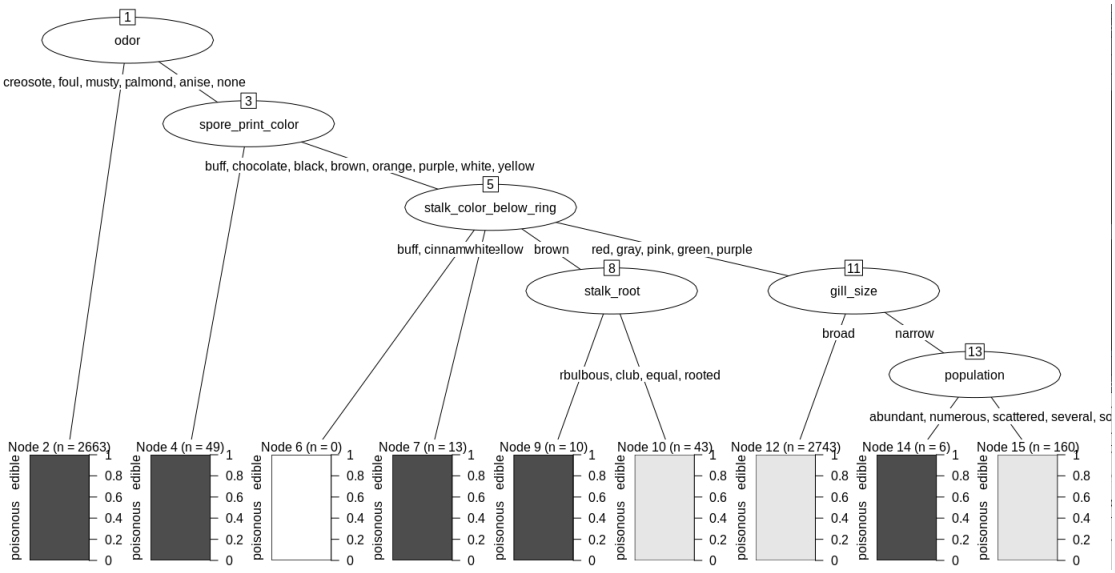

Figura 5: Árbol de Decisión

Este árbol fue generado en base a 5687 observaciones definidas en el conjunto de entrenamiento, tiene un total de 9 hojas. Las decisiones o nodos que toma el árbol se generan en base a los atributos de a continuación:

- Odor: Nodo raíz, al ser el primero se usa en un 100 %
- *spore_print_color*: Nodo intermedio, se usa un 53.17 %
- *stalk_color_below_ring*: Nodo intermedio, se usa un 52.31 %
- *gill_size*: Nodo intermedio, se usa un 51.15 %
- *population*: Nodo intermedio, se usa un 2.92 %
- *stalk_root*: Nodo intermedio, se usa un 0.93 %

El árbol de la Figura 5 se puede apreciar mejor usando la función *summary(tree)* obteniendolo de manera mas compacta como se ve en la Figura 6.

```
odor in {creosote,foul,musty,pungent,spicy,fishy}: poisonous (2663)
odor in {almond,anise,none}:
:...spore_print_color = green: poisonous (49)
    spore_print_color in {buff,chocolate,black,brown,orange,purple,white,yellow}:
    :...stalk_color_below_ring in {buff,cinnamon,yellow}: edible (0)
        stalk_color_below_ring = white: poisonous (13)
        stalk_color_below_ring = brown:
        :...stalk_root = missing: poisonous (10)
        :   stalk_root in {bulbous,club,equal,rooted}: edible (43)
        stalk_color_below_ring in {red,gray,pink,green,purple}:
        :...gill_size = broad: edible (2743)
            gill_size = narrow:
            :...population = clustered: poisonous (6)
                population in {abundant,numerous,scattered,several,solitary}: edible (160)
```

Figura 6: Árbol de Decisión más compacto

Con el árbol obtenido, se generaron entonces 8 reglas lógicas para decidir la comestibilidad de un hongo. Estas reglas se pueden ver en el Cuadro 1, como se indicó anteriormente, a la muestra de prueba *test.set* se le fueron aplicadas estas reglas, para decidir por cada dato si este era comestible o venenoso.

N°	Antecedentes	Consecuentes	Lift
1	odor=$\{$almond, anise, none$\}$ $\wedge gill\_size = broad \wedge spore\_print\_color = \{buff, black, brown, orange, white, yellow\}$	Comestible	1.9
2	odor=$\{$almond, anise, none$\}$ $\wedge stalk\_color\_below\_ring = \{gray, pink, green, purple\}$ $\wedge$ $spore\_print\_color = \{buff, chocolate, black, brown, orange, purple, white, yellow\}$ $\wedge$ $population = \{abundant, numerous, scattered, several, solitary\}$	Comestible	1.9
3	odor=$\{$almond, anise, none$\}$ $\wedge stalk\_root = \{bulbous, club, equal, rooted\}$ $\wedge$ $spore\_print\_color = \{black, brown, purple, white\}$	Comestible	1.9
4	odor=$\{$creosote, foul, musty, pungent, spicy, fishy$\}$	Venenoso	2.1
5	spore_print_color=green	Venenoso	2.0
6	gill_size=narrow $\wedge population = clustered$	Venenoso	1.9
7	stalk_color_below_ring=white	Venenoso	1.9
8	stalk_root = missing $\wedge stalk\_color\_below\_ring = brown$	Venenoso	1.9

Cuadro 1: Reglas Obtenidas

Se ve también en la Figura 7 la probabilidad que tiene un cierto dato a pertenecer a la clase comestible y venenosa.

```
          edible    poisonous
8  0.0001944533 0.9998055467
10 0.9998243526 0.0001756474
11 0.9998243526 0.0001756474
12 0.9998243526 0.0001756474
16 0.9998243526 0.0001756474
20 0.9998243526 0.0001756474
```

Figura 7: Probabilidad de predicción de clase para algunas observaciones

Con esto se obtuvo la matriz de confusión que se ve en el Cuadro 2, se puede ver que la clasificación fue correcta en la totalidad de las observaciones, no se encontró ningún error, por lo que la exactitud, sensibilidad, especificidad, valor de predicción negativo y valor de predicción positivo tienen valor 1 o 100 %. Entonces las reglas obtenidas funcionaron perfectamente en el conjunto de pruebas. Aunque no funcionó al 100 % en el conjunto de entrenamiento, ya que se encontraron 6 hongos que fueron clasificados como comestibles siendo que eran venenosos (esto se puede apreciar usando la función *summary(tree.rules)*). Por lo tanto aunque las reglas tiene una alta exactitud, no son 100 % exactas.

	Comestible	Venenoso
Comestible	1262	0
Venenoso	0	1174

Cuadro 2: Matriz de confusión

4.2. Comparación con las reglas de asociación

Comparando las reglas obtenidas del árbol de decisión, con las reglas las generadas mediante reglas de asociación, se tiene que hay 2 atributos que se repiten en ambas reglas, estos atributos son: “odor”y “gill_size”. En ambos conjuntos de reglas, el atributo “odor” es el que más datos abarca y el único que por si solo genera una reglas sin necesidad de estar con otro atributo.

Se puede analizar también la métrica de calidad lift de cada regla obtenida con el árbol de decisión, con el lift de cada una de las reglas de asociación, en las cuales en la segunda el máximo valor obtenido por el lift es de 1,8. Mientras en las reglas obtenidas por el árbol el lift varía entre 1.9 y 2.1, lo que significa que los antecedentes obtenidos son mejor aún para encontrar el consecuente indicado.

Como último punto, las reglas obtenidas en este laboratorio son mucho más completas y específicas, esto se debe a que gracias al árbol de decisión se generan reglas lógicas utilizando todos los valores de un atributo, ya que debe explorar todos los caminos para generar los nodos. En las de asociación, en cambio, solo se escoge un valor por atributo.Por ejemplo aquí se puede utilizar odor = (almond, anise, none), y en las de asociación solo odor = none, Lo que amplía las posibilidades para la predicción de la clase, y lo específica que pueden ser las reglas para obtener resultados aún más precisos.

Para finalizar el análisis, se pudo observar que la primera regla del árbol es la misma que la regla más importante propuesta por Wlodzislaw en su investigación (Wlodzislaw D, 1997), esta es $odor = (almond \vee anise \vee none)$, aunque aquí aparece acompañada de 2 atributos más.

5. Conclusión

A modo de conclusión, hay que destacar que los métodos utilizados en los laboratorios son útiles al momento de generar agrupaciones de datos y diferentes implementaciones en los datasets. Sin embargo, depende del tipo de datos que se esté trabajando para escoger que método es más útil para la aplicación de este. Se pudo analizar, que el árbol de decisión fue uno de las mejores implementaciones para el datasets de hongos, generando reglas específicas y completas. Cabe destacar, que el método va a ser más o menos útil dependiendo de la base de datos con la cual se trabaje.

También, se puede concluir que la aplicación de árboles de decisión para este estudio entrega un resultado específico en comparación a las experiencias pasadas, como por ejemplo el método de clustering o reglas de asociación, que daban una respuesta agrupando los datos pero que no eran descriptivas o específicas en si.

Por otro lado, analizando los resultados y tomando en cuenta que los datos son todos categóricos, se recomienda a los futuros estudios con distintos datasets, utilizar árboles de decisión, ya que se demostró que entregan agrupaciones correctas y con alta precisión para bases de datos completamente categóricas.

Bibliografía

Data, S. B. (Año desconocido). *Árbol de decisión en Machine Learning*.

Kuhn, M. (2019). *C5.0 Decision Trees and Rule-Based Models*.

Kuhn, M. (2020). *createDataPartition*.

Vega, J. B. M. (2018). *Árbol de decisión con R*.

Wlodzislaw D, Adamczack R, G. K. (1997). *Extraction of crisp logical rules using constructive constrained backpropagation networks*. Proceedings of International Conference on Neural Networks (ICNN'97).