

|  |
| --- |
| **FACULTAD DE MATEMÁTICAS** |
| **Minería de Datos** |

**Unidad III:Métodos de Clasificación**

**ADA 6: *Árboles de Decisión.***

*Licenciatura en Actuaria.*

**Integrantes:**

- Álvarez Herrera Samantha

- Ciau Puga Abigail

- Colonia Espinosa Cindy

- Fernández Caro Frida

- Padilla Jiménez Meybor

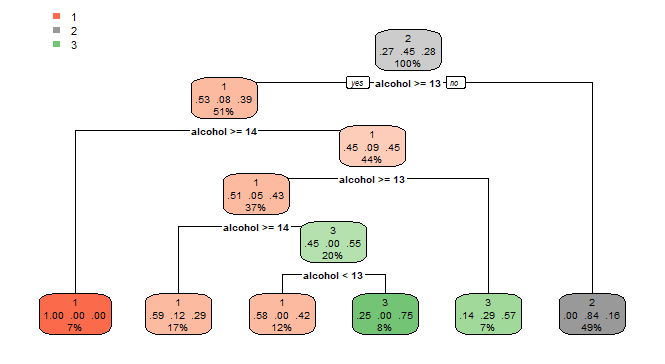
- Sobrino Bermejo Samantha

**M.C. Ernesto Guerrero Lara**

*Para los siguientes análisis se usó la base de datos vinos.csv, la cual tiene 178 observaciones y 15 variables.*

1. **Clasifica el tipo de vino considerando la variable alcohol y 100 datos de entrenamiento.**

Para este primer inciso, se buscó clasificar la variable *alcohol*, la cual es numérica y cuyos valores se encuentran entre 11.03 y 14.83. La base de entrenamiento tiene un total de 100 observaciones, las cuales son tomadas de una muestra aleatoria.

El árbol de decisión obtenido en la sección de entrenamiento es el siguiente:

En primera instancia podemos destacar que cada rectángulo presenta un nodo de nuestro árbol con su respectiva regla de clasificación. Si nos fijamos en el nodo base podemos notar que la regla de clasificación es si el alcohol es mayor o igual a 13, para los datos que cumplen que SÍ, se abre el nodo de la izquierda, y para los que NO, se abre el nodo de la derecha. Siguiendo en el nodo base podemos ver que 27% de los datos se encuentran en la categoría 1, 45% se encuentran en la categoría 2 y 28% en la categoría 3, siendo así que la mayoría de los datos que agrupo se encuentran en la categoría 2, por lo tanto, su color corresponde al del grupo 2.

Hay que destacar que el primer rectángulo de la última fila de rectángulos, de izquierda a derecha, obtuvo un 100% de clasificaciones correctas para la categoría 1 y cuyos datos clasificados representan el 7% de todos los datos tomados para la muestra de entrenamiento. El siguiente rectángulo que obtuvo el mayor porcentaje de clasificaciones correctas es el último, de la última fila de rectángulos, de izquierda a derecha, con un 84% de clasificaciones correctas para la categoría 2 y cuyos datos clasificados representan al 49% de todos los datos, es decir, casi el 50% de los datos se encuentran clasificados en la categoría 2.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 2 | 3 |
| 1 | 28 | 3 | 8 |
| 2 | 0 | 21 | 4 |
| 3 | 4 | 2 | 8 |

Posteriormente, para la etapa de predicción, se utilizaron 78 datos. La matriz de confusión fue la siguiente:

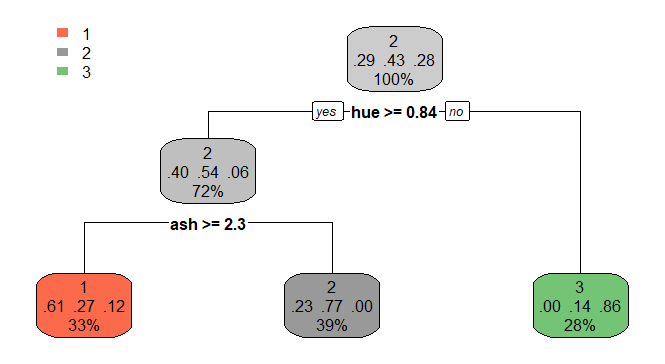
De la matriz anterior podemos destacar que de los 39 datos que pertenecen a la *categoría 1*, 28 fueron clasificados correctamente y 11 de manera incorrecta, de entre los cuales 3 se clasificaron como categoría 2 y 8 de categoría 3. Para la *categoría 2*, 21 de los 25 datos fueron clasificados correctamente, 4 fueron incorrectos y puestos en la categoría 3 y ninguno en la categoría 1. De la *categoría 3*, 8 de los 14 datos fueron clasificados correctamente y 6 de manera incorrecta de los cuales 4 se clasificaron en la categoría 1 y 2 en la categoría 2.

Para finalizar en análisis se obtuvo la **exactitud** de este modelo, la cual fue de 73.0769%, es decir, casi ¾ de los datos son clasificados de manera correcta.

1. **Clasifica el tipo de vino considerando las variables ash y hue y 100 datos de entrenamiento.**

Para este segundo inciso, se buscó clasificar las variables ash y hue (fresno y matiz), las cuales son numéricas y cuyos valores se encuentran en los intervalos (1.36, 3.23) y (0.48,1.71) respectivamente. La base de entrenamiento tiene un tamaño de 100 observaciones las cuales son tomadas a manera de muestra aleatoria.

El árbol obtenido en la sección de entrenamiento es el siguiente:

****

Fijándonos en el nodo base podemos notar que la regla de clasificación es: si el *Matiz (hue)* es **mayor o igual** **a 0.84**, para los datos que cumplen que SÍ, se abre el nodo de la izquierda y para los que NO, se abre el nodo de la derecha, el cual ya forma uno de los grupos finales. Aún en el nodo base, podemos ver que 29% de los datos se encuentran en la categoría 1, 43% se encuentran en la categoría 2 y 28% en la categoría 3, siendo así que la mayoría de los datos que agrupo se encuentran en la categoría 2 por lo tanto el color del nodo es el correspondiente al grupo 2.

En el siguiente nodo, donde los datos que fueron mayor o igual a *hue* 0.84, podemos ver que 40% de los datos fueron clasificados en la categoría 1, 54% en la categoría 2 y solo el 6% en la categoría 3. En este nodo se encuentra la siguiente ramificación, la cual toma como referencia a la variable *ash* y clasifica de acuerdo a si esta es mayor o igual a 2.3.

Como resultado se obtienen dos nodos finales más. Observamos que, junto con el que se formó al tener un valor menor a 0.84 para la variable *hue*, se tiene uno para cada categoría.

En el primer rectángulo, de izquierda a derecha, observamos que 61% de los datos fueron clasificados de manera correcta, es decir, en la primera categoría, 27% en la segunda y 12% en la tercera. Estos datos representan el 33% de los datos usados en la base de entrenamiento. Observando al nodo que se encuentra siguiente, observamos que 77% de los datos fueron clasificados de manera correcta, es decir, en la categoría 2 y el total usado en este nodo representa 39% de los datos totales empleados. El último nodo nos muestra el 86% fueron clasificados de manera correcta, es decir, en la categoría 3.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 2 | 3 |
| 1 | 21 | 11 | 2 |
| 2 | 8 | 12 | 0 |
| 3 | 1 | 5 | 18 |

Para la siguiente parte, la de predicción, se utilizaron 78 datos. La matriz de confusión fue la siguiente:

De la matriz podemos destacar que, de los 34 datos que pertenecen a la categoría 1, 21 fueron clasificados correctamente y 13 incorrectos de entre los cuales 11 se clasificaron como categoría 2 y 2 de categoría 3. Para la categoría 2, 12 de los 20 datos fueron clasificados correctamente, 8 fueron incorrectos y puestos en la categoría 1 y ninguno en la categoría 3. De la categoría 3, 18 de los 24 datos fueron clasificados correctamente y 6 de manera incorrecta de los cuales 1 se clasifico en la categoría 1 y 5 en la categoría 2.

Para finalizar en análisis se obtuvo la **exactitud** de este modelo, la cual fue de 65.3846%.

1. **Clasifica el tipo de vino considerando las variables alcohol y magnesio. Realiza 3 muestras de entrenamiento de tamaño 140.**

Se buscó clasificar las variables alcohol y magnesio, las cuales son numéricas y cuyos valores se encuentran en los intervalos (11.03, 14.83) y (70.00, 162.00) respectivamente. La base de entrenamiento tiene un tamaño de 140 observaciones las cuales son tomadas a manera de muestra aleatoria.

**Muestra 1.**

El árbol obtenido en la sección de entrenamiento es el siguiente:

En el nodo base podemos notar que la regla de clasificación es si el alcohol es mayor o igual a 12.78, para los datos que cumplen que SÍ, se abre el nodo de la izquierda, y para los que NO, se abre el nodo de la derecha. Siguiendo en el nodo base podemos ver que el 35% de los datos se encuentran en la categoría 1, 37% se encuentran en la categoría 2 y 28% en la categoría 3, siendo así que la mayoría de los datos que agrupo se encuentran en la categoría 2, por lo tanto, su color corresponde al del grupo 2. Como resultado se obtienen 8 nodos finales.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 2 | 3 |
| 1 | 7 | 0 | 3 |
| 2 | 0 | 19 | 3 |
| 3 | 3 | 0 | 3 |

Para la siguiente parte, la de predicción, se utilizaron 38 datos. La matriz de confusión fue la siguiente:

De la matriz podemos destacar que de los 10 datos que pertenecen a la categoría 1, 7 fueron clasificados correctamente y 3 incorrectos, los cuales se clasificaron como categoría 3. Para la categoría 2, 22 de los 19 datos fueron clasificados correctamente, 3 fueron incorrectos y puestos en la categoría 3 y ninguno en la categoría 1. De la categoría 3, 3 de los 6 datos fueron clasificados correctamente, los 3 datos incorrectos se clasificaron en la categoría 1.

Para finalizar en análisis se obtuvo la **exactitud** de este modelo, la cual fue de 76.3175%.

**Muestra 2.**

El árbol obtenido en la sección de entrenamiento es el siguiente:

En el nodo base podemos notar que la regla de clasificación es si el alcohol es mayor o igual a 13, para los datos que cumplen que SÍ, se abre el nodo de la izquierda, y para los que NO, se abre el nodo de la derecha. Siguiendo en el nodo base podemos ver que el 33% de los datos se encuentran en la categoría 1, 40% se encuentran en la categoría 2 y 27% en la categoría 3, siendo así que la mayoría de los datos que agrupo se encuentran en la categoría 2, por lo tanto, su color corresponde al del grupo 2. Como resultado se obtienen 7 nodos finales.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 2 | 3 |
| 1 | 10 | 0 | 1 |
| 2 | 0 | 15 | 4 |
| 3 | 3 | 0 | 5 |

Para la siguiente parte, la de predicción, se utilizaron 38 datos. La matriz de confusión fue la siguiente:

De la matriz podemos destacar que de los 11 datos que pertenecen a la categoría 1, 10 fueron clasificados correctamente y 1 incorrecto, el cual se clasificó como categoría 3. Para la categoría 2, 15 de los 19 datos fueron clasificados correctamente, 4 fueron incorrectos y puestos en la categoría 3 y ninguno en la categoría 1. De la categoría 3, 5 de los 11 datos fueron clasificados correctamente, los 3 datos incorrectos se clasificaron en la categoría 1.

Para finalizar en análisis se obtuvo la **exactitud** de este modelo, la cual fue de 78.9473%.

**Muestra 3.**

El árbol obtenido en la sección de entrenamiento es el siguiente:



En el nodo base podemos notar que la regla de clasificación es si el alcohol es mayor o igual a 12.78, para los datos que cumplen que SÍ, se abre el nodo de la izquierda, y para los que NO, se abre el nodo de la derecha. Siguiendo en el nodo base podemos ver que el 33% de los datos se encuentran en la categoría 1, 39% se encuentran en la categoría 2 y 28% en la categoría 3, siendo así que la mayoría de los datos que agrupo se encuentran en la categoría 2, por lo tanto, su color corresponde al del grupo 2.

Como resultado se obtienen 8 nodos finales.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 1 | 2 | 3 |
| 1 | 3 | 1 | 2 |
| 2 | 0 | 15 | 3 |
| 3 | 10 | 0 | 4 |

Para la siguiente parte, la de predicción, se utilizaron 38 datos. La matriz de confusión fue la siguiente:

De la matriz podemos destacar que de los 6 datos que pertenecen a la categoría 1, 3 fueron clasificados correctamente y 3 incorrectos, uno fue clasificado en la categoría 2 y dos en la categoría 3. Para la categoría 2, 15 de los 18 datos fueron clasificados correctamente, tres fueron incorrectos y puestos en la categoría 3 y ninguno en la categoría 1. De la categoría 3, 4 de los 14 datos fueron clasificados correctamente, los 10 datos incorrectos se clasificaron en la categoría 1.

Para finalizar en análisis se obtuvo la **exactitud** de este modelo, la cual fue de 57.8947%.

**Tabla 1.** Exactitud obtenida

|  |  |
| --- | --- |
| Muestra | Exactitud |
| 1 | 76.3175% |
| 2 | 78.9473% |
| 3 | 57.8947% |

La salida en R de la exactitud para cada muestra se presenta en la Tabla 1. De acuerdo con los resultados **el mejor modelo se obtiene con la muestra 2** ya que es la que tiene una **mejor exactitud**, siendo del 78.9473%.