Ejercicio 1

Patricio Guledjuan Azul Noguera Rocio Gonzalez

2023-08-22

Datos Obtenidos

Trabajaremos con una variedad de conjuntos de datos que desempeñarán un papel fundamental en nuestra experimentación. Estos conjuntos de datos incluyen los de *Churn* y *Heart*, proporcionados según las instrucciones, así como un nuevo conjunto de datos adquirido de la plataforma UC Irvine. Esta plataforma aloja 643 conjuntos de datos como parte de su servicio a la comunidad de aprendizaje automático.

Caracteristicas de los Datasets

Churn

Este conjunto de datos se recopiló al azar de la base de datos de una empresa de telecomunicaciones iraní durante 12 meses. Contiene información de 3150 clientes en 13 columnas, incluyendo detalles como fallas en llamadas, frecuencia de SMS, quejas, duración de suscripción, entre otros. Los atributos son principalmente datos agregados de los primeros 9 meses, mientras que las etiquetas de abandono representan el estado de los clientes al final de los 12 meses, con un período de planificación de tres meses. La taza de abandono *Churn* será la variable a predecir.

El enlace siguiente les dará acceso al conjunto de datos:

Enlace: https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset

Heart

El conjunto de datos *Hearts* almacena información relacionada con pacientes y su estado en términos de Enfermedades Cardiacas. Las investigaciones realizadas con la base de datos de Cleveland se han centrado en la tarea de distinguir entre la presencia y la ausencia de enfermedades cardíacas. Por lo tanto, la variable objetivo que se busca predecir es "HeartDisease", la cual indica si el paciente presenta o no una enfermedad del corazón.

El enlace siguiente les dará acceso al conjunto de datos:

Enlace: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Wine Quality

El conjunto de datos que estamos considerando está vinculado a la variante blanca del vino portugués "Vinho Verde". En aras de abordar asuntos de privacidad y facilitar aspectos logísticos, hemos optado por hacer disponibles exclusivamente las variables de naturaleza fisicoquímica (entradas) y sensorial (salida).

Resulta esencial enfocarnos en ciertos aspectos clave del conjunto de datos en cuestión:

En primer lugar, contamos con un total de 4898 observaciones. Dentro de estas observaciones, se han registrado 12 variables que desempeñan distintos roles en el análisis. Es relevante notar que todas estas variables son de tipo numérico, lo que ofrece una base sólida para un análisis cuantitativo. Adicionalmente, es importante destacar que no se ha identificado la presencia de valores faltantes en el conjunto de datos, lo cual fortalece su integridad y confiabilidad.

El enlace siguiente les dará acceso al conjunto de datos específico que hemos seleccionado para este proyecto:

Enlace: https://archive.ics.uci.edu/dataset/186/wine+quality

Predicción

Nuestro objetivo consistirá en predecir, a partir de las variables predictoras, la calidad del vino. En este contexto, un valor de 1 indicará que el vino posee una calidad favorable, mientras que un valor de 0 indicará lo contrario. Esta tarea de predicción no es de naturaleza trivial debido a que las variables empleadas para este propósito no solo presentan dificultades en su interpretación individual, sino también porque las diferentes variables pueden interactuar de manera compleja y entrelazada.

Modificaciones

Dado que la variable objetivo del conjunto de datos que descargamos no presentaba una estructura de clasificación binaria, fue necesario llevar a cabo una adaptación en nuestra base de datos. En el conjunto de datos original, se encontraba una columna denominada 'quality', la cual reflejaba la calidad del vino mediante un valor numérico que oscilaba entre 1 (representando el vino de menor calidad) y 10 (indicando el vino de mayor calidad). Ante esta situación, se optó por ajustar esta variable predictora para adecuarla a nuestra problemática de clasificación binaria.

La conversión de una variable numérica a una forma binaria implica encontrar un punto de corte que divida los valores en dos categorías claramente definidas. Para determinar el punto de corte más adecuado, se requiere un análisis exhaustivo de diversas características de los datos. Esto incluye la consideración de la distribución de los valores, la detección de valores atípicos y la comprensión de la dispersión de los datos, entre otros factores.

En este sentido, implementaremos un análisis cuantitativo que involucra el cálculo de la media y la mediana de los valores presentes en la columna *Quality* de las 4898 instancias de nuestro conjunto de datos. Estos valores centrales pueden ser puntos de referencia para ubicar el punto de corte, ya sea en uno de ellos o en una posición intermedia entre los cuartiles.

Además, procederemos con la elaboración de un histograma que permitirá examinar la distribución de los valores numéricos en *Quality*. Este enfoque busca identificar regiones en las que los valores tiendan a agruparse en dos grupos diferenciados. Si encontramos una separación clara en estos grupos, ese punto podría ser considerado como un posible candidato para el umbral.

La utilización de un boxplot, también conocido como diagrama de caja, será una herramienta valiosa para identificar valores atípicos y la variabilidad de los datos. En particular, pondremos especial atención en buscar indicios de una división entre los valores inferiores y superiores. No es improbable que el valor que se ubica en el extremo superior de la "caja" pueda ser evaluado como una opción viable para el punto de corte.

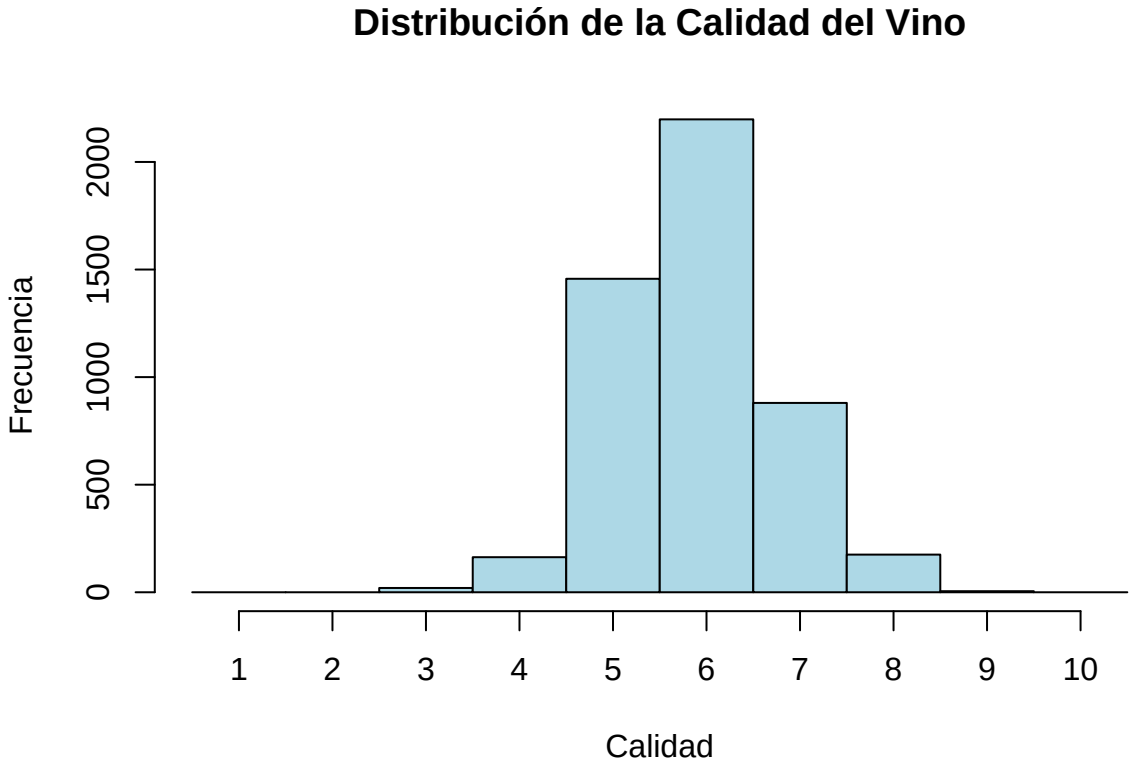
Mediana La mediana de la calidad del vino es:

```
## [1] 6
```

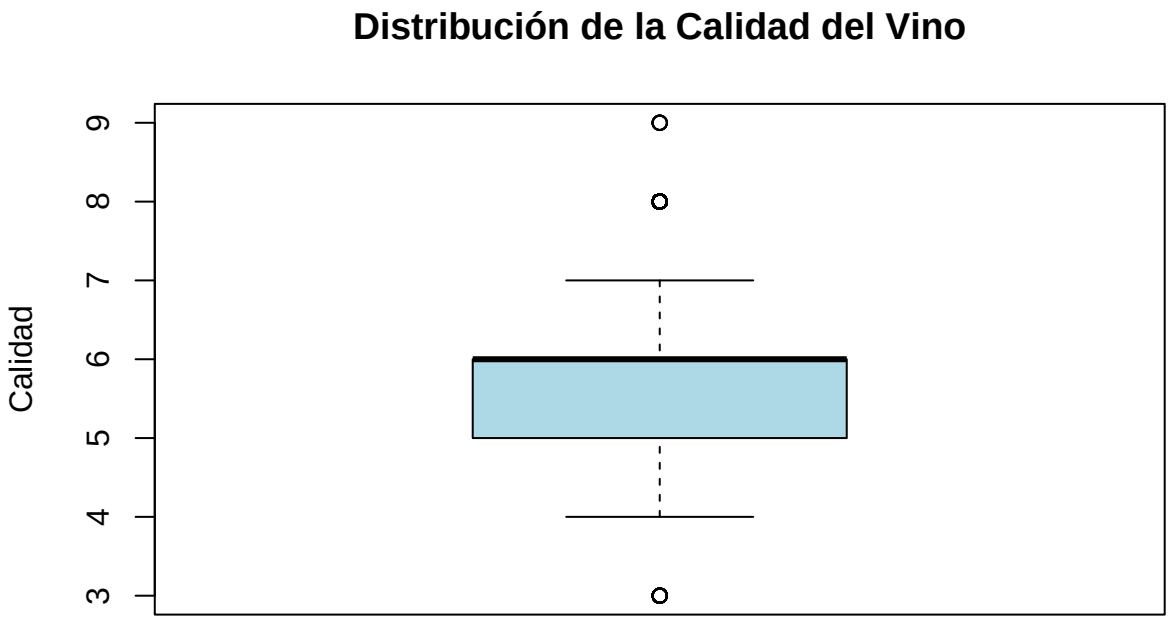
Promedio El promedio de la calidad del vino es:

```
## [1] 5.877909
```

Histograma Examinemos la distribución de los datos a través de un histograma.



BoxPlot Analizaremos la variabilidad de los datos a través de un diagrama de caja.



Debido a que los datos exhiben una distribución continua y no muestran una segregación evidente en grupos distintos a través del histograma o el boxplot, hemos optado por determinar el umbral adecuado para la transformación de la variable numérica en una variable binaria basándonos en el cálculo del promedio y la mediana. Esta elección nos pareció acertada debido a que estas medidas centrales proporcionan una representación equilibrada de la distribución de los datos.

Seleccionar el promedio y la mediana como puntos de corte nos brinda ciertas ventajas. El promedio ofrece una evaluación de la tendencia central de los datos, mientras que la mediana proporciona una división en la mitad de la distribución. La combinación de ambas perspectivas nos permite obtener una comprensión más completa de la ubicación y la dispersión de los valores.

Basándonos en estos resultados, hemos llegado a la conclusión de que el valor 6 es un umbral adecuado para llevar a cabo nuestra clasificación binaria. Esta división demarcara la distinción entre un vino de calidad superior y otro de calidad inferior. La selección de este punto de corte fue llevada a cabo con el objetivo de lograr una distribución equitativa de los datos en ambas categorías de calidad.

En la siguiente etapa, empleamos los conocimientos obtenidos a partir de los cálculos anteriores para transformar nuestra variable objetivo. Esta variable tomará el valor 1 si 'quality' es mayor a 6, y el valor 0 en caso contrario.

Código de modificación

Los valores de *Quality* solian presentarse de esta forma:

```
##  [1] 6 6 6 6 6 6 6 6 6 6 5 5 5 7 5 7 6 8 6 5
```

```
# Recorremos la columna Quality y modificar los valores
for (i in 1:length(datos$quality)) {
  if (datos$quality[i] >= 6) {
    datos$quality[i] <- 1
  } else {
    datos$quality[i] <- 0
  }
}
suppressWarnings({
  ruta_archivo_nuevo <- "./data/winequality_modificados.csv"
  write.csv(datos, file=ruta_archivo_nuevo, sep = ',', row.names=FALSE)
})
```

Con la siguiente modificación podemos ver como los valores mayores o iguales a 6 fueron reemplazados por un 1, y los valores menores a 6 por un 0.

```
##  [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 1 0 1 1 1 1 0
```

Nuestro dataset no vino en formato tradicional ya que el separador usado era ';'. De todas maneras, al modificar la columna como hicimos en el ultimo código, aprovechamos para escribir el nuevo csv con separador ','. Luego de eso, el dataset quedo en formato tradicional.

A continuación, presentamos una vista previa inicial del conjunto de datos mediante la visualización de las primeras filas:

```
##    fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1            7.0             0.27        0.36           20.7     0.045
## 2            6.3             0.30        0.34            1.6     0.049
## 3            8.1             0.28        0.40            6.9     0.050
## 4            7.2             0.23        0.32            8.5     0.058
## 5            7.2             0.23        0.32            8.5     0.058
## 6            8.1             0.28        0.40            6.9     0.050
## 7            6.2             0.32        0.16            7.0     0.045
```

```
## 8            7.0             0.27        0.36           20.7     0.045
## 9            6.3             0.30        0.34            1.6     0.049
## 10           8.1             0.22        0.43            1.5     0.044
##    free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                   45                  170  1.0010 3.00      0.45     8.8
## 2                   14                  132  0.9940 3.30      0.49     9.5
## 3                   30                   97  0.9951 3.26      0.44    10.1
## 4                   47                  186  0.9956 3.19      0.40     9.9
## 5                   47                  186  0.9956 3.19      0.40     9.9
## 6                   30                   97  0.9951 3.26      0.44    10.1
## 7                   30                  136  0.9949 3.18      0.47     9.6
## 8                   45                  170  1.0010 3.00      0.45     8.8
## 9                   14                  132  0.9940 3.30      0.49     9.5
## 10                  28                  129  0.9938 3.22      0.45    11.0
##    quality
## 1        1
## 2        1
## 3        1
## 4        1
## 5        1
## 6        1
## 7        1
## 8        1
## 9        1
## 10       1
```