Práctica 2: Limpieza y análisis de datos

Estudiante: Edison Vicente

Descripción del dataset

El presente conjunto de datos es obtenido desde el repositorio de Kaggle, el conjunto de datos corresponde a variantes rojas y blancas del vino portugués "Vinho Verde", contiene variables fisicoquímicas y sensoriales. El conjunto de datos está compuesto por 12 variables y 1599 registros. Las variables que incluye el conjunto de datos son los siguientes:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality

Importancia

El conjunto de datos puede ser utilizado para tareas de clasificación y regresión. Para este caso práctico se desea determinar la calidad en base a sus propiedades fisicoquímicas. Los análisis pueden ser de utilidad en el sector vitivinícola o determinar la calidad de otros productos que también incluyan variables fisicoquímicas como el cacao, por ejemplo.

Selección de datos de interés

En este caso se utiliza todos los registros del conjunto de datos ya que tiene un número de registros manejable para la práctica así también se considera todos los atributos.

Limpieza de datos

A continuación, se realiza procesos de limpieza para ello vamos a utilizar la herramienta RStudio. En primera instancia se realiza la lectura del fichero en formato CSV:

```
> # Lectura de datos
> dataset <- read.csv("D:/DATOS MASTER/winequality-red.csv", header = TRUE)
> head(dataset)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density
1           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978
2           7.8             0.88        0.00            2.6     0.098                  25                   67  0.9968
3           7.8             0.76        0.04            2.3     0.092                  15                   54  0.9970
4          11.2             0.28        0.56            1.9     0.075                  17                   60  0.9980
5           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978
6           7.4             0.66        0.00            1.8     0.075                  13                   40  0.9978
     pH sulphates alcohol quality
1 3.51      0.56     9.4       5
2 3.20      0.68     9.8       5
3 3.26      0.65     9.8       5
4 3.16      0.58     9.8       6
5 3.51      0.56     9.4       5
6 3.51      0.56     9.4       5
```

Luego se verifica el tipo de datos que incluye cada variable la cual corresponde al dominio de las mimas:

```
> #Tipo de datos de cada atributo
> sapply(dataset, function(x) class(x))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
           "numeric"           "numeric"            "numeric"           "numeric"            "numeric"
 free.sulfur.dioxide total.sulfur.dioxide              density                  pH            sulphates
           "numeric"           "numeric"            "numeric"           "numeric"            "numeric"
             alcohol             quality
           "numeric"           "integer"
```

Verificación de datos vacíos

Es común utilizar el cero para denotar valores vacíos en otros casos NA. Para este conjunto de datos no es necesario aplicar aquello ya que los campos no incluyen valores vacíos, como se observa a continuación:

```
> #Varificación de valores vacios
> sapply(dataset, function(x) sum(is.na(x)))
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
                   0                   0                    0                   0                    0
 free.sulfur.dioxide total.sulfur.dioxide              density                  pH            sulphates
                   0                   0                    0                   0                    0
             alcohol             quality
                   0                   0
```

Valores extremos

Los valores extremos son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Con el propósito de identificar estos valores nos valdremos de la función **`boxplot.stats()`** de R.

```
> #Valores extremos
> boxplot.stats(dataset$fixed.acidity)$out
 [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6
[24] 12.5 13.0 12.5 13.3 12.4 12.5 12.9 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9 13.3
[47] 12.9 12.6 12.6
> boxplot.stats(dataset$volatile.acidity)$out
 [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035 1.025 1.115 1.020 1.020 1.580 1.180 1.040
> boxplot.stats(dataset$citric.acid)$out
[1] 1
> boxplot.stats(dataset$residual.sugar)$out
  [1]  6.10  6.10  3.80  3.90  4.40 10.70  5.50  5.90  5.90  3.80  5.10  4.65  4.65  5.50  5.50  5.50  5.50  7.30  7.20
 [20]  3.80  5.60  4.00  4.00  4.00  4.00  7.00  4.00  4.00  6.40  5.60  5.60 11.00 11.00  4.50  4.80  5.80  5.80  3.80
 [39]  4.40  6.20  4.20  7.90  7.90  3.70  4.50  6.70  6.60  3.70  5.20 15.50  4.10  8.30  6.55  6.55  4.60  6.10  4.30
 [58]  5.80  5.15  6.30  4.20  4.20  4.60  4.20  4.60  4.30  4.30  7.90  4.60  5.10  5.60  5.60  6.00  8.60  7.50  4.40
 [77]  4.25  6.00  3.90  4.20  4.00  4.00  4.00  6.60  6.00  6.00  3.80  9.00  4.60  8.80  8.80  5.00  3.80  4.10  5.90
 [96]  4.10  6.20  8.90  4.00  3.90  4.00  8.10  8.10  6.40  6.40  8.30  8.30  4.70  5.50  5.50  4.30  5.50  3.70  6.20
[115]  5.60  7.80  4.60  5.80  4.10 12.90  4.30 13.40  4.80  6.30  4.50  4.50  4.30  4.30  3.90  3.80  5.40  3.80  6.10
[134]  3.90  5.10  5.10  3.90 15.40 15.40  4.80  5.20  5.20  3.75 13.80 13.80  5.70  4.30  4.10  4.10  4.40  3.70  6.70
[153] 13.90  5.10  7.80
```

```
> boxplot.stats(dataset$chlorides)$out
  [1] 0.176 0.170 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.610 0.360 0.270 0.039 0.337 0.263
 [20] 0.611 0.358 0.343 0.186 0.213 0.214 0.121 0.122 0.122 0.128 0.120 0.159 0.124 0.122 0.122 0.174 0.121 0.127 0.413
 [39] 0.152 0.152 0.125 0.122 0.200 0.171 0.226 0.226 0.250 0.148 0.122 0.124 0.124 0.143 0.222 0.039 0.157 0.422 0.034
 [58] 0.387 0.415 0.157 0.157 0.243 0.241 0.190 0.132 0.126 0.038 0.165 0.145 0.147 0.012 0.012 0.039 0.194 0.132 0.161
 [77] 0.120 0.120 0.123 0.123 0.414 0.216 0.171 0.178 0.369 0.166 0.166 0.136 0.132 0.132 0.123 0.123 0.123 0.403 0.137
 [96] 0.414 0.166 0.168 0.415 0.153 0.415 0.267 0.123 0.214 0.214 0.169 0.205 0.205 0.039 0.235 0.230 0.038
> boxplot.stats(dataset$free.sulfur.dioxide)$out
 [1] 52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52 55 55 48 48 66
> boxplot.stats(dataset$total.sulfur.dioxide)$out
 [1] 145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165 124 124 134 124 129 151 133
[30] 142 149 147 145 148 155 151 152 125 127 139 143 144 130 278 289 135 160 141 141 133 147 147 131 131 131
> boxplot.stats(dataset$density)$out
 [1] 0.99160 0.99160 1.00140 1.00150 1.00150 1.00180 0.99120 1.00220 1.00220 1.00140 1.00140 1.00140 1.00140 1.00320
[15] 1.00260 1.00140 1.00315 1.00315 1.00315 1.00210 1.00210 0.99170 0.99220 1.00260 0.99210 0.99154 0.99064 0.99064
[29] 1.00289 0.99162 0.99007 0.99007 0.99020 0.99220 0.99150 0.99157 0.99080 0.99084 0.99191 1.00369 1.00369 1.00242
[43] 0.99182 1.00242 0.99182
> boxplot.stats(dataset$pH)$out
 [1] 3.90 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89 2.89 2.92 3.90 3.71 3.69 3.69 3.71 3.71
[24] 2.89 2.89 3.78 3.70 3.78 4.01 2.90 4.01 3.71 2.88 3.72 3.72
```

```
> boxplot.stats(dataset$sulphates)$out
 [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36
[24] 1.18 1.13 1.04 1.11 1.13 1.07 1.06 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18 1.07
[47] 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
> boxplot.stats(dataset$alcohol)$out
 [1] 14.00000 14.00000 14.00000 14.00000 14.90000 14.00000 13.60000 13.60000 13.60000 14.00000 14.00000 13.56667 13.60000
> boxplot.stats(dataset$quality)$out
 [1] 8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8
```

Como se puede observar la mayoría de los tributos posee varios valores extremos. Sin embargo, debido a la naturaleza de los datos, estos se considerarán como valores legítimos, por lo que no se modificarán y serán contemplados en el análisis.

Análisis de datos

Comprobación de la normalidad

Con el objetivo de verificar la normalidad de los datos, ahora se procede a utilizar el test de Shapiero-Wilk considerado uno de los mejores para este cometido, esta prueba nos indica que si el p-valor es menor al nivel de significancia que por lo general corresponde α = 0.05, se concluye que los datos no cuentan con distribución normal, caso contrario contienen una distribución normal. Desde R lo podemos implementar de la siguiente manera para cada una de los atributos:

```
> #Comprobacion de la normalidad
> shapiro.test(dataset$fixed.acidity)

        Shapiro-Wilk normality test

data:  dataset$fixed.acidity
W = 0.94203, p-value < 2.2e-16

> shapiro.test(dataset$volatile.acidity)

        Shapiro-Wilk normality test

data:  dataset$volatile.acidity
W = 0.97434, p-value = 2.693e-16

> shapiro.test(dataset$citric.acid)

        Shapiro-Wilk normality test

data:  dataset$citric.acid
W = 0.95529, p-value < 2.2e-16

> shapiro.test(dataset$residual.sugar)

        Shapiro-Wilk normality test

data:  dataset$residual.sugar
W = 0.56608, p-value < 2.2e-16

> shapiro.test(dataset$chlorides)

        Shapiro-Wilk normality test

data:  dataset$chlorides
W = 0.48425, p-value < 2.2e-16
```

```
> shapiro.test(dataset$free.sulfur.dioxide)

        Shapiro-Wilk normality test

data:  dataset$free.sulfur.dioxide
W = 0.90184, p-value < 2.2e-16

> shapiro.test(dataset$total.sulfur.dioxide)

        Shapiro-Wilk normality test

data:  dataset$total.sulfur.dioxide
W = 0.87322, p-value < 2.2e-16

> shapiro.test(dataset$density)

        Shapiro-Wilk normality test

data:  dataset$density
W = 0.99087, p-value = 1.936e-08

> shapiro.test(dataset$pH)

        Shapiro-Wilk normality test

data:  dataset$pH
W = 0.99349, p-value = 1.712e-06

> shapiro.test(dataset$sulphates)

        Shapiro-Wilk normality test

data:  dataset$sulphates
W = 0.83304, p-value < 2.2e-16
```

```
> shapiro.test(dataset$alcohol)

        Shapiro-Wilk normality test

data:  dataset$alcohol
W = 0.92884, p-value < 2.2e-16

> shapiro.test(dataset$quality)

        Shapiro-Wilk normality test

data:  dataset$quality
W = 0.85759, p-value < 2.2e-16
```

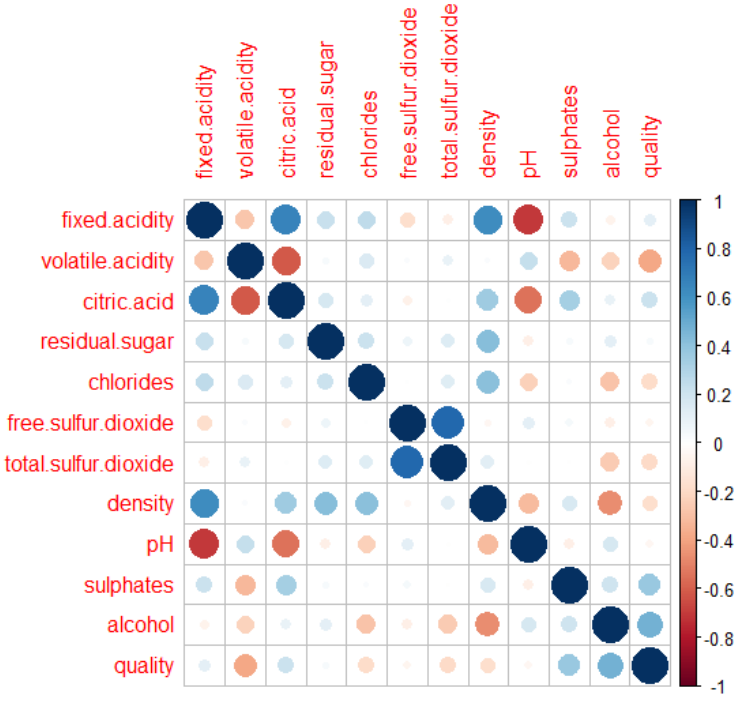
Como se puede observar según los resultados del test para cada atributo, ninguno sigue una distribución normal. Se podría normalizar estos datos, pero para este caso práctico se la conservara tal como están.

Aplicación de pruebas estadísticas

A continuación, se realizan diferentes análisis de los datos, en primera instancia se pretende conocer la correlación entre las distintas variables, con el objetivo de conocer su influencia sobre la calidad del vino. Los datos no contienen una distribución normal por lo tanto se hará uso de la correlación de Spearman considerada una alternativa no paramétrica para este caso.

```
> #Correlacion de variables
> cor(dataset, method = "spearman")
                      fixed.acidity volatile.acidity  citric.acid residual.sugar     chlorides free.sulfur.dioxide
fixed.acidity           1.00000000      -0.27828222  0.661708417     0.22070086  0.2509041064       -0.1751365613
volatile.acidity       -0.27828222       1.00000000 -0.610259467     0.03238560  0.1587702548        0.0211626414
citric.acid             0.66170842      -0.61025947  1.000000000     0.17641731  0.1125765077       -0.0764515753
residual.sugar          0.22070086       0.03238560  0.176417306     1.00000000  0.2129592419        0.0746178640
chlorides               0.25090411       0.15877025  0.112576508     0.21295924  1.0000000000        0.0008051686
free.sulfur.dioxide    -0.17513656       0.02116264 -0.076451575     0.07461786  0.0008051686        1.0000000000
total.sulfur.dioxide   -0.08841741       0.09411014  0.009399602     0.14537506  0.1300333418        0.7896978767
density                 0.62307076       0.02501412  0.352285261     0.42226586  0.4113896972       -0.0411776800
pH                     -0.70667359       0.23357152 -0.548026276    -0.08997095 -0.2343612736        0.1156791779
sulphates               0.21265375      -0.32558398  0.331074404     0.03833200  0.0208254792        0.0458623500
alcohol                -0.06657566      -0.22493168  0.096455544     0.11654813 -0.2845039422       -0.0813673063
quality                 0.11408367      -0.38064651  0.213480914     0.03204817 -0.1899223356       -0.0569006455
                      total.sulfur.dioxide     density          pH     sulphates      alcohol      quality
fixed.acidity              -0.0884174083   0.62307076 -0.706673595   0.2126537506 -0.06657566   0.11408367
volatile.acidity            0.0941101376   0.02501412  0.233571519  -0.3255839818 -0.22493168  -0.38064651
citric.acid                 0.0093996024   0.35228526 -0.548026276   0.3310744040  0.09645554   0.21348091
residual.sugar              0.1453750584   0.42226586 -0.089970954   0.0383320002  0.11654813   0.03204817
chlorides                   0.1300333418   0.41138970 -0.234361274   0.0208254792 -0.28450394  -0.18992234
free.sulfur.dioxide         0.7896978767  -0.04117768  0.115679178   0.0458623500 -0.08136731  -0.05690065
total.sulfur.dioxide        1.0000000000   0.12933210 -0.009841438  -0.0005038194 -0.25780603  -0.19673508
density                     0.1293321018   1.00000000 -0.312055078   0.1614782344 -0.46244458  -0.17707407
pH                         -0.0098414382  -0.31205508  1.000000000  -0.0803060380  0.17993243  -0.04367193
sulphates                  -0.0005038194   0.16147823 -0.080306038   1.0000000000  0.20732955   0.37706020
alcohol                    -0.2578060251  -0.46244458  0.179932427   0.2073295535  1.00000000   0.47853169
quality                    -0.1967350754  -0.17707407 -0.043671935   0.3770601991  0.47853169   1.00000000
```

```
> library("corrplot")
> corrplot(cor(dataset, method = "spearman"))
```



Como podemos observar la correlación existente de las variables con la calidad no es muy relevante. El alcohol el más relacionado con la calidad.

Modelo de regresión lineal

En esta práctica se planteó el objetivo de tener la posibilidad de conocer la calidad del vino en base a sus características fisicoquímicas. A continuación, se empleará la regresión lineal para obtener un modelo que permita hacer predicciones sobre la calidad del vino.

Para la generación del modelo se creará dos, el primero (Modelo 1) donde se incluye todas las variables y el segundo (Modelo 2) donde incluya el top de las variables algo correladas con respecto a la calidad y luego seleccionaremos el que tengo mejor coeficiente de determinación (R^2).

(Modelo 1)

```
> #Regresion lineal
> ntrain <- nrow(dataset)*0.8
> ntest <- nrow(dataset)*0.2
> set.seed(1)
> index_train <- sample(1:nrow(dataset), size = ntrain)
> train <- dataset[index_train,]
> test <- dataset[-index_train,]
> model_full <- lm(formula = quality~., data = train)
> summary(model_full)

Call:
lm(formula = quality ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.52213 -0.35627 -0.04738  0.44215  2.00517

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.342e+01  2.323e+01   1.008 0.313441
fixed.acidity         2.725e-02  2.862e-02   0.952 0.341210
volatile.acidity     -1.032e+00  1.353e-01  -7.625 4.77e-14 ***
citric.acid          -1.857e-01  1.626e-01  -1.142 0.253523
residual.sugar        2.223e-02  1.623e-02   1.370 0.170964
chlorides            -1.456e+00  4.849e-01  -3.002 0.002731 **
free.sulfur.dioxide   4.483e-03  2.452e-03   1.828 0.067725 .
total.sulfur.dioxide -3.086e-03  8.191e-04  -3.767 0.000173 ***
density              -1.976e+01  2.372e+01  -0.833 0.404986
pH                   -3.457e-01  2.165e-01  -1.597 0.110515
sulphates             9.192e-01  1.257e-01   7.311 4.69e-13 ***
alcohol               2.834e-01  2.909e-02   9.741  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6474 on 1267 degrees of freedom
Multiple R-squared:  0.3536,    Adjusted R-squared:  0.348
F-statistic: 63.02 on 11 and 1267 DF,  p-value: < 2.2e-16
```

(Modelo 2)

```
> #Modelo parcial (variables mas correladas)
> model_par <- lm(formula = quality~volatile.acidity+citric.acid+sulphates+alcohol, data = train)
> summary(model_par)

Call:
lm(formula = quality ~ volatile.acidity + citric.acid + sulphates +
    alcohol, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41985 -0.37922 -0.06241  0.44713  2.22102

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.49732    0.22382  11.158  < 2e-16 ***
volatile.acidity -1.18117    0.12623  -9.357  < 2e-16 ***
citric.acid      -0.07383    0.11503  -0.642    0.521
sulphates         0.74502    0.11329   6.576 7.03e-11 ***
alcohol           0.31531    0.01737  18.157  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6549 on 1274 degrees of freedom
Multiple R-squared:  0.335,     Adjusted R-squared:  0.3329
F-statistic: 160.4 on 4 and 1274 DF,  p-value: < 2.2e-16
```

Como podemos observar los modelos no son buenos, el coeficiente R^2 es un poco mejor cuando se utiliza todas las variables con un valor de 0.3536. Sin embargo, está lejos de lo óptimo. Ahora procederemos a verificar la predicción del modelo uno:

```
> #verificando prediccion
> pre_calidad <- predict(model_full, test, type="response")
> comparacion <- data.frame(Original = test$quality, Predicho = as.integer(pre_calidad),
+                           Igual = (test$quality-as.integer(pre_calidad)))
> head(comparacion)
  Original Predicho Igual
1        5        5     0
2        5        5     0
3        5        5     0
4        5        5     0
5        5        5     0
6        6        5     1
> resultados <- data.frame(Correctos = sum(comparacion$Igual == 0), Incorrectos = sum(comparacion$Igual != 0) )
> resultados
  Correctos Incorrectos
1       152         168
```

Se puede observar que un poco más de la mitad no se predijo de manera correcta, esto era de esperar debido a que el modelo no es bueno.

Resolución del problema

Conclusiones

En primera instancia de verifico la normalidad de los datos para ello se utilizó el test de Shapiero-Wilk cuyos resultados indicaron que ninguna de las variables presenta una distribución normal. Luego se procedió a analizar la correlación a través de Spearman de las variables para conocer su influencia con la calidad del vino, teniendo como resultado una deficiente correlación de las variables con el vino.

Por último, se utiliza el método de regresión lineal donde se obtuvo dos modelos el primero consta del uso de todas las variables y el segundo con las variables más cercanas en correlación. Ninguno de los modelos obtuvo un R^2 considerable. Sin

embargo, el primer modelo obtuvo un valor de coeficiente mayor que el segundo, con este modelo se realizó un proceso de predicción de datos, para verificar la efectividad cuyos resultados no fueron satisfactorios.

Código

El código de esta práctica se encuentra disponible en el siguiente enlace: https://github.com/devleo0595/limpieza_analisis.git