Parcial 1: Modelos lienales generales y no paramétricos

Cesar A. Saavedra Vanegas

10/23/2020

Parcial 1

Este conjunto de datos de vino tinto consta de 1599 observaciones y 12 variables, 11 de las cuales son sustancias químicas. Las variables son:

1. **Acidez fija:** La mayoría de los ácidos implicados en el vino son fijos o no volátiles (no se evaporan fácilmente).
2. **Acidez volátil:** La cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.
3. **Ácido cítrico:** Encontrado en pequeñas cantidades, el ácido cítrico puede agregar "frescura" y sabor a los vinos.
4. **Azúcar residual:** Es la cantidad de azúcar que queda después de que se detiene la fermentación, es raro encontrar vinos con menos de 1 gramo / litro y los vinos con más de 45 gramos / litro se consideran dulces.
5. **Cloruros:** Es la cantidad de sal del vino.
6. **Dióxido de azufre libre:** La forma libre de SO_2 existe en equilibrio entre el SO_2 molecular (como gas disuelto) y el ion bisulfito; Previene el crecimiento microbiano y la oxidación del vino.
7. **Dióxido de azufre total:** Es la cantidad de formas libres y unidas de SO_2; en concentraciones bajas, el SO_2 es mayormente indetectable en el vino, pero en concentraciones de SO_2 libre superiores a 50 ppm, el SO_2 se hace evidente en la nariz y el sabor del vino.
8. **Densidad:** La densidad es cercana a la del agua dependiendo del porcentaje de alcohol y contenido de azúcar.
9. **pH:** Describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH.
10. **Sulfatos:** Aditivo del vino que puede contribuir a los niveles de dióxido de azufre (SO_2), que actúa como antimicrobiano y antioxidante.
11. **Alcohol:** El porcentaje de contenido de alcohol del vino.
12. **Calidad:** Variable de respuesta (basada en datos sensoriales, puntuación entre 0 y 10).

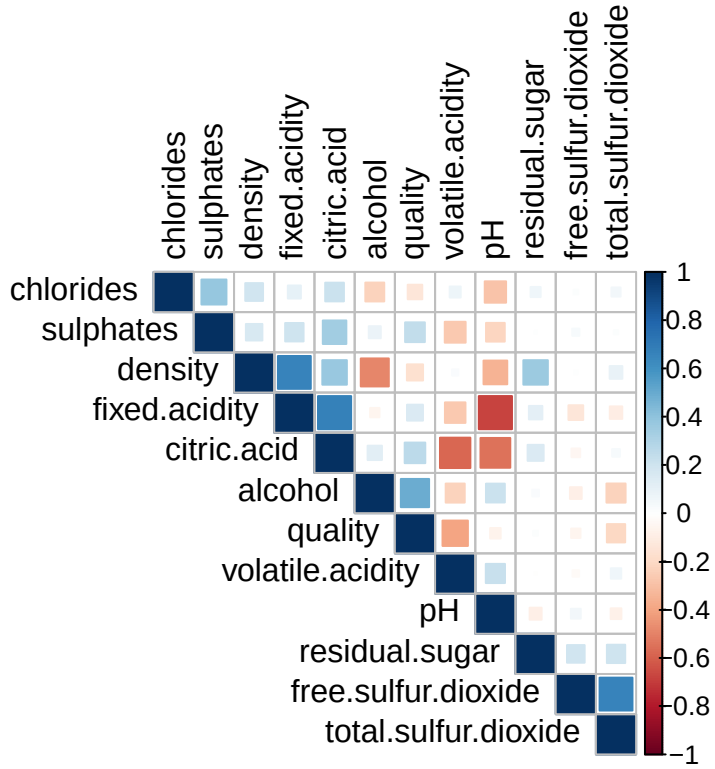
Base de datos vinos

```
Datos <- read.table("Datos.txt",header=T,sep = ",")
```

Se selecciona un tamaño de muestra de 1200 vinos para realizar el modelo, tal y como se muestra a continuación:

```
#Tamaño de la muestra
n <- 1200
```

```r
corrplot(cor(muestra), method="square", type="upper", order="hclust", tl.col="black")
```



Variables indicadoras

Se convierten las variables "alcohol" en una variable indicadora con dos niveles: "Bajo" y "Alto" y la variable "quality" como dicotómica con sus niveles en "0" y "1", estas nuevas variables se denominan "alcoholAB" y "calidadAB" respectivamente. Se realiza el siguiente procedimiento:

```r
# Variable indicadora alcoholAB
alcoholAB <- vector()
alcoholAB[muestra$alcohol < 12] <- "Bajo"
alcoholAB[muestra$alcohol >= 12] <- "Alto"
alcoholAB <- as.factor(alcoholAB)
table(alcoholAB)
```

```
## alcoholAB
## Alto Bajo
##  127 1073
```

```r
# Variable indicadora quality
calidadAB <- vector()
calidadAB[muestra$quality <= 6] <- "0"
calidadAB[muestra$quality > 6] <- "1"
calidadAB <- as.factor(calidadAB)
table(calidadAB)
```

```
## calidadAB
##    0    1
## 1037  163
```

Modelo con variable indicadora y dicotómica

En esta sección, se procede a generar un modelo logístico con variable de respuesta ordinal (CalidadAB), ya que la variable de respuesta "calidadAB" ha sido dicotomisada, esto es, una puntuación entre 0 y 1, donde 0 representa una mala calidad y 1 una buena calidad del vino.

```
modelo.logit <- glm(calidadAB ~ fixed.acidity + alcoholAB, data = muestra, family = "binomial")
```

```
summary(modelo.logit)
```

```
##
## Call:
## glm(formula = calidadAB ~ fixed.acidity + alcoholAB, family = "binomial",
##     data = muestra)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8050  -0.4828  -0.4000  -0.3569   2.4226
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.21874    0.42027  -5.279 1.30e-07 ***
## fixed.acidity  0.26300    0.04791   5.489 4.04e-08 ***
## alcoholABBajo -2.29167    0.21797 -10.513  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 953.58  on 1199  degrees of freedom
## Residual deviance: 828.67  on 1197  degrees of freedom
## AIC: 834.67
##
## Number of Fisher Scoring iterations: 5
```