UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

ANÁLISIS DE DATOS

LABORATORIO 1:

MODELANDO PREFERENCIAS DE VINO DESDE CARACTERÍSTICAS FISICOQUÍMICAS

Angelo Carlier

Juan Giglio

Profesor: Felipe Andrés Bello

Santiago – Chile

2019

TABLA DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

1.1 INTRODUCCIÓN

Actualmente el vino es disfrutado por un amplio rango de consumidores. Portugal está dentro de los diez países exportadores con una participación del 3,17% del mercado en 2005 FAO (2015). Las exportaciones de la variedad *vinho verde* se ha incrementado en un 36% desde 1997 a 2007. La industria del vino para sostener este crecimiento está invirtiendo en nuevas tecnologías para mejorar sus procesos de marketing y producción. Dentro de este contexto la certificación y evaluación de calidad son claves. La certificación previene que los vinos sean nocivos para la salud y asegura su calidad en el mercado. La evaluación de calidad la mayoría de las veces es parte de este proceso de certificación y puede ser utilizada para mejorar el proceso de elaboración del vino y para identificar su nicho dentro del mercado. La certificación generalmente consta de dos etapas una fisicoquímica y otra sensorial. En la etapa fisicoquímica una muestra del vino es entregada a un laboratorio y este determina su densidad, grado alcohólico y pH entre otros, mientras que la evaluación sensorial es llevada a cabo por catadores expertos. La relación entre la evaluación fisicoquímica y la sensorial es compleja. Cortez et al. propone utilizar minería de datos para predecir la preferencia de sabor de un vino basado en las evaluaciones fisicoquímicas que se obtienen durante su certificación. Se consideran dos muestras de textitvinho verde una de vino blanco y otra de vino rojo. Para efectos de este laboratorio utilizaremos el mismo conjunto de datos que Cortez et al. lo describiremos con técnicas estadísticas y buscaremos las variables de la evaluación fisicoquímica que mejor caracterizan la evaluación sensorial.

CAPÍTULO 2. DESCRIPCIÓN DEL PROBLEMA

Considerando que la relación entre el análisis fisicoquímico y la evaluación sensorial de vinos es compleja y desconocida, podemos plantear el problema de la siguiente manera: ¿Cuáles son las características fisicoquímicas que mejor representan la evaluación sensorial de vinos?

2.1 DESCRIPCIÓN DE LA BASE DE DATOS

Cortez et al. considera en su estudio el vino *vinho verde* de Minho del noreste de Portugal. Analiza sus dos variedades más comunes blanco y rojo. Los datos se recolectaron entre Mayo 2004 y Febrero 2007, solo muestras con denominación de origen fueron analizadas por la entidad oficial de certificación (CVRVV). CVRVV es una organización que tiene como meta mejorar la calidad y marketing del *vinho verde*. Por solicitud de los productores evalúa fisicoquímicamente y sensorialmente muestras de vino. Cada registro considera estas dos mediciones. Se seleccionaron las evaluaciones fisicoquímicas más comunes para evitar descartar muestras. Dado que el vino rojo y el blanco tienen perfiles de sabor diferentes se separaron en dos conjuntos de datos con 1599 y 4898 muestras respectivamente. Todas las muestras fueron evaluadas por al menos tres asesores considerando pruebas a ciegas, los cuales evaluaron los vinos en una escala de 0 (muy malo) a 10 (excelente). El puntaje final corresponde a la media de estas evaluaciones.

2.2 DESCRIPCIÓN DE CLASES Y VARIABLES

Variables fisicoquímicas:

1. Acidez fija: gramos de ácido tartárico por decímetro cúbico (g/dm^3)
2. Acidez volátil: gramos de ácido acético por decímetro cúbico (g/dm^3)
3. Ácido cítrico: gramos de ácido cítrico por decímetro cúbico (g/dm^3)
4. Azúcar residual: gramos de azúcar por decímetro cúbico (g/dm^3)
5. Cloruros: gramos de cloruro de sodio por decímetro cúbico (g/dm^3)
6. Dióxido de azufre libre: miligramos por decímetro cúbico (mg/dm^3)

7. Dióxido de azufre total: miligramos por decímetro cúbico (mg/dm^3)
8. Densidad: gramo por centímetro cúbico (g/cm^3)
9. pH: métrica de acidez/alcalinidad sin unidad de medida
10. Sulfatos: gramos de sulfato de potasio por decímetro cúbico (g/dm^3)
11. Alcohol: grados de alcohol

Variables sensoriales:

1. Calidad: Evaluación de percepción humana de la calidad del vino en escala de 1 a 10

2.3 ANÁLISIS ESTADÍSTICO:

Para obtener descripción completa del set de datos, desde el punto de vista de estadística descriptiva, utilizamos la función *summary*, nativa de R y obtenemos la siguiente información:

Estadística descriptiva para dataset red wine, figura 2.1:

```
> summary(df_redwine)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
 total.sulfur.dioxide    density             pH           sulphates         alcohol         quality
 Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Figura 2.1: Dataset red wine.
Fuente: Elaboración propia, 2019.

Estadística descriptiva para dataset white wine, figura 2.2:

```
> summary(df_whitewine)
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200   Median :0.04300   Median : 34.00
 Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
 3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
 Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800   Max.   :0.34600   Max.   :289.00
 total.sulfur.dioxide    density             pH           sulphates         alcohol         quality
 Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
 Median :134.0        Median :0.9937   Median :3.180   Median :0.4700   Median :10.40   Median :6.000
 Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
 3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
 Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Figura 2.2: Dataset white wine.
Fuente: Elaboración propia, 2019.

En dichas descripciones podemos encontrar:

1. Min.: Valor mínimo asignado a la variable
2. 1st Qu.: primer percentil
3. Median: Mediana de la variable
4. Mean: Media de la variable
5. 3rd Qu.: Tercer percentil
6. Max.: Valor máximo asignado a la variable

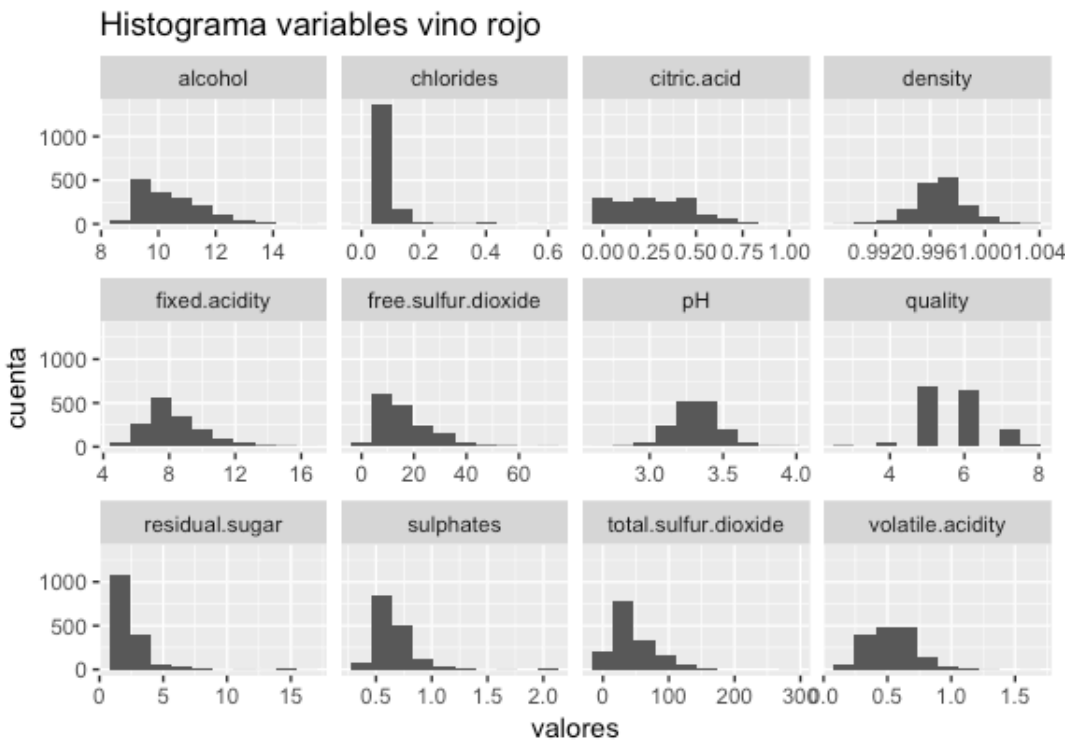

Figura 2.3: Histogramas vino rojo.
Fuente: Elaboración propia, 2019.

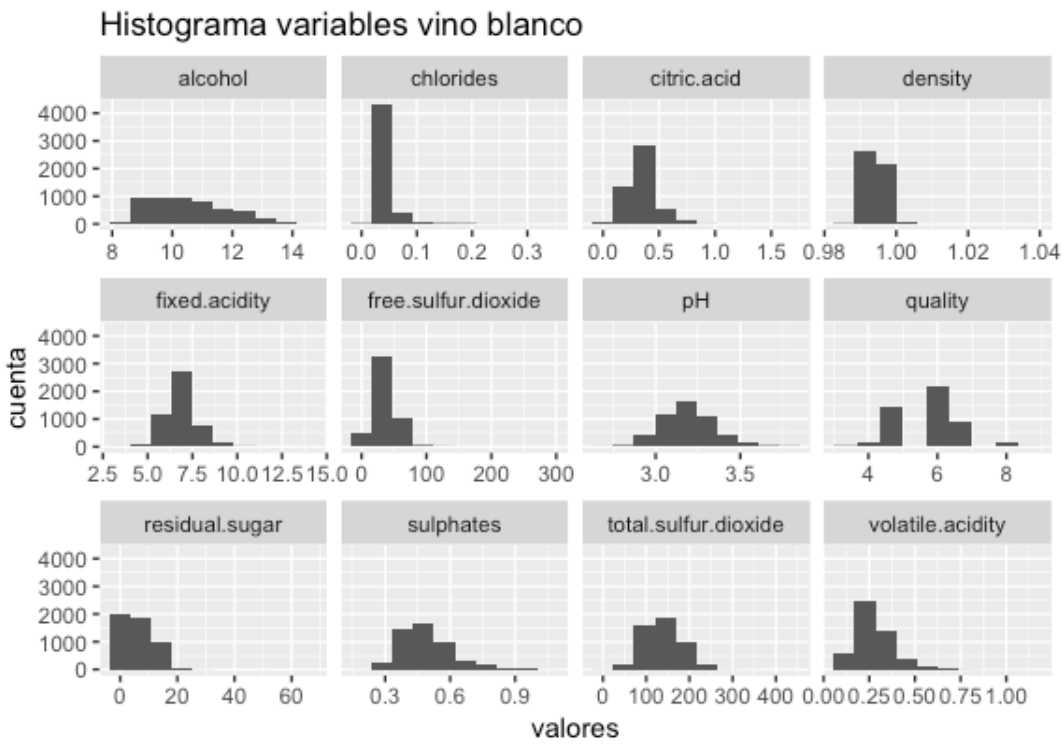

Figura 2.4: Histogramas vino blanco.
Fuente: Elaboración propia, 2019.

De las matrices de correlación podemos observar que:

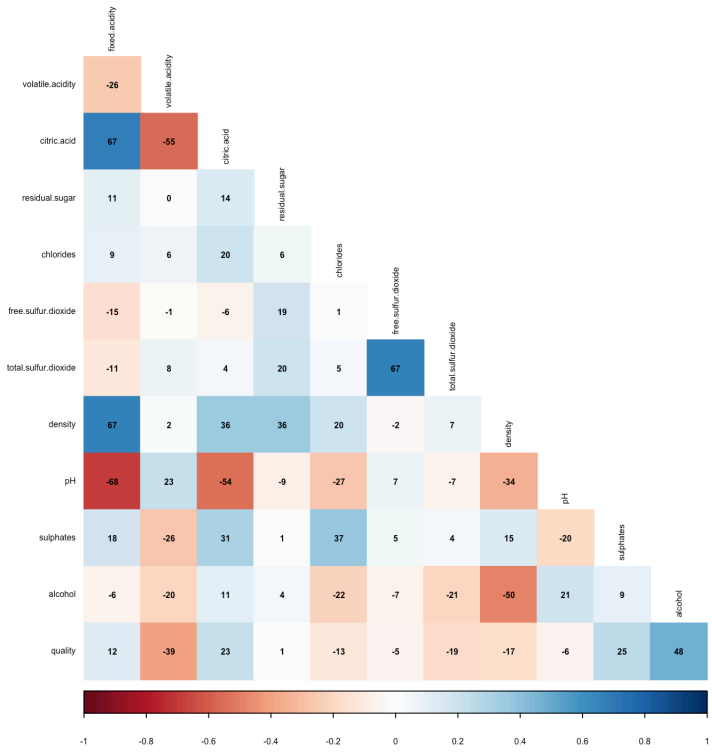

Figura 2.5: Matriz de correlación vino rojo.
Fuente: Elaboración propia, 2019.

Para el vino rojo figura 2.5 podemos observar que la calidad está correlaciona positivamente con: alcohol, pH y sulfatos y negativamente con: densidad, cloruros, dióxido de azufre total, acidez volátil y azúcar residual.

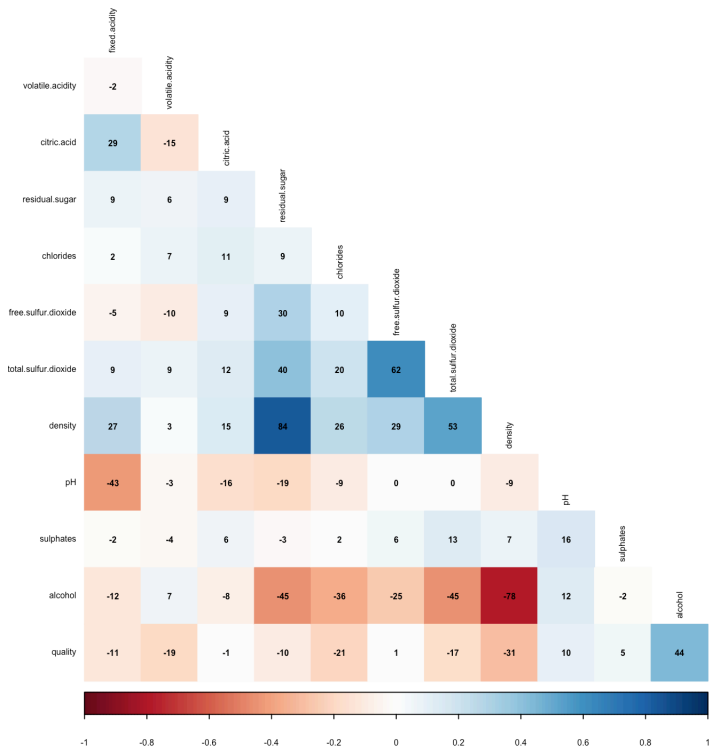

Figura 2.6: Matriz de correlación vino blanco.
Fuente: Elaboración propia, 2019.

Para el vino blanco figura 2.6 podemos observar que la calidad está correlaciona positivamente con: alcohol, sulfatos, ácido cítrico y acidez fija y negativamente con: acidez volátil, dióxido de azufre total, densidad y cloruros.

Adicional a los análisis ya realizados, utilizaremos Random Forest (RF) sobre el conjunto de datos. Este análisis nos permite obtener la importancia de cada variable sobre la clase, en esta ocasión la calificación otorgada por los expertos, utilizando el mean decreasing gini. En su implementación obtenemos los siguientes resultados (2.7, 2.8):

```
> importance(fit_rf)
                      IncNodePurity
fixed.acidity             60.90116
volatile.acidity         129.12208
citric.acid               71.17692
residual.sugar            55.40250
chlorides                 67.39943
free.sulfur.dioxide       49.47494
total.sulfur.dioxide      81.50977
density                   83.76102
pH                        57.09789
sulphates                134.87932
alcohol                  197.42732
```

Figura 2.7: RF sobre datos de vino rojo.
Fuente: Elaboración propia, 2019.

```
> importance(fit_rf)
                      IncNodePurity
fixed.acidity             232.9765
volatile.acidity          395.3314
citric.acid               254.3736
residual.sugar            281.6550
chlorides                 306.6896
free.sulfur.dioxide       387.7417
total.sulfur.dioxide      297.1475
density                   425.4647
pH                        257.9789
sulphates                 221.0017
alcohol                   597.6857
```

Figura 2.8: RF sobre datos de vino blanco.
Fuente: Elaboración propia, 2019.

Estos resultados se condicen con el análisis de correlación explicado previamente. Y se interpretan utilizando los valores más altos de ”IncNodePurity” como variables que poseen mayor importancia.

Escalamiento multidimensional o MDS por sus siglas en inglés *multidimensional scaling* es un conjunto de métodos para representar los datos en un mapa de acuerdo a una distancia, para este laboratorio consideramos la distancia correlacional. En este caso lo aplicamos como una técnica de análisis exploratorio para observar los datos sin conocer la estructura de estos a priori.

En el caso del vino rojo podemos ver en la figura 2.9 que los vinos con puntaje 7 y 8 en el análisis sensorial se encuentran dispersos alrededor de la mayoría de los vinos sin formar una agrupación. Podemos inferir que: los datos del análisis sensorial no necesariamente están relacionados con todas las variables el análisis fisicoquímico o que los expertos no tienen un perfil de sabor en común, prefieren la variedad. En el caso del vino blanco en la figura 2.10 no es posible observar un patrón dado que los datos se encuentran aglomerados.

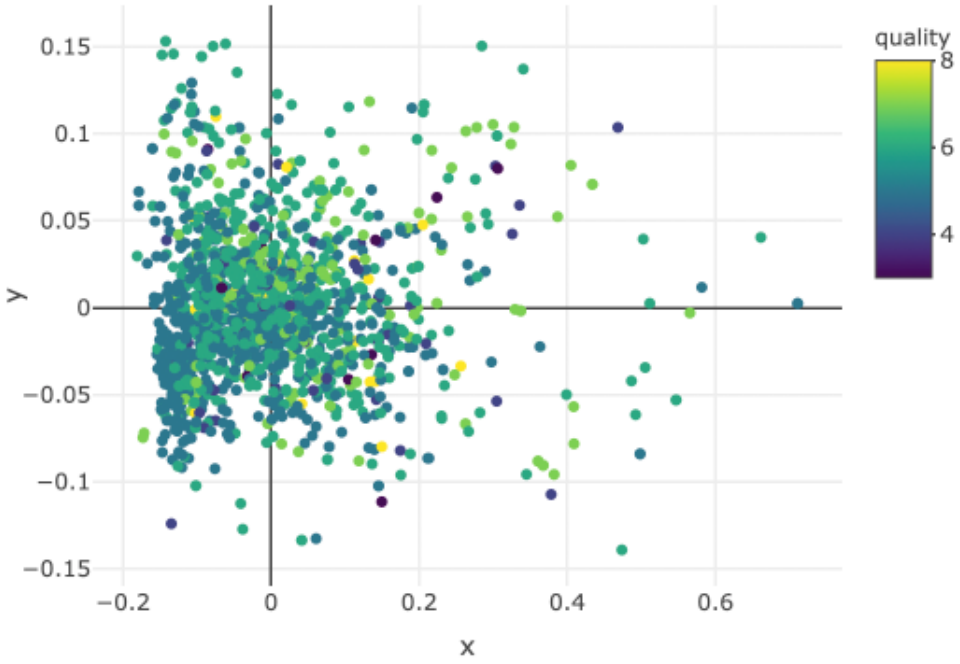

Figura 2.9: MDS vino rojo.
Fuente: Elaboración propia, 2019.

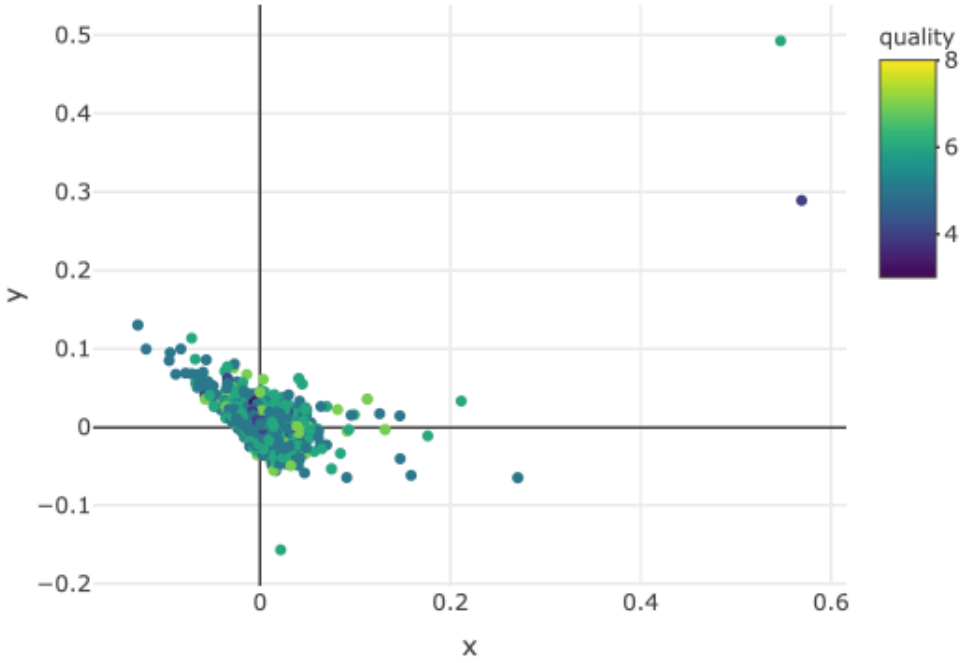

Figura 2.10: MDS vino blanco.
Fuente: Elaboración propia, 2019.

CAPÍTULO 3. CONCLUSIONES

3.1 ANÁLISIS ESTADÍSTICO Y LA RESOLUCIÓN DEL PROBLEMA

El problema abordado en la investigación consiste, en primera instancia, en obtener el conjunto de variables dentro del perfil fisicoquímico, más influyentes en la calificación entregada por lo expertos para cada uno de los vinos en evaluación. Nuestro estudio abordó este problema mediante análisis de correlación entre variables y la utilización de Random Forest. Estos análisis nos permiten obtener las siguientes conclusiones respecto a las variables del dataset:

1. La variable más influyente directamente proporcional en la calificación es ”alcohol”
2. Las variables más influyentes de forma inversamente proporcional en la calificación son ”Density” y ”volatile.acidity”
3. El uso de técnicas de selección de características facilitan la clasificación en los conjuntos de datos, disminuyendo el ruido que genera la consideración del total de variables y eliminando aquellas que aportan poca información (o nula) a la clasificación
4. Desde el punto de vista computacional, la reducción de variables impacta directamente en los tiempos requeridos para alcanzar una solución utilizando técnicas de minería de datos, dicho esto, la reducción de variables disminuye los tiempos de procesamiento

Finalmente, nuestro conjunto de variables junto a su importancia se detalla en 3.1 y 3.2

```
alcohol              0.19572829
sulphates            0.13780597
volatile.acidity     0.13185772
density              0.08730588
total.sulfur.dioxide 0.08292060
chlorides            0.07019412
citric.acid          0.06970012
fixed.acidity        0.06188827
pH                   0.05823370
residual.sugar       0.05532529
free.sulfur.dioxide  0.04904003
```

Figura 3.1: Importancia de variables fisicoquimicas red wine.
Fuente: Fabricación propia, 2019

alcohol	0.16822904
density	0.11522838
volatile.acidity	0.10762329
free.sulfur.dioxide	0.10496062
chlorides	0.08168971
total.sulfur.dioxide	0.08077873
residual.sugar	0.07688681
pH	0.06996698
citric.acid	0.06971729
fixed.acidity	0.06367100
sulphates	0.06124814

Figura 3.2: Importancia de variables fisicoquimicas white Wine.
Fuente: Fabricación propia, 2019

3.2 COMPARACIÓN CON LA LITERATURA

Comparando nuestros resultados con los de Cortez et al. podemos observar que no llegamos al mismo resultado, en orden de importancia obtuvimos para el vino rojo que las variables en figura 3.1 y para el vino blanco en figura 3.2, las variables consideradas por Cortez et al. están representadas en la figura 3.3. En este laboratorio no se considera la creación del modelo predictivo por lo que no es posible evaluar su precisión utilizando la configuración de variables encontradas. Dado esto no podemos compararnos con el estudio de Cortez et al.

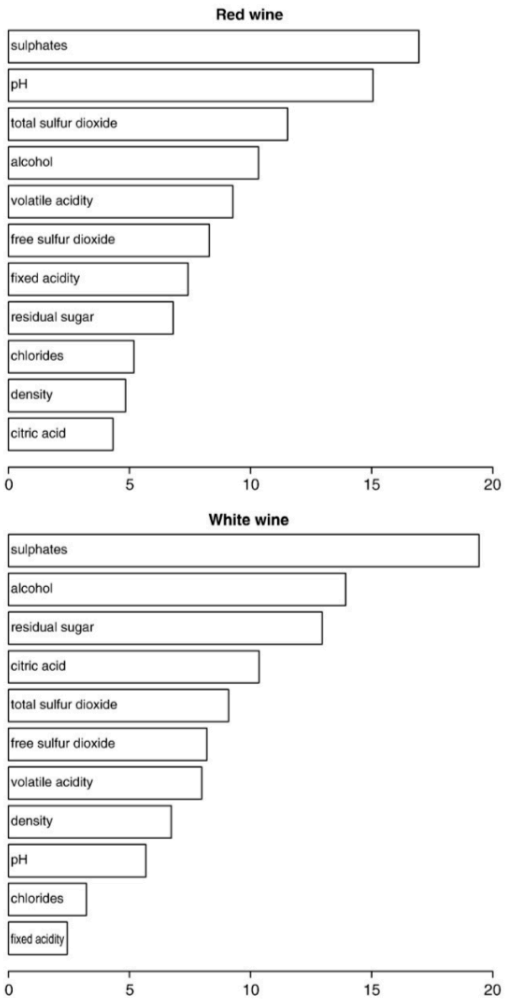

Figura 3.3: Importancia de variables fisicoquimicas modelo SVM en porcentaje (%).
Fuente: Cortez et al. (2009)

REFERENCIAS BIBLIOGRÁFICAS

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. URL https://doi.org/10.1016/j.dss.2009.05.016

FAO (2015). Crops and livestock products. URL http://www.fao.org/faostat/en/#data/TP

ANEXO A. CÓDIGO FUENTE EXPLICADO

```r
#librerias
library(plotly)
library(ggbiplot)
library(amap)
library(tidyr)
library(ggplot2)
library(mvnormtest)
library(corrplot)
library(randomForest)

#Importación de archivos
wineRed<-read.csv('./datasets/winequality-red.csv',sep=';')
wineWhite<-read.csv('./datasets/winequality-white.csv',sep=';')

#Descripcion de dataset
summary(wineRed)
summary(wineWhite)

#Análisis MDS
dataMDSRedWine<-as.data.frame(cbind(cmdscale(Dist(wineRed[,-12],method='
    correlation')),wineRed[,12],1:nrow(wineRed)))

names(dataMDSRedWine)<-c('x','y','quality','id')

p <- plot_ly(data = dataMDSRedWine, x = ~x, y = ~y,color= ~quality ,type='
    scatter',mode='markers')
p

dataMDSWhiteWine<-as.data.frame(cbind(cmdscale(Dist(wineWhite[,-12],method='
    correlation')),wineRed[,12],1:nrow(wineRed)))

names(dataMDSWhiteWine)<-c('x','y','quality','id')
p <- plot_ly(data = dataMDSWhiteWine, x = ~x, y = ~y,color= ~quality ,type='
    scatter',mode='markers')
p

#Matrices de correlación
corrplot(cor(wineRed),method = 'color',type='lower',addCoefasPercent = TRUE,
    diag=FALSE,tl.cex=0.5,tl.col='black',addCoef.col = 'black',cl.cex=0.5,
    number.cex = 0.5)
corrplot(cor(wineWhite),method = 'color',type='lower',addCoefasPercent = TRUE,
    diag=FALSE,tl.cex=0.5,tl.col='black',addCoef.col = 'black',cl.cex=0.5,
    number.cex = 0.5)

#Histogramas
ggplot(gather(wineRed), aes(value)) +
      geom_histogram(bins = 10) +
       facet_wrap(~key, scales = 'free_x')+
      ggtitle("Histograma variables vino rojo")+
  xlab("valores")+
  ylab("cuenta")

ggplot(gather(wineWhite), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = 'free_x')+
  ggtitle("Histograma variables vino blanco")+
  xlab("valores")+
```

```
  ylab("cuenta")

# Analisis RF
fit_rf = randomForest(quality~., data=wineRed)
# Importance based on mean decreasing gini
importance(fit_rf)

fit_rf = randomForest(quality~., data=wineWhite)
# Importance based on mean decreasing gini
importance(fit_rf)
```