UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

Análisis de datos

Laboratorio 1: Análisis estadístico e inferencial

Richard Torti

Profesor: Max Chacón

Ayudante: Ignacio Ibañez

Santiago – Chile

2018

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

Una de las bebidas alcohólicas más comunes a nivel internaciones es el vino, por lo tanto sería interesante ver las variables que actúan en este para ver las diferencias y elementos comunes que tienen entre los vinos analizados. Es por esto, que el presente informe hace un estudio general aplicando análisis estadístico e inferencial para determinar correlaciones entre variables, gráficos, etc; con el fin de descubrir información que pueda ayudar a tomar futuras decisiones al respecto de la elaboración de éstos.

En este laboratorio se abordan temas con respecto a análisis estadístico, describiendo la base de dato entregada como también sus atributos para posteriormente ver los datos que tienen éstas. Para los datos se sacaron estadísticas básicas como lo son quintiles, media, mediana, mínimo y máximo, variación y coeficiente de variación. Luego, se hace un análisis más profundo viendo la matriz de correlación para ver la relación entre las variables, para proseguir con un histograma a cada una de ellas. También, se hacen test de Shapiro-Wilk y gráficos de frecuencia.

CAPÍTULO 2. DESCRIPCIÓN DEL PROBLEMA

2.1 DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos fue elaborada por Riccardo Leardi y son una colección de resultados de un análisis químico de vinos crecidos en la misma región de Italia pero enviados desde tres cultivos diferentes. El análisis determinó cuantitativamente trece de los componentes que trae el vino y que fueron encontrado en cada uno de los tres tipos de vino (de los tres cultivos diferentes) Leardi (1991).

Las características de la base de datos se muestran a continuación:

- **Características del conjunto de variables:** Multivariable
- **Número de instancias:** 178.
- **Área:** Física
- **Características de los atributos:** Enteros y reales.
- **Número de atributos:** 13.
- **Fecha:** 01-07-1991.
- **Tareas asociadas:** Clasificación.
- **Datos faltantes:** Ninguno.

También, los archivos incluidos en la base de datos son:

- **wine.data:** Contiene todos los datos recolectados en el proceso del análisis.
- **wine.names:** Contiene la descripción y las variables que fueron analizadas en la recolección de datos

2.2 DESCRIPCIÓN DE CLASES Y VARIABLES

La base de datos se compone de trece atributos:

- **Alcohol:** Grados de alcohol que posee el vino
- **Malic acid:** El ácido málico se identifica por un peculiar olor en el vino, que recuerda al olor de las manzanas verdes. Este atributo muestra la cantidad de ácido málico en gramos por litro de vino.
- **Ash:** Cantidad de cenizas en gramo por cada litro de vino.
- **Alcalinity of ash:** La ceniza tiene un PH alcalino, por lo cual se data su valor pero no en PH, si no que la medición se expresa en gramos por litro de carbonato de potasio.
- **Magnesium:** Cantidad de magnesio presente en el vino.
- **Total phenols:** El fenol o ácido fénico le entrega cuerpo al vino, midiéndose en gramos por litro de vino.
- **Flavanoids:** Es un compuesto de la uva que le da el color, pudiendo ser típicamente de rojo-púrpura o amarillo. Esta medido en gramos por litro de vino.
- **Nonflavanoid phenols:** Compuesto del vino, medido en gramos por litro de vino.
- **Proanthocyanidins:** Compuesto de la uva y por tanto también presente en el vino. Su medición es de gramos por litro de vino.
- **Color intensity:** Intensidad de color medido según la distorsión de color de un objeto al poner el vino frente a éste.
- **Hue:** Mide el tipo de color del vino.
- **OD280/OD315 of diluted wines**
- **Proline:** La prolina es un aminoácido presente en el vino, medido en miligramos por litro de vino.

Así mismo, la recolección de los datos como se mencionó anteriormente, proviene de tres cultivos diferentes, por lo cual hay tres tipos de instancias:

- **Class 1:** Vinos analizados provenientes del cultivo numero uno.
- **Class 2:** Vinos analizados provenientes del cultivo numero dos.
- **Class 3:** Vinos analizados provenientes del cultivo numero tres.

Cabe destacar que todas las variables son continuas.

CAPÍTULO 3. ANÁLISIS ESTADÍSTICO E INFERENCIAL

Durante este capítulo, se ven gráficamente los datos que son proporcionados por separados, para ver relaciones y aplicar métodos de análisis estadístico e inferencia estadística con ayuda del software R.

3.1 RESUMEN

A continuación, en la tabla 3.1 se muestra el valor mínimo, primer quintil, mediana, promedio tercer quintil y el valor máximo en cada una de las variables. Notar que Var es la varianza y CV es el coeficiente de variación.

Tabla 3.1: Summary para las variables

Atributo	Mín	1 qu.	Med	Media	3 qu.	Máx	Var	CV
Identificador	1	1	2	1.984	3	3	0.601	0.399
Alcohol	11.03	12.36	13.05	13.00	13.68	14.83	0.659	0.062
Malic acid	0.740	1.603	1.865	2.336	3.083	5.800	1.248	0.478
Ash	1.360	2.210	2.360	2.367	2.558	3.230	0.075	0.115
Alkalinity of ash	10.60	17.20	19.50	19.49	21.50	30.00	11.153	0.171
Magnesium	70.00	88.00	98.00	99.74	107.00	162.00	204	0.143
Total phenols	0.980	1.742	2.355	2.295	2.800	3.880	0.392	0.272
Flavanoids	0.340	1.205	2.135	2.029	2.875	5.080	0.998	0.492
Nonflavanoids	0.1300	0.2700	0.3400	0.3619	0.4375	0.6600	0.015	0.343
Proanthocyanins	0.410	1.250	1.555	1.591	1.950	3.580	0.328	0.359
Color intensity	1.280	3.220	4.690	5.058	6.200	13.000	5.374	0.458
Hue	0.4800	0.7825	0.9650	0.9574	1.1200	1.7100	0.052	0.238
0D280/0D315	1.270	1.938	2.780	2.612	3.170	4.000	0.504	0.271
Proline	278.0	500.5	673.5	746.9	985.0	1680.0	99166	0.421

3.2 MATRIZ DE CORRELACIÓN

La información de la tabla no dice mucho, por lo que información sobre la correlación entre variables podría mostrar algo más, y para esto se muestra la matriz de correlaciones en la tabla 3.2.

Tabla 3.2: Matriz de correlaciones de las variables

Var	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	-.33	.44	-.05	.52	-.21	-.72	-.85	.49	-.5	.27	-.62	-.79	-.63
2	-.33	1	.09	.21	-.31	.27	.29	.24	-.16	.14	.55	-.07	.07	.64
3	.44	.09	1	.16	.29	-.05	-.34	-.41	.29	-.22	.25	-.56	-.37	-.19
4	-.05	.21	.16	1	.44	.29	.13	.12	.19	.01	.26	-.07	0	.22
5	.52	-.31	.29	.44	1	-.08	-.32	-.35	.36	-.2	.02	-.27	-.28	-.44
6	-.21	.27	-.05	.29	-.08	1	.21	.2	-.26	.24	.2	.06	.07	.39
7	-.72	.29	-.34	.13	-.32	.21	1	.86	-.45	.61	-.06	.43	.7	.5
8	-.85	.24	-.41	.12	-.35	.2	.86	1	-.54	.65	-.17	.54	.79	.49
9	.49	-.16	.29	.19	.36	-.26	-.45	-.54	1	-.37	.14	-.26	-.5	-.31
10	-.5	.14	-.22	.01	-.2	.24	.61	.65	-.37	1	-.03	.3	.52	.33
11	.27	.55	.25	.26	.02	.2	-.06	-.17	.14	-.03	1	-.52	-.43	.32
12	-.62	-.07	-.56	-.07	-.27	.06	.43	.54	-.26	.3	-.52	1	.57	.24
13	-.79	.07	-.37	0	-.28	.07	.7	.79	-.5	.52	-.43	.57	1	.31
14	-.63	.64	-.19	.22	-.44	.39	.5	.49	-.31	.33	.32	.24	.31	1

La correlación muestra la relación **lineal** entre las variables, y busca saber si es positiva o negativa, junto con la fuerza de éstas.

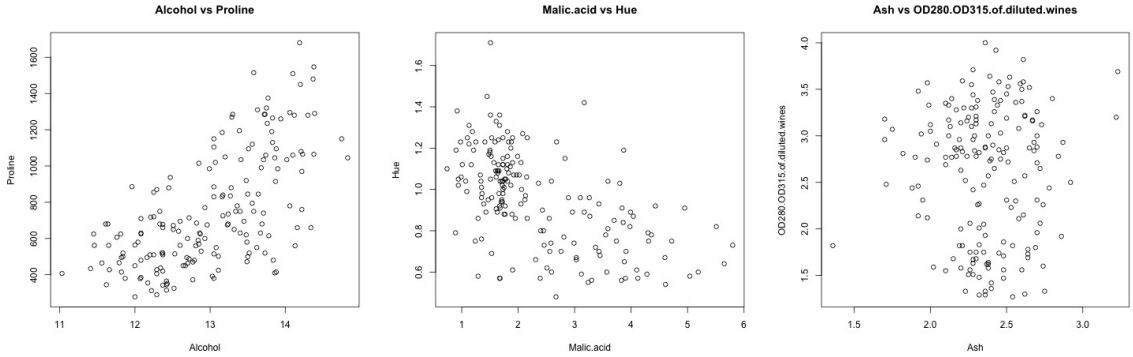
Las variables más interesantes de analizar con esta matriz son:

- **Ash vs OD280/OD315 of diluted wines:** La matriz muestra que estas dos variables son totalmente independientes, al menos, linealmente.
- **Total phenols vs Flavanoids:** La matriz ha mostrado que el $r = -0{,}86$ tiene una relación negativa, por lo tanto se puede concluir que, cuando hay más fenoles, normalmente baja la cantidad de flavanoides. Esto dándole el significado de los atributos significa que entre más cuerpo tenga un vino, menos color tiene.

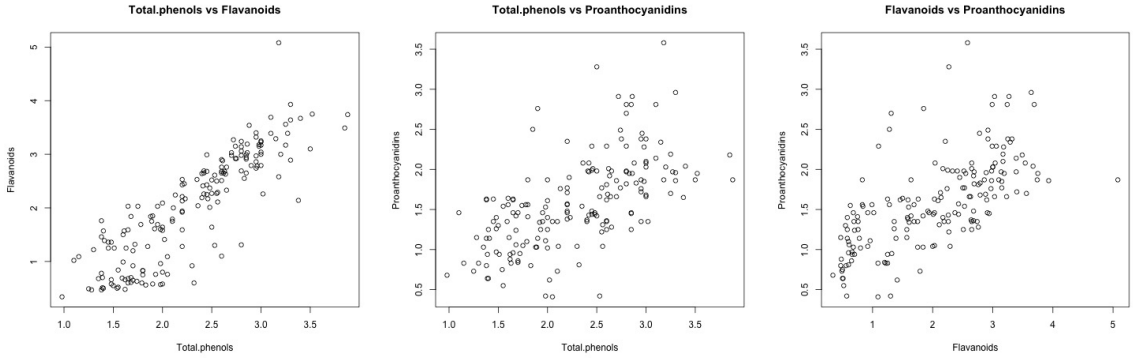
3.3 RELACIÓN DE ATRIBUTOS

Al ver las relaciones graficadas, hay que notar que algunas funciones pueden tener un comportamiento totalmente independiente como otras que sí dependen en algun grado de la otra. En la figura 3.1 se muestran las más importantes.

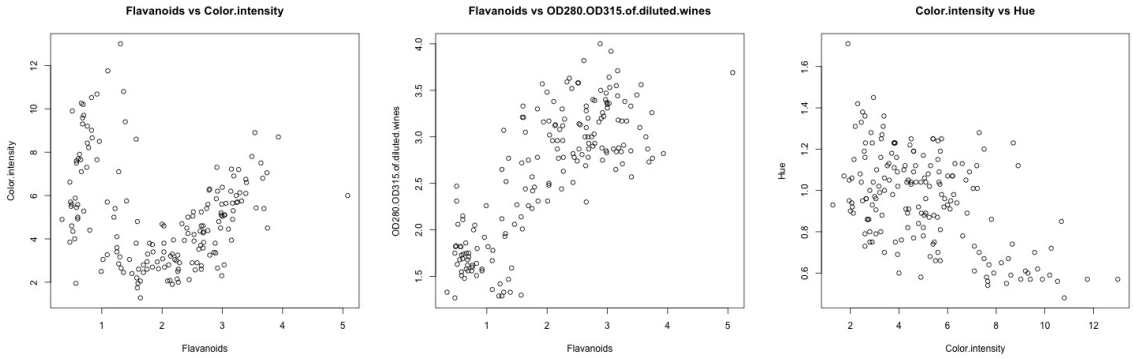
Como se puede apreciar, la matriz de correlación da una gran iniciativa de que relaciones son dependientes, aunque hay casos como en la subfigura (g) que podría ser una



(a) Atributo 2 y 14, $r = 0{,}21$ (b) Atributo 3 y 12, $r = 0{,}56$ (c) Atributo 4 y 13, $r = 0$



(d) Atributo 7 y 8, $r = 0{,}86$ (e) Atributo 7 y 10, $r = 0{,}61$ (f) Atributo 8 y 10, $r = 0{,}65$



(g) Atributo 8 y 11, $r = 0.{,}17$ (h) Atributo 8 y 13, $r = 0{,}79$ (i) Atributo 11 y 12, $r = 0{,}52$

Figura 3.1: Atributos con gráficos destacados.

relación cuadrática pero el resultado entregado es bastante pequeño, indicando que linealmente no tienen una relación fuerte.

3.4 HISTOGRAMA PARA LAS VARIABLES

Por supuesto que los datos se visualizan mejor con gráfico adecuado, por lo que se adjuntan imágenes donde aparece el histograma de las variables más importantes en la figura 3.2. Notar que la mayoría de los histogramas esta dividido en 10 barras para una fácil visualización.

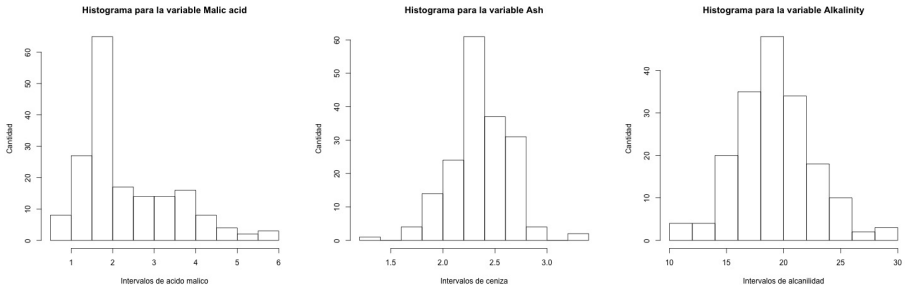
3.5 TEST DE SHAPIRO-WILK

Una pregunta frecuente, es saber si los datos obtenidos provienen de una distribución normal o no, y por tanto, es conveniente ocupar algún método para comprobar estos fines. A continuación se aplica el test de Shapiro-Wilk para ver la normalidad de los datos. Notar que R (p-value $<$ 0.1) significa que se concluye que los datos no vienen de una distribución normal, mientras que A (p-value $>$ 0.1) concluye que no se puede rechazar la hipótesis nula, la cual es que la población está distribuida normalmente.

Tabla 3.3: Resultados entregado al aplicar test de Shapiro-Wilk

Atributo	**W**	**p-value**	**Conclusión**
Alcohol	0.88878	2.946e-10	R
Malic acid	0.98395	0.03868	R
Ash	0.99023	0.2639	A
Alkanility of ash	0.93833	6.346e-07	R
Magnesium	0.97668	0.004395	R
Total phenols	0.95453	1.679e-05	R
Nonflavanoids phenols	0.96252	0.0001055	R
Proanthocyanidins	0.98072	0.01445	R
Color intensity	0.94032	9.229e-07	R
Hue	0.98134	0.01743	R
OD280/OD315 of diluted wines	0.94505	2.316e-06	R
Proline	0.93119	1.741e-07	R

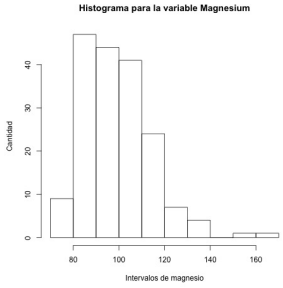
Estos resultados dicen que la mayoría de los datos no se distribuye normalmente entre los datos medidos, por lo que no se pueden asumir suposiciones que son parte de la distribución anteriormente mencionada.



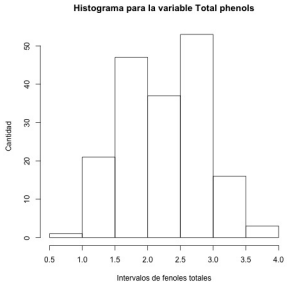
(a) Atributo 3

(b) Atributo 4

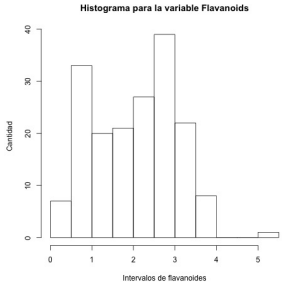
(c) Atributo 5



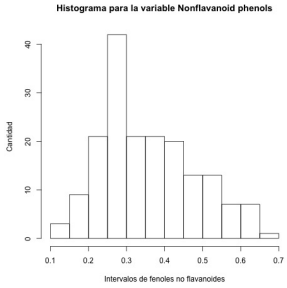
(d) Atributo 6



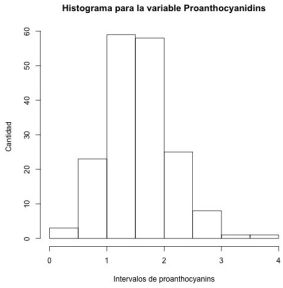
(e) Atributo 7



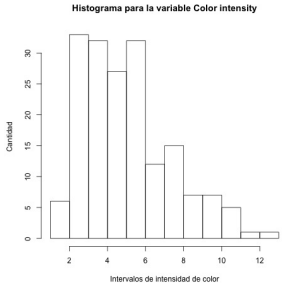
(f) Atributo 8



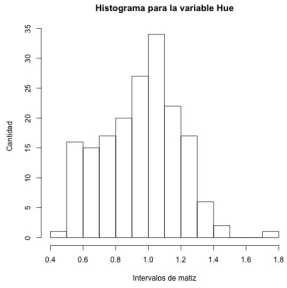
(g) Atributo 9



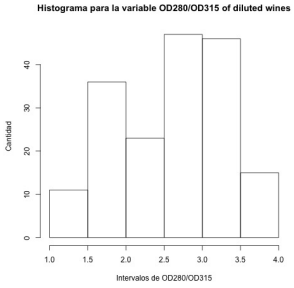
(h) Atributo 10



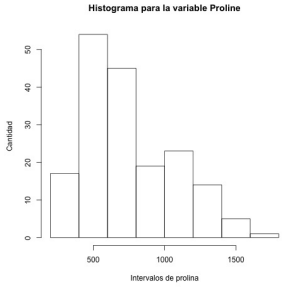
(i) Atributo 11



(j) Atributo 12



(k) Atributo 13



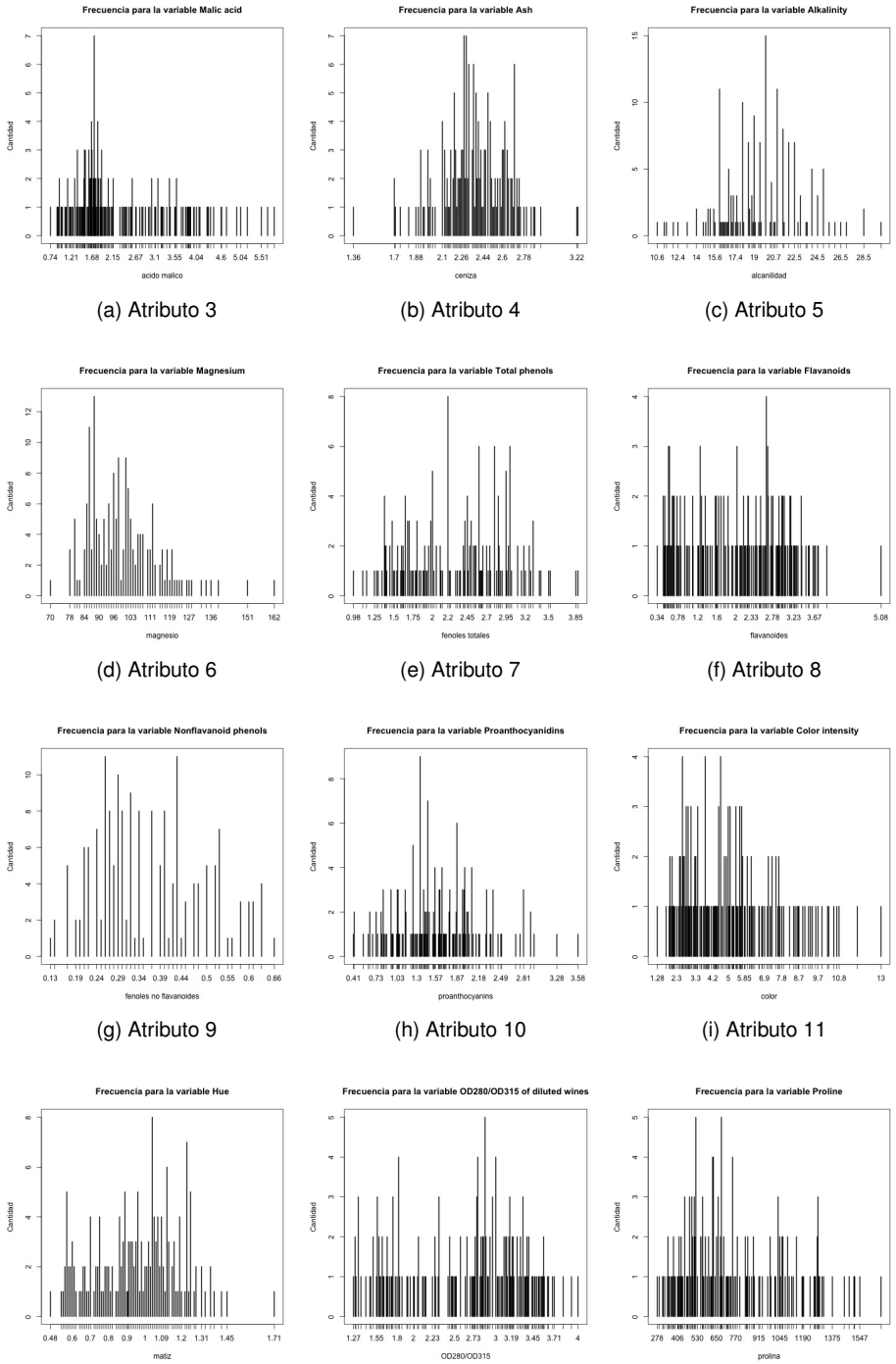
(l) Atributo 14

Figura 3.2: Histograma para las variables más importantes.

3.6 GRÁFICOS DE FRECUENCIA

En la sección anterior se vio que solamente uno de los atributos se distribuye normalmente, sin embargo, los histogramas muestran que hay varios otros atributos que sí se asemejan a la normal. Para resolver estas contradicciones, se realiza un gráfico de frecuencia para todas las variables y así ver realmente la distribución. Esto se encuentra en la figura 3.3.

Gracias a estos gráficos es que se puede notar que ya no pertenecen a una distribución normal, aun cuando los histogramas insinuaban lo contrario. Esto pasó por el intervalo elegido para éstos, el cual no siempre es adecuado y no hay formula exacta para saber cuantos intervalos ocupar (existen referencias, pero no siempre son pertinentes).



(a) Atributo 3 (b) Atributo 4 (c) Atributo 5

(d) Atributo 6 (e) Atributo 7 (f) Atributo 8

(g) Atributo 9 (h) Atributo 10 (i) Atributo 11

(j) Atributo 12 (k) Atributo 13 (l) Atributo 14

Figura 3.3: Gráfico de frecuencia para las variables más importantes.

CAPÍTULO 4. CONCLUSIONES

Durante el laboratorio, se realizaron análisis básicos que pudieron ser realizados fácilmente con ayuda del lenguaje de programación R, el cual tiene funciones predefinidas que ayudaron a evitar cálculos tediosos y optimizar el tiempo.

Para el análisis estadístico básico se empleó la función $summary()$, la cual entregó los datos referentes al promedio, mediana, quintiles, etc. Sin embargo, este análisis es muy poco útil cuando se quiere concluir algo, el promedio indicaría una probabilidad a priori que no es efectiva prediciendo resultados.

Luego se realizaron otros tipos de análisis más avanzados que incluían regresiones lineales para ver la dependencia de variables y su comportamiento.

Sin dejar a un lado, se hizo un histograma el cual fue interpretado como que varios de los atributos se comportaban de forma normal, sin embargo, al hacer un test de Shapiro-Wilk el cual da un índice para ver la posibilidad de si los datos pertenecen o no a una normal, fue todo lo contrario, ya que solo uno dio positivo y se podía asumir su normalidad. Ante la contradicción, un diagrama de frecuencia fue lo mejor para ver que tan normal eran los datos de forma visual, sin formulas.

Un problema presentado fue el anteriormente mencionado, pero después de ver el gráfico de frecuencia de las variables, se recopiló información sobre la normal y en base al promedio, entre el rango de $-\sigma$ y $+\sigma$ deberían estar aproximadamente el 67 % de los datos, y no cumplía eso. Al hacer esto se descubrió otra forma más visual de ver la normalidad de los datos, que no fue explícitamente estudiada anteriormente.

BIBLIOGRAFÍA

Leardi, R. (1991). Wine data set. Recuperado desde `https://archive.ics.uci.edu/ml/datasets/wine"`.