UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 1 - Análisis Estadístico

Integrantes:	Jaime Yefi
	Mónica Araneda
Curso:	Análisis de Datos
Sección	13116-0 V-35
Profesor:	Max Chacón Pacheco
Ayudante:	Javier Arredondo

21 de Noviembre de 2020

Tabla de contenidos

1. Introducción **1**

2. Descripción del Problema **2**

2.1. Descripción de la base de datos 2

2.1.1. Revisión preliminar de los datos 2

2.1.2. Análisis exploratorio 2

2.2. Descripción de base de datos 4

2.3. Descripción de clases y variables 5

2.3.1. Clase objetivo 5

2.3.2. Variables cualitativas 5

2.3.3. Variables cuantitativas 8

3. Análisis Estadístico e Inferencial **9**

3.1. Análisis variable de clase 9

3.2. Análisis variables cuantitativas 9

3.3. Análisis variables cualitativas 13

3.4. Análisis de correlación 19

4. Conclusiones **21**

Bibliografía **23**

1. Introducción

Dentro del análisis de datos, uno de los temas trascendentales corresponde al conocimiento del problema en todo su contexto. Es así como en el presente trabajo, se presenta una serie de síntomas, con una clase objetivo, que es determinar si la persona va a vivir (LIVE) o va a morir (DIE). Referente al problema de base, cabe señalar que la Hepatitis es una enfermedad que afecta al hígado en distintas formas, y se clasifica en 5 tipos: A, B, C, D y E (*también conocidas como VHA, VHB, VHC, VHD y VHE*). El hígado es un organismo vital que se encarga de realizar distintos procesos tales como procesar nutrientes, filtrar la sangre, descomponer sustancias químicas, por lo que un hígado inflamado o enfermo, interrumpe dichas funciones y es un escenario complejo (y en algunos casos puede ser fulminante y llevar a la muerte). (11)

Dentro de todos los tipos de hepatitis existen algunos, que son asíntomáticos, es decir, no presentan síntomas evidentes de la enfermedad, sino hasta cuando esta se agrava. Por otra parte y un dato no menor, es que esta enfermedad es más común de lo que uno piensa, y es que al menos en este año la OMS indicó que 9 de cada 10 personas tienen hepatitis, muchos de ellos sin saber que tienen dicha enfermedad. Lo anterior sumado a que muchos de los síntomas pueden confundirse con otras dolencias. (12) Para el desarrollo de las experiencias del presente laboratorio, se desarrollaron las rutinas en un ambiente con Jupyter Noteebook, con el lenguaje de programación Python.

Todo lo anterior, hace que sea tremendamente relevante determinar en base a los síntomas que presenta una persona sospechosa de tener algún tipo de Hepatitis (independiente de su tipo), establecer un diagnóstico temprano, de forma de tener un tratamiento adecuado, de forma temprana, dado que los tratamientos pueden tomar incluso, hasta tres meses de duración. Otro punto importante a considerar, es que de acuerdo a los datos de la muestra, el 20,65 % de las personas que tuvieron síntomas de Hepatitis (uno de cada cinco personas), lo que hace más relevante el estudio para determinar cuáles son las causas que provocan la muerte en un paciente.

2. Descripción del Problema

El problema planteado en la base de datos Hepatitis, es a partir de los síntomas presentados determinar en qué casos pudieran finalizar con la clase objetivo, con el valor DIE, es decir, se debe determinar qué características son las determinantes en que un paciente fallezca cuando este presente síntomas asociados a Hepatitis. Para lo anterior, se aplicarán técnicas de estadística descriptiva e inferencial, junto con la descripción de cada una de las variables y su posible correlación con la clase objetivo. Se realizará una breve descripción de de cada una de las variables para entender el contexto del problema.

2.1. Descripción de la base de datos

Para el presente laboratorio, trabajamos con el *Data Set* "Hepatitis" del repositorio de *Machine Learning* de la Universidad de California (4).

2.1.1. Revisión preliminar de los datos

Como podemos ver en la Figura 1, el conjunto de datos tiene 155 filas correspondientes al número de pacientes incluidos en este estudio, y 20 columnas, correspondientes a las características recolectadas para cada paciente.

```
# Métricas de individuos, variables
df.shape
```

```
(155, 20)
```

Figura 1: Individuos y variables presentes en la base de datos

2.1.2. Análisis exploratorio

Una parte importante de hacer predicciones con técnicas de aprendizaje automático es realizar análisis de datos exploratorios. Esto es útil para conocer sus datos, mirarlos desde diferentes perspectivas, describirlos y resumirlos sin hacer ninguna suposición para detectar posibles problemas.

Primero, podemos inspeccionar nuestros datos para ver si necesitamos limpiarlos. Comenzaremos usando el comando *head*, que nos mostrará las primeras 5 filas de nuestro *DataFrame*, tal como se muestra en la figura 2.

```
# Primeras columnas
print(df.head())
```

```
   Class  AGE  SEX STEROID  ...  SGOT ALBUMIN PROTIME HISTOLOGY
0      2   30    2       1  ...    18     4.0       ?         1
1      2   50    1       1  ...    42     3.5       ?         1
2      2   78    1       2  ...    32     4.0       ?         1
3      2   31    1       ?  ...    52     4.0      80         1
4      2   34    1       2  ...   200     4.0       ?         1

[5 rows x 20 columns]
```

Figura 2: Revisión preliminar de los datos

Sin embargo al revisar los datos en la base, nos damos cuenta que existen valores sin especificar ("?") en distintas variables, lo que hace que los tipos de datos sean clasificados como de tipo *object*, tal como se muestra en la Figura 3.

```
[7]  # Tipos de datos del universo muestral
     df.dtypes
```

```
Class              int64
AGE                int64
SEX                int64
STEROID           object
ANTIVIRALS         int64
FATIGUE           object
MALAISE           object
ANOREXIA          object
LIVER BIG         object
LIVER FIRM        object
SPLEEN PALPABLE   object
SPIDERS           object
ASCITES           object
VARICES           object
BILIRUBIN         object
ALK PHOSPHATE     object
SGOT              object
ALBUMIN           object
PROTIME           object
HISTOLOGY          int64
dtype: object
```

Figura 3: Tipos de datos Iniciales

2.2. Descripción de base de datos

Variable	Descripción	Niveles
Class	Supervivencia del paciente	DIE, LIVE
AGE	Edad del paciente	10, 20, 30, 40, 50, 60, 70, 80
SEX	Sexo del paciente	male, female
STEROID	¿El paciente estaba recibiendo esteroides?	no, yes
ANTIVIRALS	¿El paciente está en tratamiento antiviral	no, yes
FATIGUE	¿El paciente sufre fatiga crónica?	no, yes
MALAISE	¿El paciente tiene dolor abdominal?	no, yes
ANOREXIA	¿El paciente presenta anorexia nerviosa?	no, yes
LIVER BIG	¿El paciente tiene agrandamiento del hígado?	no, yes
LIVER FIRM	¿El paciente tiene endurecimiento en el hígado?	no, yes
SPLEEN PALPABLE	¿El paciente tiene endurecimiento en el bazo?	no, yes
SPIDERS	¿El paciente presenta arañas vasculares?	no, yes
ASCITES	¿El paciente muestra signos de ascitis o acumulación de líquido?	no, yes
VARICES	¿El paciente presenta síntomas de varices esofágicas ?	no, yes
BILIRUBIN	Niveles de Bilirrubina en la sangre, medida en mg/dL	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
ALK PHOSPHATE	Niveles de fosfatasa alcalina en la sangre (medida en (UI/L)	33, 80, 120, 160, 200, 250
SGOT	Niveles de fostatasa alcalina en la sangre, medida en (UI/L)	13, 100, 200, 300, 400, 500
ALBUMIN	Niveles de albúmina en la sangre, medidos en g/dL	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
PROTIME	Tiempo de protrombina	10, 20, 30, 40, 50, 60, 70, 80, 90
HISTOLOGY	¿El paciente, se realizó un hemograma?	no, yes

2.3. Descripción de clases y variables

Revisamos las descripciones de cada uno de los individuos presentes en la muestra de datos, de acuerdo a las descripciones presentes en el repositorio de la base de datos (*Data Set Description*) (4).

2.3.1. Clase objetivo

La Clase Objetivo corresponde a la variable *Class*, y el objetivo es a partir de las variables presentes en el universo muestral, predecir el valor DIE.

2.3.2. Variables cualitativas

Variable SEX

Corresponde al sexo de los individuos, en donde el valor 1 corresponde a sexo femenino y 2 al sexo masculino.

Variable STEROID

Indica si hubo interacción previas con esteroides anabolizantes, los cuales aumentan entre un 20 y un 30 % el riesgo de sufrir una hepatitis tóxica. Los esteroides anabolizantes suelen utilizarse como fármacos que incrementan la masa muscular, y son utilizados por fisicocultoristas y deportistas de alto rendimiento.(1). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable ANTIVIRALS

Esta variable ndica si hubo tratamiento con antivirales (como Paritaprevir o ombitasvir). (22). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable FATIGUE

La fatiga crónica o cansancio es uno de los primeros síntomas que aparecen en los cuadros de hepatitis. Si bien es cierto la fatiga crónica es un síntoma que está presente en muchas enfermedades, es común en aquellas de origen inflamatorio. (14). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable MALAISE

Corresponde a si tiene o no dolor (o malestar) en la zona abdominal. Este dolor se ubica caracteristicamente en la parte superior derecha del abdomen, debajo de las costillas. Incluso en algunas circunstancias, un simple examen del abdomen permite detectar la inflamación hepática. (13). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable ANOREXIA

La anorexia nerviosa es uno de los síntomas que pueden aparecer en conjunto con la hepatitis o fallo hepático agudo. (2) . En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable LIVER BIG

Este síntoma corresponde al crecimiento anormal del hígado, la cual es síntoma de muchas enfermedades, una de las cuales es la hepatitis. En medicina este síntoma se denomina "hepatomegalia". (19).En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable LIVER FIRM

Es el endurecimiento del hígado, la cual tiene otro término en medicina, que es la enfermedad hepática crónica con fibrosis, o comúnmente llamada Cirrosis hepática. Dicho síntoma tiene por origen, las hepatitis virales B o C. (15). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable SPLEEN PALPABLE

Es un síntoma similar a LIVER BIG, con la diferencia que corresponde al Bazo. Es un crecimiento anormal del Bazo, lo cual es muy difícil de detectar, dado que en exámenes de rayos x, el 16 % son de tamaño normal. Cabe señalar que el Bazo es un órgano que pesa 250 gramos, que disminuye con la edad y que comúnmente no se palpa. (17). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable SPIDERS

Esta variable corresponde a las comúnmente llamadas "arañas vasculares", las cuales corresponden a lesiones cutáneas que se presentan en la piel. Estas son frecuentes en mujeres embarazadas, desórdenes hormonales (derivado del uso de anticonceptivos), enfermedades hepáticas. Cerca del 33 % de los pacientes que tienen cirrosis, presentan arañas vasculares. (16). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable ACITES

Corresponde al fenómeno de Ascitis o más conocido como acumulación de líquido en el abdomen y piernas. Este síntoma es derivado de una enfermedad que causa daño hepático grave, como la Hepatitis B o C, consumo de alcohol durante muchos años, o hígado graso. (15). (20). En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable VARICES

Corresponde a las varices esofágicas, las cuales son una complicación común derivada de la cirrosis. Una de las complicaciones de este síntoma es la altísima mortalidad en los casos que estas varices sufran una ruptura. (23)En la base de datos se presenta con los valores 1, que equivale a "no" y 2 que equivale a "yes".

Variable HISTOLOGY

Corresponde a lo que comúnmente se llama hemograma, lo cual es un examen a la sangre, el cual indica distintas métricas. En el caso de la hepatitis, existe un bajo nivel de hematíes, lo que se traduce como Anemia. (Concepció Bartres Viñas)

2.3.3. Variables cuantitativas

Variable AGE

Esta variable corresponde a la edad de los individuos, la cual se presenta como un valor entero.

Variable BILIRUBIN

La Bilirrubina es un pigmento de color amarillo anaranjado que se almacena en en hígado formando parte de la bilis. Cuando esta se acumula, es responsable del color amarillo en la piel o en los ojos (ictericia). Es frecuente en las hepatitis A, B y E. (18)(Concepció Bartres Viñas). Los valores normales de la Bilirrubina son entre 0,1 a 1,2 mg/dL (6).

Variable ALK PHOSPHATE

Corresponde al resultado del examen de fostatasa alcalina (*Alkaline Phosphatase*), la cual es una enzima que está presente en todo el cuerpo, sin embargo ésta se concentra en el hígado. Cuando existe una enfermedad hepática, esta enzima se filtra al torrente sanguíneo, aumentando su presencia. (21). En la base de datos está medida en unidades internacionales por litro y los valores normales están entre 44 a 147 (UI/L) (7).

Variable SGOT

Es otra enzima presente en el organismo, llamada *Aspartato Aminostransferasa*. Su disminución está asociada a enfermedades hepáticas, sin embargo también está asociada a otras no hepáticas como el hipertiroidismo o enfermedad celíaca. (3). Para medir el nivel de esta variable, se utiliza la prueba de sangre de aspartato aminotransferasa, la cual en resultados normales, tiene valores entre 8 a 33 U/L (9).

Variable ALBUMIN

Corresponde a la Albúmina, la cual es la principal proteína de la sangre con múltiples funciones. Se produce en el hígado y su descenso en sangre indica que el hígado no está funcionando de forma correcta. (Concepció Bartres Viñas). Los valores normales en la sangre van de 3.4 a 5.4 g/dL (8)

Variable PROTIME

Es una prueba que se realiza en el laboratorio con una muestra de sangre que mide la capacidad de coagulación de la sangre. Las proteínas relacionadas con la coagulación sanguínea se producen en el hígado. Si el hígado no funciona bien debido a un daño agudo o crónico, se puede alterar la coagulación sanguínea al no producir dichas proteínas y la sangre tarda más tiempo en coagular. En el caso de una hepatitis aguda su alteración indica que la hepatitis es grave y que existe un fallo hepático (Concepció Bartres Viñas). Los valores normales van de 11 a 13.5 segundos (10).

3. Análisis Estadístico e Inferencial

3.1. Análisis variable de clase

A partir de la información existente en la BD, obtenemos los valores de mortalidad de la clase objetivo:

```
# Pacientes fallecidos (para obtener la tasa de mortalidad)
pacientes_fallecidos = np.sum(df['Class'] == 1)
tasa_mortalidad = round((pacientes_fallecidos/df.shape[0])*100,2)
#pacientes que viven

total_de_pacientes = df.shape[0]
pacientes_vivos = (np.sum(df['Class'] == 2)/total_de_pacientes)*100
print("Tasa de mortalidad = ",tasa_mortalidad,"%")
print("Tasa de pacientes vivos:", round(pacientes_vivos,2),"%")
```

```
Tasa de mortalidad =  20.65 %
Tasa de pacientes vivos: 79.35 %
```

Figura 4: Cálculo de tasa de mortalidad

Ahora podemos comprobar si nuestro conjunto de datos sufre un desequilibrio de clases, podemos calcular qué porcentaje de los datos pertenece a cada categoría.

3.2. Análisis variables cuantitativas

Se calculó tanto las medidas de tendencia central como las medidas de dispersión de nuestra variable de clase, y podemos obtener un resumen estadístico con las principales medidas. Pandas nos ofrece el método describe, un comando con el cual se ve un resumen de

las principales medidas estadísticas de las variables cuantitativas, tal como se muestra en la figura 5. Esto se abordará en detalle en cada una de las variables.

```
# Descripción de variables cuantitativas
cont_vars = ['AGE', 'BILIRUBIN', 'PROTIME', 'ALBUMIN', 'ALK PHOSPHATE', 'SGOT']
df[cont_vars].describe()
```

	AGE	BILIRUBIN	PROTIME	ALBUMIN	ALK PHOSPHATE	SGOT
count	155.000000	149.000000	88.000000	139.000000	126.000000	151.00000
mean	41.200000	1.427517	61.852273	3.817266	105.325397	85.89404
std	12.565878	1.212149	22.875244	0.651523	51.508109	89.65089
min	7.000000	0.300000	0.000000	2.100000	26.000000	14.00000
25%	32.000000	0.700000	46.000000	3.400000	74.250000	31.50000
50%	39.000000	1.000000	61.000000	4.000000	85.000000	58.00000
75%	50.000000	1.500000	76.250000	4.200000	132.250000	100.50000
max	78.000000	8.000000	100.000000	6.400000	295.000000	648.00000

Figura 5: Máximos, mínimos y medidas centrales (variables cuantitativas)

Variable AGE

Los datos de la variable AGE están agrupados entre el primer y tercer cuartil. No se detectan valores atípicos, tal como se aprecia en la figura 6.

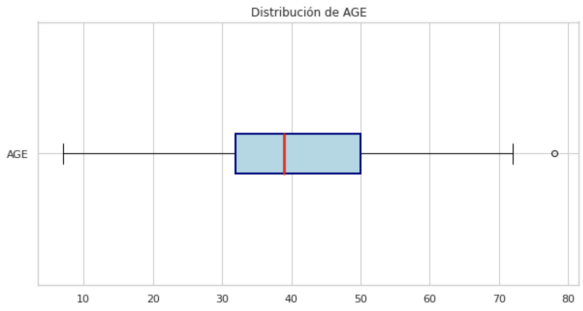

Figura 6: Distribución de AGE

Variable BILIRUBIN

Los datos de la variable BILIRUBIN están agrupados entre el primer cuartil. Se detectan valores atípicos, tal como se aprecia en la figura 7.

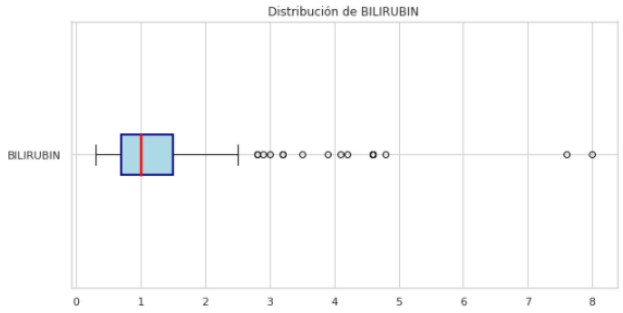

Figura 7: Distribución de BILIRUBIN

Variable PROTIME

Los datos de la variable PROTIME están agrupados entre el segundo y tercer cuartil. No se detectan valores atípicos, tal como se aprecia en la figura 8.

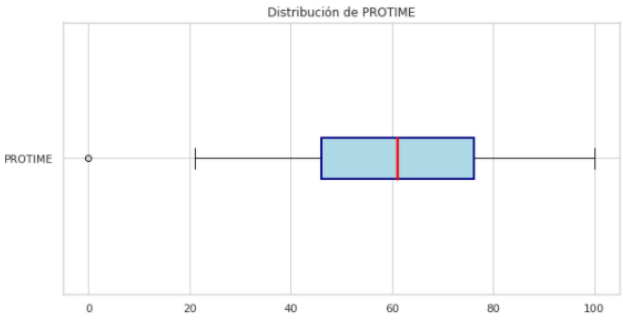

Figura 8: Distribución de PROTIME

Variable ALBUMIN

Los datos de la variable ALBUMIN están agrupados entre el segundo y tercer cuartil. No se detectan valores atípicos, tal como se aprecia en la figura 9.

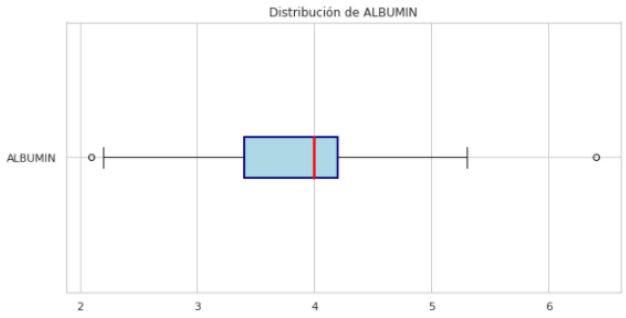

Figura 9: Distribución de ALBUMIN

Variable ALK PHOSPHATE

Los datos de la variable ALK PHOSPHATE están agrupados entre el segundo y tercer cuartil. No se detectan valores atípicos, tal como se aprecia en la figura 10.

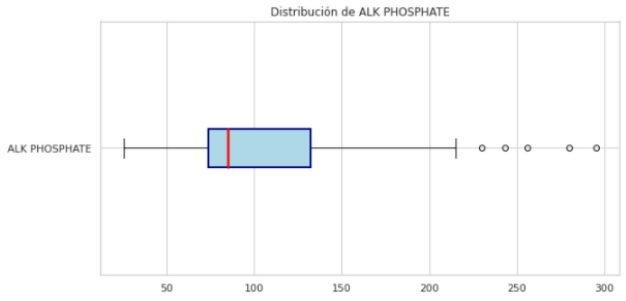

Figura 10: Distribución de ALK PHOSPHATE

Variable SGOT

Los datos de la variable SGOT están agrupados entre el primer cuartil. Se detectan valores atípicos, tal como se aprecia en la figura 11.

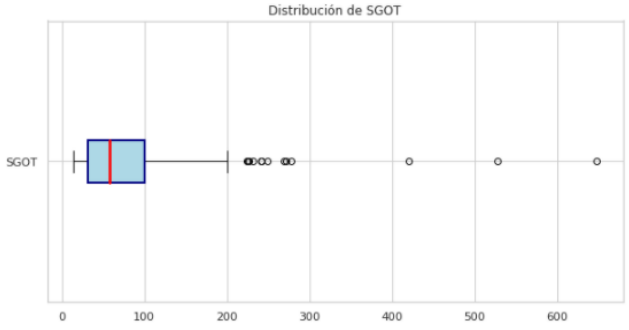

Figura 11: Distribución de SGOT

3.3. Análisis variables cualitativas

Para efectos de análisis, no tiene mucho sentido analizar medidas de media central sobre las variables cualitativas, por lo que procedimos a contar la presencia de cada una de estas variables en la base de datos, tal como se presenta en la figura 12:

```
disc_vars = ['SEX', 'STEROID', 'ANTIVIRALS', 'FATIGUE','MALAISE','ANOREXIA','LIVER BIG','LIVER FIRM','SPLEEN PALPABLE', 'SPIDERS',
             'ASCITES','VARICES', 'HISTOLOGY']
datos = df[disc_vars].apply(pd.Series.value_counts)
#datos.index=['DIE', 'LIVE']
datos
```

	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER BIG	LIVER FIRM	SPLEEN PALPABLE	SPIDERS	ASCITES	VARICES	HISTOLOGY
1.0	139	76	24	100	61	32	25	60	30	51	20	18	85
2.0	16	78	131	54	93	122	120	84	120	99	130	132	70

Figura 12: Resumen de Variables cuantitativas

No obstante lo anterior, es interesante revisar la interacción de las variables cualitativas frente a la Clase objetivo. Para ello analizaremos cada una de las variables en particular.

Relación entre SEX y la clase objetivo

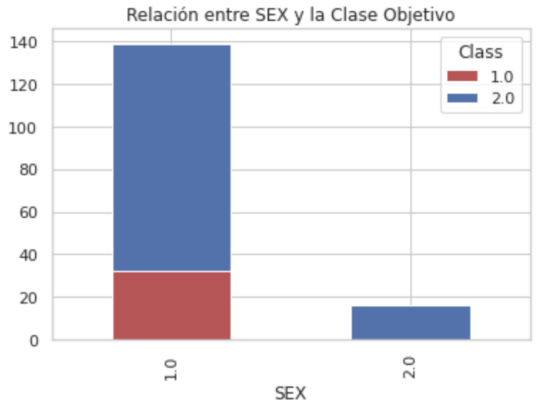

Figura 13: Relación entre SEX y la Clase objetivo

De la figura 13, se puede inferir que si el paciente es mujer (SEX = 2), entonces la clase objetivo = 2 (Class = "LIVE"), es decir, que en toda la BD, ***no existen mujeres fallecidas de hepatitis***.

Relación entre STEROID y la clase objetivo

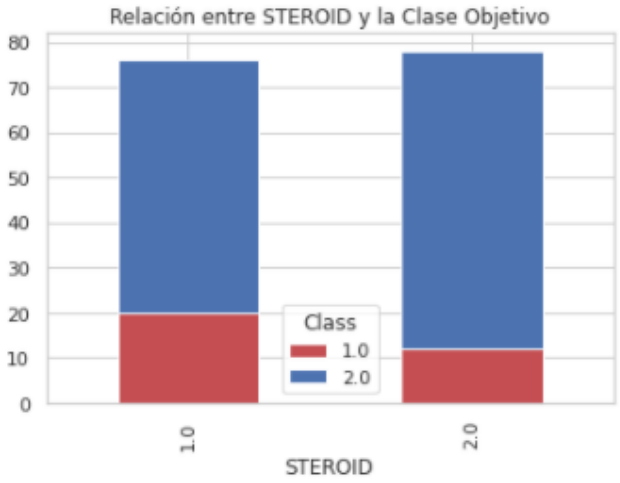

Figura 14: Relación entre la variable STEROID y la Clase objetivo

De la figura 14, se puede inferir que si el paciente usa o no esteroides, se ve afectado casi de igual manera en la clase objetivo.

Relación entre la variable FATIGUE y la clase objetivo

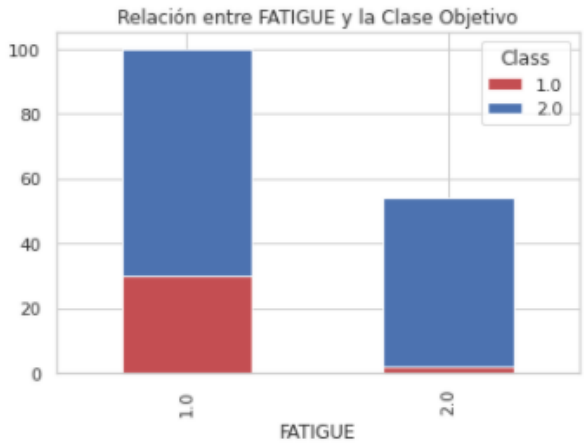

Figura 15: Relación entre FATIGUE y la Clase objetivo

De la figura 15, se puede inferir que si el paciente no presenta síntomas de fatiga, tiene mayor presencia el resultado de muerte, sin embargo no es tan categórico como en la variable AGE.

Relación entre la variable MALAISE y la clase objetivo

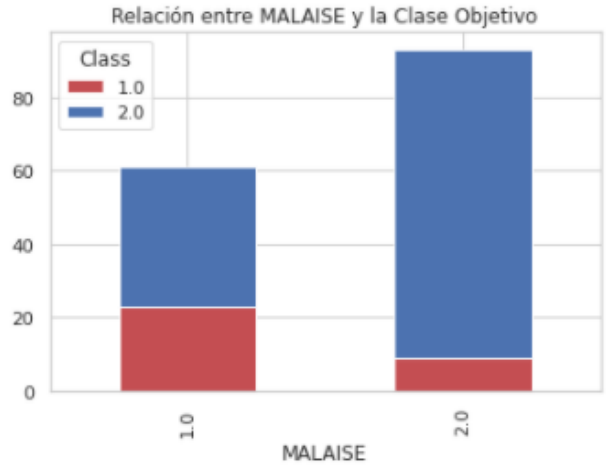

Figura 16: Relación entre MALAISE y la Clase objetivo

De la figura 16, se puede inferir que si el paciente no presenta síntomas de malestar abdominal, tiene mayor presencia el resultado de muerte.

Relación entre la variable ANOREXIA y la clase objetivo

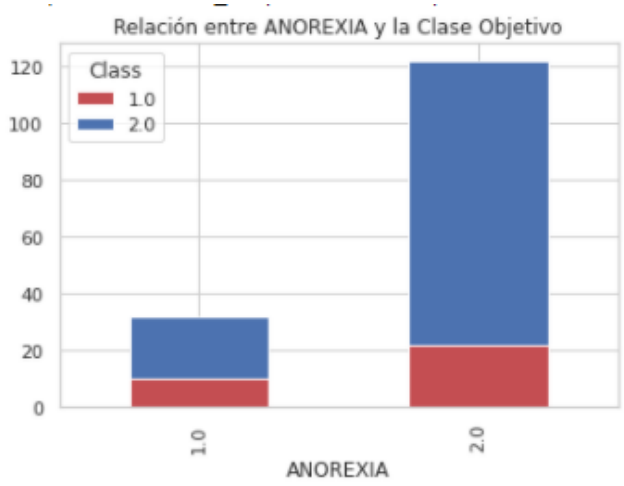

Figura 17: Relación entre ANOREXIA y la Clase objetivo

De la figura 17, se puede inferir que de la base de datos, existen más pacientes que presentan síntomas de anorexia nerviosa, de los cuales cuantitativamente son más que los que no presntan dicho síntoma.

Relación entre la variable LIVER BIG y la clase objetivo

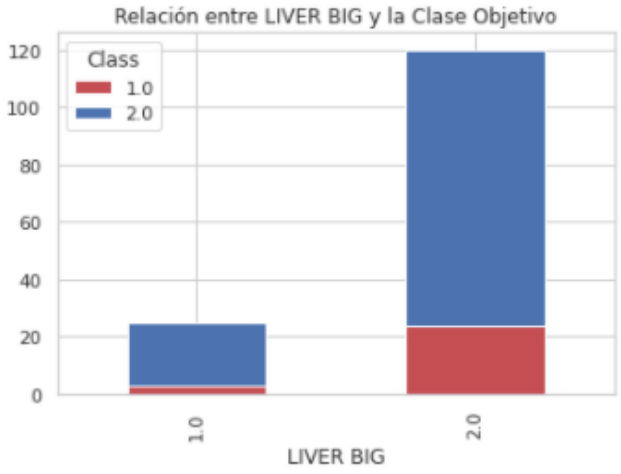

Figura 18: Relación entre LIVER BIG y la Clase objetivo

De la figura 18, se puede inferir que de la base de datos, existen más pacientes que presentan síntomas de anorexia nerviosa, de los cuales cuantitativamente son más que los que no presntan dicho síntoma.

Relación entre la variable LIVER FIRM y la clase objetivo

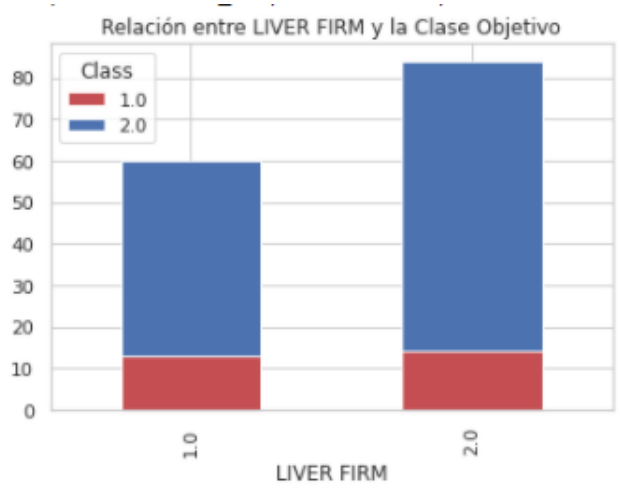

Figura 19: Relación entre LIVER FIRM y la Clase objetivo

De la figura 19, se puede inferir que de la base de datos, existen más individuos que presentan síntomas de endurecimiento de hígado, sin embargo cuantitativamente ambos se presentan con fallecimiento en la clase objetivo.

Relación entre la variable ASCITES y la clase objetivo

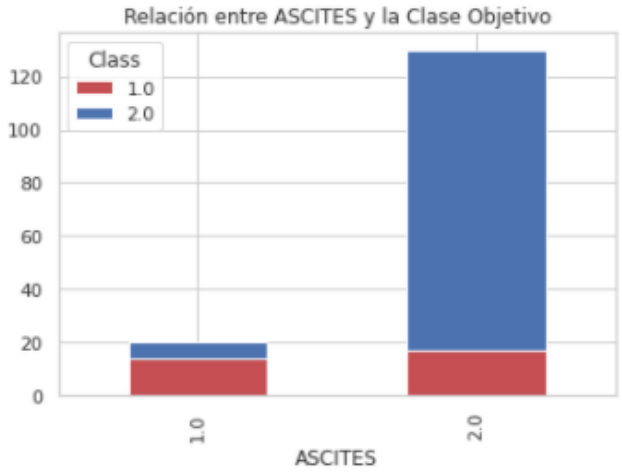

Figura 20: Relación entre ASCITES y la Clase objetivo

De la figura 20, se puede inferir que de la base de datos, existen más individuos que presentan síntomas de retención de líquido, sin embargo cuantitativamente ambos se presentan con fallecimiento en la clase objetivo.

Relación entre la variable VARICES y la clase objetivo

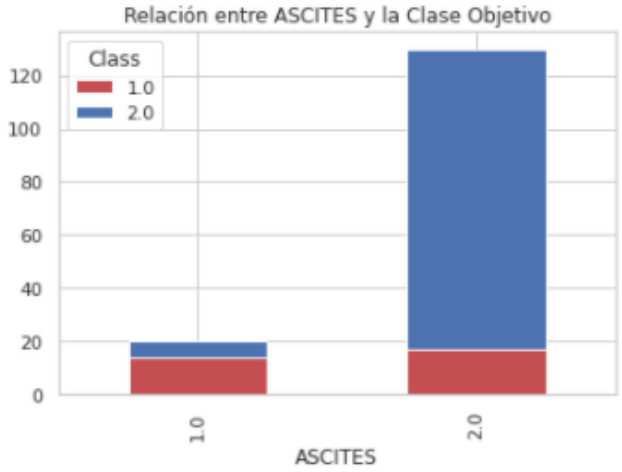

Figura 21: Relación entre VARICES y la Clase objetivo

De la figura 21, se puede inferir que de la base de datos, existen más individuos que presentan varices esofágicas, sin embargo cuantitativamente ambos son equivalentes en fallecimiento en la clase objetivo.

Relación entre la variable ANTIVIRALS y la clase objetivo

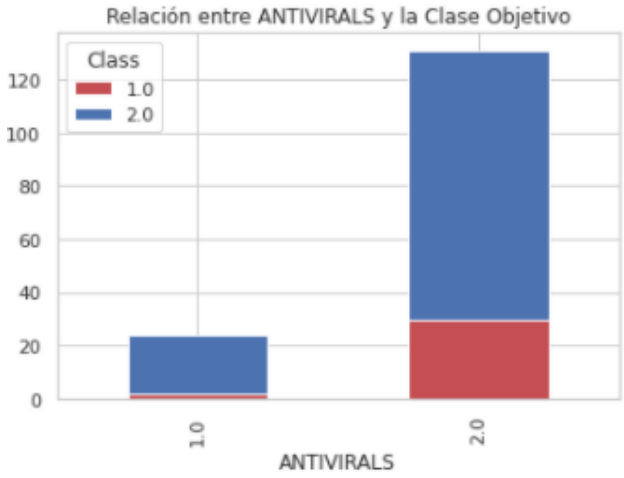

Figura 22: Relación entre ANTIVIRALS y la Clase objetivo

De la figura 22, se puede inferir que de la base de datos, existen más individuos que presentan tratamientos antivirales. Asimismo, cuantitativamente es mucho más significativo el resultado de muerte cuando existe un tratamiento con antivirales.

Relación entre la variable HISTOLOGY y la clase objetivo

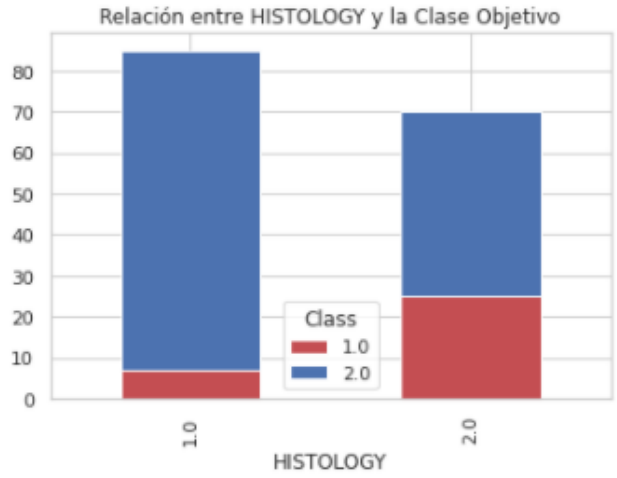

Figura 23: Relación entre HISTOLOGY y la Clase objetivo

De la figura 23, se puede inferir que de la base de datos, existen menos individuos que se practicaron un hemograma. Asimismo, cuantitativamente es mucho más significativo el resultado de muerte cuando se realiza este examen.

3.4. Análisis de correlación

Para esta tarea, usaremos el coeficiente de correlación de Pearson porque es un buen parámetro para evaluar la fuerza de la relación lineal entre dos variables. Para realizar el análisis de correlación con todas nuestras variables.

El gráfico presente en la figura 24 una matriz de 20 x 20 y llena de color cada celda en base al coeficiente de correlación del par que la representa. El valor en la posición (a, b) representa el coeficiente de correlación entre los elementos de la fila a y la columna b. Será igual al valor en la posición (b, a).

Los marcadores de los ejes denotan el rasgo que representa cada uno de ellos.

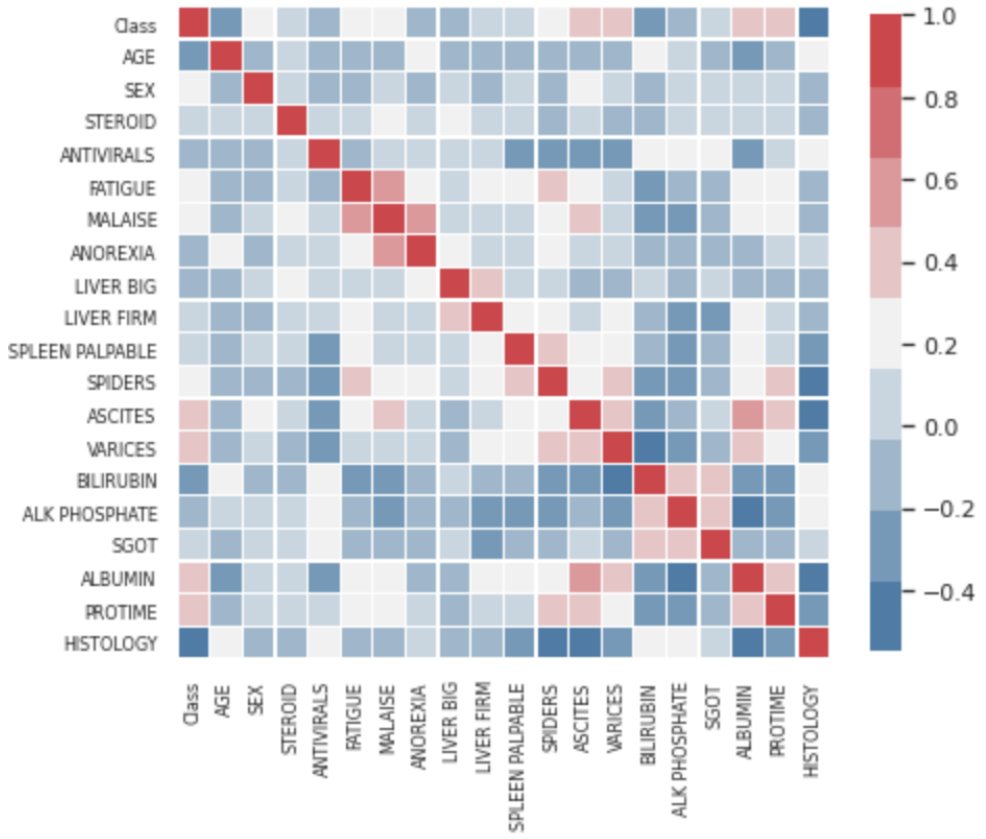

Figura 24: Gráfico de correlación - Método Pearson

- Un valor positivo cercano a 1,0 indica una fuerte correlación positiva, es decir, si el valor de una de las variables aumenta, el valor de la otra variable aumenta también.
- Un valor negativo grande (cercano a -1,0) indica una fuerte correlación negativa, es decir, que el valor de una de las variables disminuye al aumentar el de la otra y viceversa.
- Un valor cercano a 0 (tanto positivo como negativo) indica la ausencia de cualquier correlación entre las dos variables, y por lo tanto esas variables son independientes entre sí.
- Cada celda de la matriz anterior también está representada por sombras de un color. En este caso, los tonos más oscuros del color indican valores más pequeños, mientras que los tonos más brillantes corresponden a valores más grandes (cerca de 1). Esta escala se da con la ayuda de una barra de color en el lado derecho del gráfico.

4. Conclusiones

Respecto al problema y al análisis del conjunto de datos podemos observar que aunque nuestro conjunto de datos no está perfectamente equilibrado, el 79,35 % de los pacientes está contenido en la clase LIVE mientras que solo el 20,65 % está en la clase DIE, no hay un desequilibrio de clase por ende podemos continuar con nuestro análisis.

Podemos observar en la figura 6, que los pacientes pertenecen a un tramo de edad de 7 a 78 años, con una media de 41,2 y una mediana de 39. Faltan valores en la mayoría de las variables pero en particular en PROTIME, donde solo tenemos 88 observaciones. Si prestamos atención a las medias de las distintas variables, es interesante notar que presentan una varianza moderada; el rango que va de 1,42 (BILIRUBIN) a 105,35 (ALK PHOSPHATE). Además, las variables SGOT y ALK PHOSPHATE muestran una alta desviación estándar y su distribución podría estar sesgada a la derecha debido a que la media es mayor que la mediana.

En el caso de las variables categóricas, hay un marcado predominio de observaciones pertenecientes al nivel 1 en la variable SEXO lo que significa que el conjunto de datos incluye más pacientes mujeres que hombres. Asimismo, existen más observaciones en la clase 1 que en la clase 2 en las variables ANOREXIA, ASCITES y VARICES. Esto podría señalar que estas características están presentes de manera diferencial en los pacientes y podrían ser variables interesantes que influyan en su supervivencia.

No existe una tendencia para establecer una relación lineal perfecta entre las variables presentes en los gráficos, aunque en algunas de ellas podemos observar una tendencia a una interacción (SGOT y ALK PHOSPHATE, SGOT y BILIRUBIN, PROTIME y ALBUMIN, BILIRUBIN y ALK PHOSPHATE)

Luego, podemos analizar la relación entre nuestras variables categóricas y nuestras variables numéricas (ver figura 24).

La importancia de realizar un análisis de correlación en nuestro conjunto de datos radica en el hecho de que las variables altamente correlacionadas pueden dañar algunos modelos o, en otros casos, podrían proporcionar poca información adicional y considerarlas puede ser computacionalmente costoso sin ningún beneficio real. Además, saber si nuestras

variables muestran una relación lineal puede ayudarnos a elegir qué algoritmo de aprendizaje automático es más adecuado para nuestros datos.

Para finalizar el predecir la supervivencia a la hepatitis

1	instancias	detalle
2	155	filas del conjunto de datos
3	20	atributos (columnas) del conjunto de datos
4	2	valores de clase

Cuadro 1: Características disponibles

Se puede observar que ALK PHOSPHATE, BILIRUBIN, MALAISE, ASCITES, son algunas de las variables más importantes para nuestro modelo.

En este laboratorio, se desarrolló un análisis exploratorio de los datos. Adicionalmente se entregó una vista preliminar sobre la calidad de datos de la muestra y hasta puede determinar la continuidad de nuestro análisis.

Podemos observar en el mapa de calor que algunas de las variables muestran un coeficiente de correlación cercano a 0,8 o -0,4, pero la mayoría de ellas presentan un coeficiente de correlación muy bajo. Entonces podemos concluir que no existe una fuerte correlación lineal entre nuestras variables.

Finalmente todas nuestras experiencias fueron desarrolladas bajo el entorno colab de Google y están disponibles en la siguiente dirección web:

https://colab.research.google.com/drive/1equedkr$_x mNJhog4GL0CCwDkTHLIeDth?usp =$ *sharing*.

Bibliografía

[1] abc.es (2014). Se triplican los casos de hepatitis por consumo de anabolizantes ilegales. https://www.abc.es/salud/noticias/20140222/abci-hepatitis-anabolizantes-201402212030.html.

[2] Beatriz Voltas-Arribas, Juan-Carlos Ferrer-García, C. S.-J. C. M.-A. P. S.-R. L. G.-B. (2018). Anorexia nerviosa como causa de fallo hepático agudo. a propósito de un caso. http://scielo.isciii.es/scielo.php?script=sci_arttextpid=S0212-16112018000100245.

[3] Carolina Althausen K., L. M. (2018). ¿cómo enfrentar un paciente con alteración de las pruebas hepáticas? http://medicinafamiliar.uc.cl/html/articulos/155.html.

[4] Cestnik, B. (1983). Hepatitis domain. https://archive.ics.uci.edu/ml/datasets/hepatitis.

[Concepció Bartres Viñas] Concepció Bartres Viñas, Sergio Rodríguez Tajes, X. F. B.-Z. M. M. Diagnóstico de la hepatitis vírica.

[6] Dugdale, D. C. (2019a). Examen de bilirrubina en sangre. https://medlineplus.gov/spanish/ency/article/003479.htm.

[7] Dugdale, D. C. (2019b). Examen de fosfatasa alcalina en sangre. https://medlineplus.gov/spanish/ency/article/003470.htm.

[8] Dugdale, D. C. (2019c). Prueba de albúmina en la sangre. https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-albumina-en-la-sangre/.

[9] Dugdale, D. C. (2019d). Prueba de sangre de aspartato aminotransferasa. https://medlineplus.gov/spanish/ency/article/003472.htm.

[10] Dugdale, D. C. (2019e). Tiempo de protrombina (tp). https://medlineplus.gov/spanish/ency/article/003652.htm.

[11] eldiario.es (2019). Tipos de hepatitis los cinco conocidos y como se produce la infección. https://www.eldiario.es/consumoclaro/cuidarse/tipos-hepatitis-como-se-contagia_1_1475325.html.

[12] elmostrador.cl (2020). Se estima que 9 de cada 10 personas conviven con hepatitis virales y no lo saben. https://www.elmostrador.cl/agenda-pais/2020/07/28/se-estima-que-9-de-cada-10-personas-conviven-con-hepatitis-virales-y-no-lo-saben/.

[13] Enfermeria21.com (2018). Hepatitis: síntomas, tratamiento y todo lo que debes saber. https://www.enfermeria21.com/diario-dicen/hepatitis-sintomas-tratamiento-y-todo-lo-que-debes-saber-DDIMPORT-057824/.

[14] Estrada, E. (2020). La fatiga como síntoma de enfermedad hepática crónica: nuevos conocimientos y enfoques terapéuticos. https://asscat-hepatitis.org/la-fatiga-como-sintoma-de-enfermedad-hepatica-cronica-nuevos-conocimientos-y-enfoques-terapeuticos/.

[15] fesemi.org (2020). Cirrosis hepática. https://www.fesemi.org/informacion-pacientes/conozca-mejor-su-enfermedad/cirrosis-hepatica.

[16] Juan Fernández-Somoza, Isidro Rodríguez, S. T. M. B.-F. G. E. B. A. G.-Q. (2014). Precisión diagnóstica de las arañas vasculares para detectar enfermedad hepática en alcohólicos. https://galiciaclinica.info/PDF/26/587.pdf.

[17] Mar Noguerol Álvarez, C. R. M. (2015). Esplenomegalia. https://amf-semfyc.com/web/article_ver.php?id=1493.

[18] Mayo-Clinic (2018). Análisis de bilirrubina. https://www.mayoclinic.org/es-es/tests-procedures/bilirubin/about/pac-20393041.

[19] Mayo-Clinic (2020). Agrandamiento del hígado.

https://www.mayoclinic.org/es-es/diseases-conditions/enlarged-liver/symptoms-causes/syc-20372167.

[20] MedlinePlus (2018a). Ascitis.
https://medlineplus.gov/spanish/ency/article/000286.htm.

[21] MedlinePlus (2018b). Fosfatasa alcalina.
https://medlineplus.gov/spanish/pruebas-de-laboratorio/fosfatasa-alcalina/.

[22] MINSAL (2015). Manejo y tratamiento de la infección crónica por el virus de la hepati c (vhc).
https://www.minsal.cl/wp-content/uploads/2016/04/GUIA-VHC.-2015-Editada.pdf.

[23] Villena, E. Z. (2007). Várices esofagogástricas.
http://www.scielo.org.pe/scielo.php?script=sci_arttextpid=S1728-59172007000100011.