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

Laboratorio 1

Integrantes:	Gustavo Aguilar Morita
	Sandra Hernández Yamunaqué
Curso:	Inteligencia Computacional
Programa:	Magíster en Ingeniería Informática
Profesor Cátedra:	Max Chacón Pacheco
Profesor Laboratorio:	Héctor Rojas Pescio

4 de Mayo de 2021

Tabla de contenidos

1. Introducción

La hepatitis en su definición etimológica, viene del antiguo griego donde cuyas raíces *hepar* significa hígado e *itis* significa inflamación, por lo que se comprende como una lesión al hígado con inflamación de células hepáticas (1). El hígado es un órgano de color marrón rojizo oscuro con forma de triángulo que pesa alrededor de 1,36kg, situada en la parte superior derecha de la cavidad abdominal, debajo del diafragma y por encima del estómago, el riñón derecho y los intestinos. Sus principales funciones son: creación de proteínas y otras sustancias, eliminar productos de desecho y toxinas de la sangre, metabolizar fármacos, almacenar energía, secreción de bilis, endocrinas, entre otras (2)(3).

La hepatitis puede ser causadas por bacterias, virus, medicamentos, alcohol, toxinas, entre otras . Esta enfermedad se puede clasificar como una infección aguda o crónica . La infección aguda, es más dolorosa para los pacientes, pero tiene una duración corta (o limitada) de alrededor de no más 1 o 2 meses. La crónica, puede tener una duración superior a los 6 meses. Existen diferentes tipos de hepatitis virales, como la A y la E consideradas como agudas, y las hepatitis de tipo B, C y D como crónicas, y dentro de este grupo, también se encuentra la hepatitis alcohólica, causadas por el consumo excesivo de alcohol lo que conlleva a una cirrosis severa y a un daño de las células parenquimatosas (4).

La hepatitis C es un virus de transmisión sanguínea que infecta de forma crónica a 160-180 millones de personas en todo el mundo. La infección podría terminar con una cirrosis o carcinoma hepatocelular en aproximadamente 20 a 33 % y 1 a 5 % de los pacientes, respectivamente y se estima que más de 350.000 personas mueren cada año a causa de enfermedades hepáticas relacionadas con la hepatitis C, en la figura 1 se observa diferentes biopsias del tejido hepático, realizadas a un grupo de pacientes que fueron diagnosticado con el virus y que generaron diferentes niveles de fibrosis, y cirrosis (cuadro D) (5).

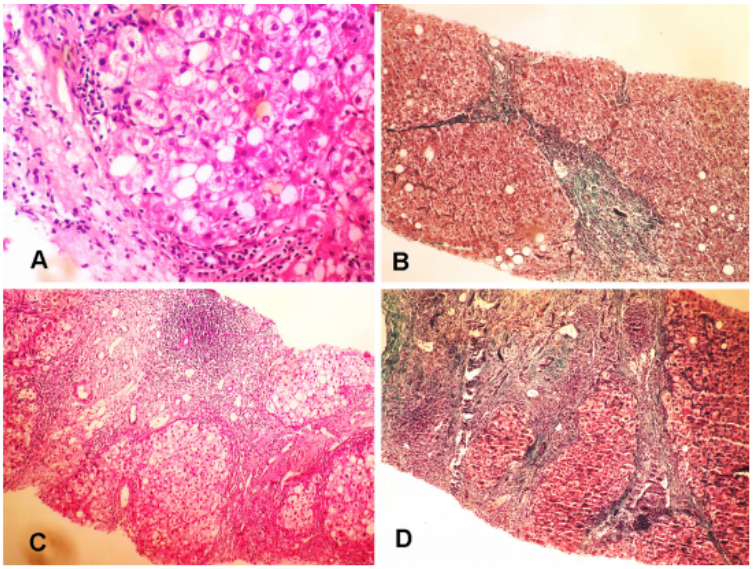

Figura 1: Niveles de fibrosis, en una biopsia de tejido hepático en pacientes diagnosticados con Hepatitis C. A: fibrosis insignificante, BC: fibrosis significante, D: cirrosis

1.1. Descripción del problema

El *dataset* contiene un grupo de 155 pacientes con hepatitis aguda y crónica, de las cuales 33 fallecieron y 122 sobrevivieron, además cuenta con 19 atributos que describen a cada uno de estos, tales como: la edad, el sexo, presencia de fatiga, anorexia, esteroides, varices, bilirrubina y entre otras mediciones bioquímicas, que en las secciones posteriores de este documento serán detalladas. El foco del problema será encontrar una metodología apropiada en conjunto con los datos, la posibilidad de predecir si el paciente tiene posibilidades de sobrevivir (6).

1.2. Objetivos

1.2.1. Objetivo General

Comprender el contexto de estudio en cuanto a la hepatitis e interpretar los datos recogidos (7) mediante un análisis exploratorio.

1.2.2. Objetivos Específicos

1. Describir de forma detallada los atributos, clases y valores del *dataset*.
2. Ejecutar técnicas de preprocesado para trabajar el *dataset*.
3. Analizar el *dataset* con herramientas de estadística descriptiva e inferencial.

1.3. Alcances

Los autores del presente documento no poseen conocimientos respecto a la hepatitis, sin embargo, en la investigación de casos de estudio, se logra identificar que hay gran variedad de documentos que utilizan el *dataset* en diferentes niveles de tratamiento y así como también material científico respecto a la morfología de la enfermedad.

1.4. Metodología y herramientas utilizadas

1.4.1. Metodología

Para la realización de este laboratorio se utiliza un mezcla de estrategias de investigación, que aportarán a una mejor comprensión de los análisis propuestos (8).

1. Experimental, que es propia de la programación del curso.
2. Casos de estudio, que permite obtener una amplia comprensión del contexto de la investigación y los procesos involucrados.
3. Exploratorio, descriptivo e inferencial, cuyo foco es la comprensión del problema, entorno al *dataset*.

1.4.2. Herramientas de desarrollo

La base de datos de hepatitis será trabajada en un ambiente cloud-base llamada Kaggle, subdiaria de Google LLC, permitiendo trabajar colaborativamente de forma eficiente y con versiones controladas de los avances del equipo. El Kernel de Kaggle, permite trabajar con lenguajes de Python y R, y además integra markdown para una mejor comodidad en la documentación.

2. Desarrollo

2.1. Descripción Base de Datos

El dataset utilizado para el desarrollo de la presente experiencia se obtuvo desde el repositorio de base de datos de la University of California, Irvine. El cual es un repositorio público que cuenta con más de 500 conjuntos de datos para el servicio de la comunidad científica[1]. En particular el dataset utilizado fue donado por Peter Turney, y corresponde a datos de pacientes con hepatitis aguda o crónica, el dataset cuenta con 155 muestras con 20 atributos por caso, entre los cuales se encuentran tipos de datos numéricos (6 atributos) y categóricos (14 atributos), finalmente se tiene una etiqueta de clase, la cual predice si el paciente con hepatitis vivirá o no.

2.2. Descripción de clases y variables

Variable	Tipo	Descripción
Clase	Clase	Variable que tiene dos posibles valores: DIE, para aquellos pacientes que no sobrevivirán a la enfermedad (Hepatitis) y LIVE para los sobrevivientes
AGE	Numérico	Variable que representar la edad de los pacientes, esta variable toma valores enteros entre 10 a 80
SEX	Categórica	Esta variable representa el sexo del paciente siendo male, para el sexo masculino y female para el sexo femenino. El dataset representan estas categorías como 1 y 2 respectivamente
STEROID	Categórica	Variable que representa la presencia de esteroides en el paciente, los valores posibles son NO y YES. Estos valores están representados en el dataset como 1 y 2 respectivamente.
ANTIVIRALS	Categórica	Representa la presencia de antivirales en el paciente, los valores posibles son NO y YES. Estos valores están representados en el dataset como 1 y 2 respectivamente.

FATIGUE	Categórica	Representa la presencia de síntomas de fatiga en el paciente, los valores posibles son NO y YES. Estos valores están representados en el dataset como 1 y 2 respectivamente.
MALAISE	Categórica	Representa la presencia de síntomas de malestar en el paciente, los valores posibles son NO y YES. Estos valores están representados en el dataset como 1 y 2 respectivamente.
ANOREXIA	Categórica	Representa el diagnóstico de anorexia en el paciente, los valores posibles son NO y YES, es un tipo de dato categórico. Estos valores están representados como 1 y 2 respectivamente.
LIVER BIG	Categórica	Representa la presencia de un hígado grande en el paciente, los valores posibles son NO y YES. Estos valores inicialmente están representados como 1 y 2 respectivamente.
LIVER FIRM	Categórica	Representa la presencia de durezas en el hígado en el paciente, los valores posibles son NO y YES. Estos valores están representados como 1 y 2 respectivamente.
SPLEEN PALPABLE	Categórico	Indica si el bazo del paciente es o no palpable, los valores posibles son NO y YES. Estos valores están representados como 1 y 2 respectivamente.
SPIDERS	Categórica	Representa la presencia de arañas vasculares los valores posibles son NO y YES, es un tipo de dato categórico. Valores que están representados como 1 y 2 respectivamente.
ASCITES	Categórica	Representa la presencia de acumulación de líquido seroso en la zona abdominal, los valores posibles son NO y YES. Valores que están representados como 1 y 2 respectivamente.
VARICES	Categórica	Representa la presencia de varices, los valores posibles son NO y YES. Estos valores están representados como 1 y 2 respectivamente.

BILIRUBIN	Numérico	Representa el nivel de bilirrubina en el paciente, los valores de los datos van en el rango de (0.39-4.00), es una variable continua. Cabe mencionar que los valores normales de bilirrubina van de 0.1 a 1.2 mg/dLm (9)
ALK PHOSPHATE	Numérico	Representa el nivel de fosfatasa alcalina presente en el paciente, los valores posibles van en el rango de (33-250), es una variable discreta. Se considera normal valores entre los 44 a 147 unidades internacionales por litro (UI/L)
SGOT	Numérico	Representa el nivel de aspartato presente en el paciente, los valores posibles van en el rango de (13-500), variable discreta. El rango considerado normal es de 8 a 33 U/L.
ALBUMIN	Numérico	Representa el nivel de albúmina en el paciente, los valores posibles van en el rango de (2.1-6.0), es una variable continua, siendo el rango normal entre 3.4 a 5.4 g/dL
PROTIME	Numérico	Representa la tendencia de la sangre a coagularse, los valores posibles van en el rango de (10-90), variable discreta.
HISTOLOGY	Categórica	Representa si se realizó un estudio histológico del tejido hepático del paciente, los valores posibles son NO y YES. Estos valores están representados como 1 y 2 respectivamente.

3. Análisis de la base de datos

3.1. Análisis estadístico e inferencial

3.1.1. Análisis de las variables de estudio

Es necesario antes de realizar los diferentes procedimientos, estudiar el comportamiento de las variables presentes en el dataset, como se menciona en el capitulo anterior, contamos con datos numéricos como categóricos por lo que para cada no de ellos se utilizo diagramas de cajas y bigotes e histogramas respectivamente para su posterior análisis. A continuación se muestran los diferentes gráficos generados en conjunto con un breve análisis de ambos tipos de variables presentes en el estudio (numéricas y categóricas). Cabe mencionar que se ha escogido el diagrama de cajas dado que nos permite resumir las características principales de los datos como su dispersión, asimetría, identificar valores atípicos, entre otros.

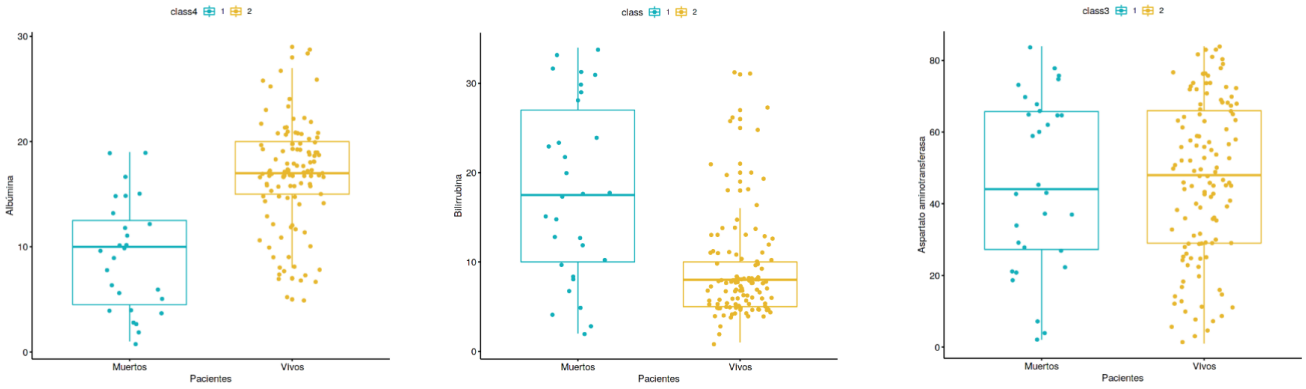

Figura 2: Diagrama de cajas de las variables numéricas, Albumina, Bilirrubina y Aspartato(SGOT) de acuerdo a los pacientes que sobrevivieron y los que no a la hepatitis.

En la Figura (2) se puede observar como los dos primeros diagramas de caja presentan una gran diferencia de acuerdo a los valores promedios entre los pacientes de hepatitis que sobrevivieron a la enfermedad versus a los que fallecieron, así como a la distribución de los valores centrales entre cada uno de ellos, luego podemos decir que las variables albumina y bilirrubina que se observan en el primer y segundo gráfico respectivamente podrían ser consideras buenas variables predictoras para decir si un paciente tiene o no probabilidades

de sobrevivir a la hepatitis. Luego para el tercer gráfico el cual representa la variable Aspartato(SGOT) si bien se puede apreciar que el rango de valores para los pacientes con hepatitis es superior a los rangos normales (8-30 U/L) no difiere entre los pacientes que sobrevivieron y lo que fallecieron a la hepatitis, por lo que se puede considerar un dato poco concluyente para hacer algún tipo de predicción con respecto a si el paciente superará o no la enfermedad.

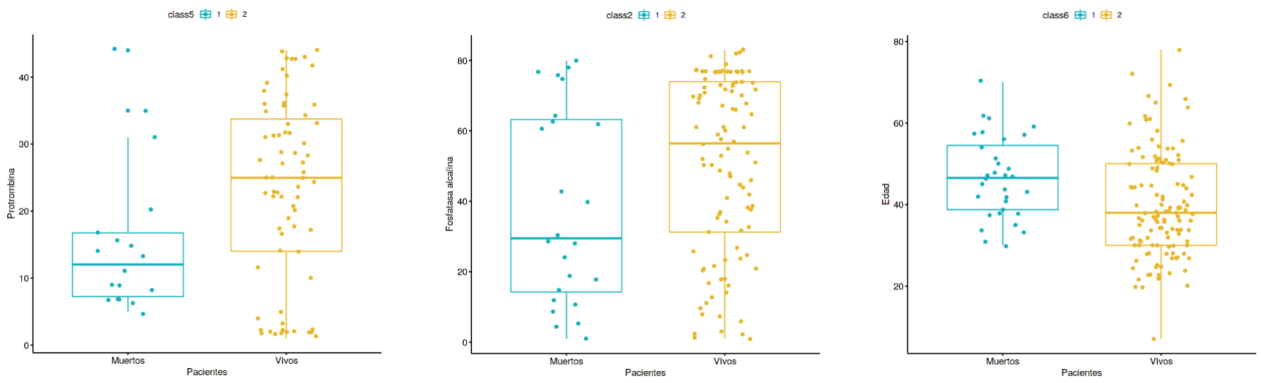

Figura 3: Diagrama de cajas de las variables numéricas, Prothrombin Time(PROTIME), fosfatasa alcalina(ALK PHOSPHATE) y edad (AGE) de acuerdo a los pacientes que sobrevivieron y los que no a la hepatitis.

La Figura (3) muestra las ultimas tres variables numéricas presentes en el estudio, las cuales son, Prothrombin Time(PROTIME), fosfatasa alcalina(ALK PHOSPHATE) y edad respectivamente, siendo Prothrombin Time(PROTIME) la que destaca entre los gráficos, pues a simple vista se puede apreciar que exite una diferencia entre los pacientes que sobrevivieron versus a los que fallecieron. El gráfico muestra que más del 50 % de los pacientes que murieron se ubica en un rango menor a los pacientes que sobrevivieron a la hepatitis, por lo que esta variable también podría tener algo que decir a la hora de predecir la sobrevivencia de un paciente con hepatitis.

Cabe señalar que las variables a considerar como predictorias o que mejor describen el modelo serán escogidas luego de hacer los estudiós de correlación y los análisis pertinentes según sea el caso y los cuales se encuentran contemplados en desarrollo de este informe.

A continuación se muestran una serie de imágenes correspondientes a las variables categóricas del modelo. Cada una ellas se explica al pie de la figura de manera de reducir la

extinción del documento.

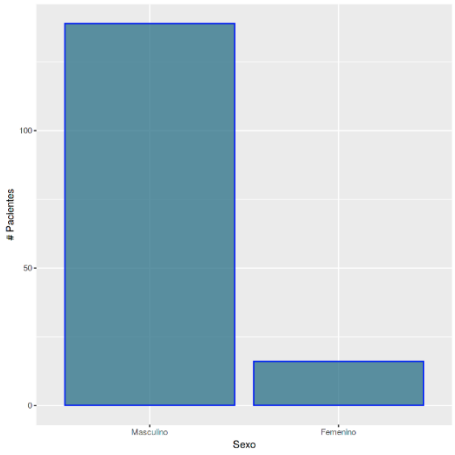

Figura 4: La figura muestra la frecuencia del sexo de los pacientes que se consideraron para el estudio y donde se puede notar que la presencia masculina predomina dentro de los pacientes de estudio

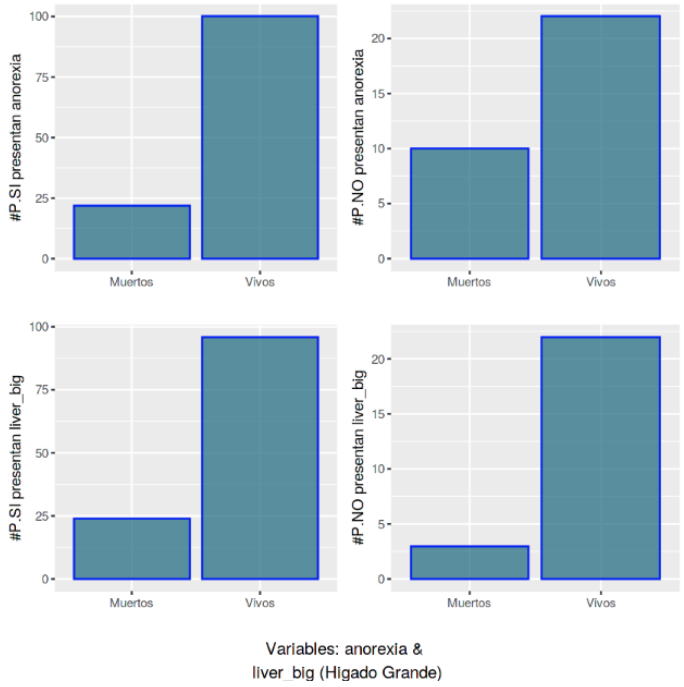

Figura 5: En la figura se muestran las variables categóricas anorexia e hígado grande donde se ve contrastado con la vida del paciente y como su ausencia o presencia se ve reflejado entre los que murieron y sobrevivieron respectivamente.

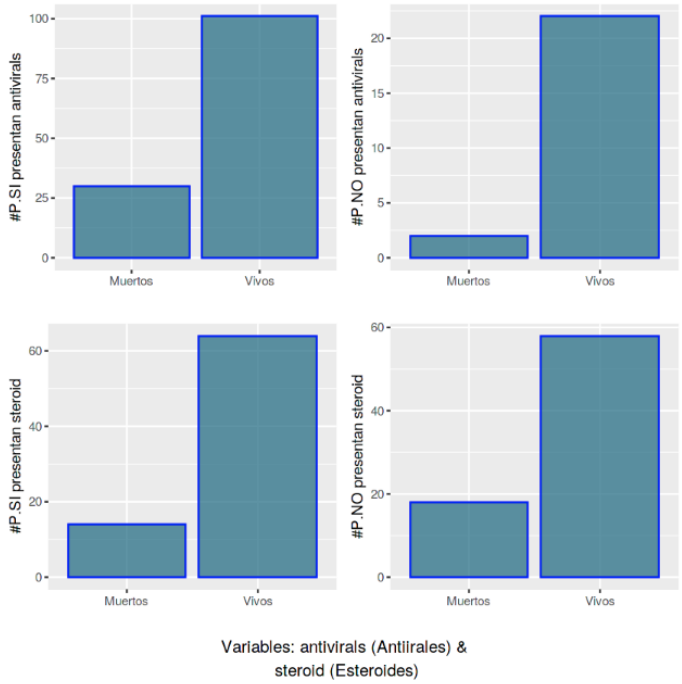

Figura 6: En la figura se muestra dos variables categóricas, en la primera fila se observan los pacientes con presencia de antivirales y los que no la presentan en conjunto a la frecuencia de los pacientes que murieron y vivieron de igual manera se puede observar para la presencia o ausencia de esteroides en los pacientes.

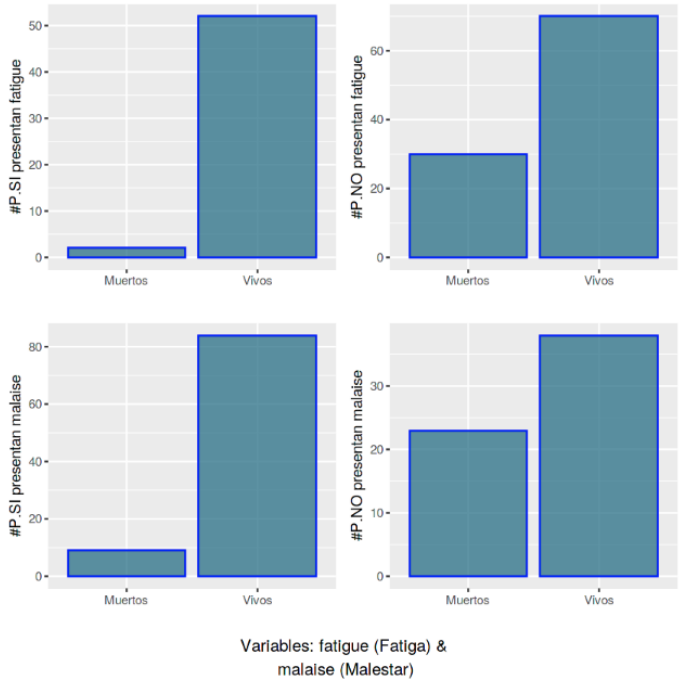

Figura 7: En la figura se muestra dos variables categóricas, fatigue y malaise comparadas con la sobrevivencia del paciente.

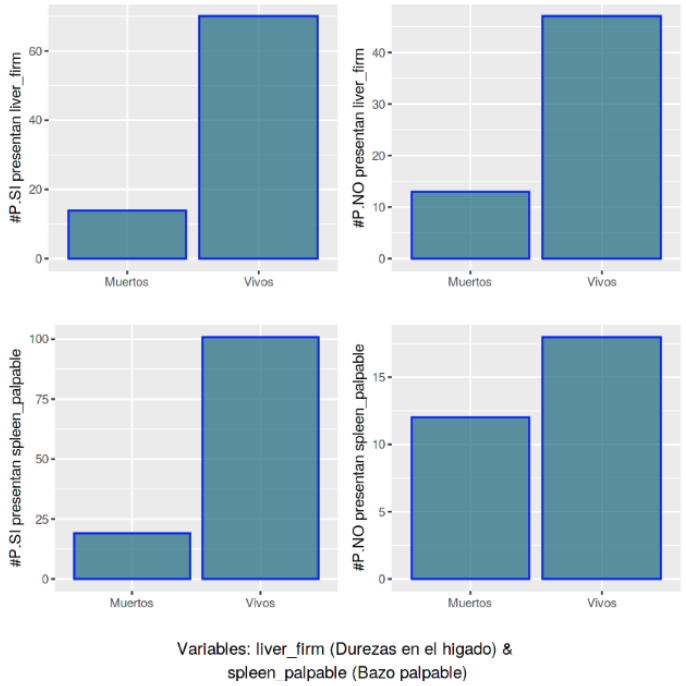

Figura 8: En la figura se muestra dos variables categóricas, liver-firm y spleen palpable comparadas con sobrevivencia del paciente.

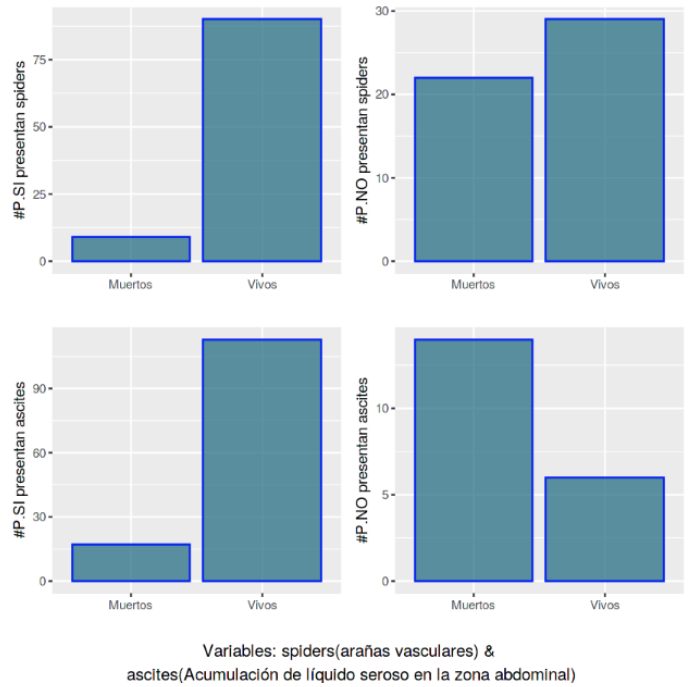

Figura 9: En la figura se muestra dos variables categóricas, spiders y ascites comparadas con la sobrevivencia del paciente.

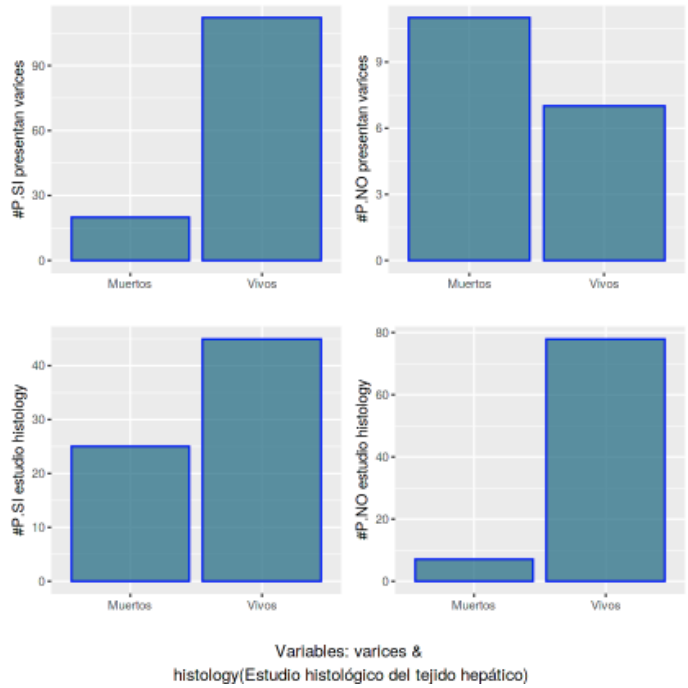

Figura 10: En la figura se muestra dos variables categóricas, varices y histology comparadas con la sobrevivencia del paciente.

3.1.2. Preprocesados relevantes

La base de datos cuenta con diversas proporciones de datos no encontrado (7), por lo que se aplica Multivariate Imputation by Chained Equations (MICE) para imputar datos perdidos.
Para ello se consideran 3 grupos acorde a la naturaleza de las variables y sus datos perdidos:

1. Grupo 1: variables que no tienen datos vacios y que fueron descartados de MICE, correspondientes a: *class*, *age*, *sex*, *antivirals*, *histology*.

2. Grupo 2: aquellos atributos categóricos (factor) que tienen dos niveles o tipos de respuesta y que se le aplicó una regresión logística.

3. Grupo 3: aquellos atributos categóricos (factor) que tienen 3 o más níveles o tipos de respuesta, y que se les aplicó regresión logística politómica.

Posteriormente se realiza una conversión de todos los atributos (a excepción del tipo class que se excluye del análsis de PCA) a tipo numérico o entero según el caso, para poder trabajar los datos en PCA.

3.1.3. Análisis de componentes principales (PCA)

Antes de estudiar el PCA, es importante analizar las correlaciones de los atributos con el fin de detectar multicolinealidad.
Para tener un análisis preliminar, de qué atributos contribuyen a la variable objetivo (class: vivo-muerto), se utilizan los coeficientes de correlación de *Pearson*, además considerando el valor absoluto de las correlaciones, se obtiene la figura (11), donde se puede destacar que existe una gran proporción de atributos que pudiesen contribuir.

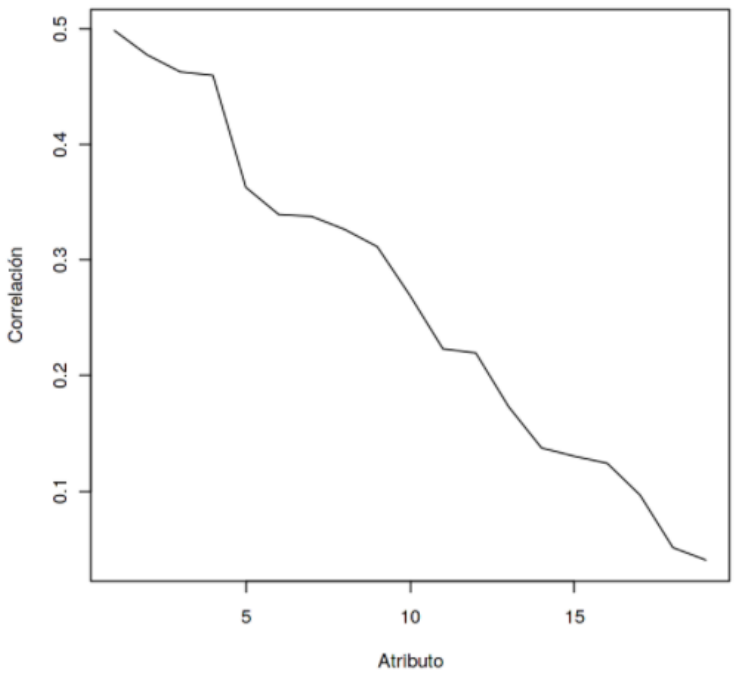

Figura 11: Correlación absoluta entre variable objetivo y los atributos.

Por otra parte, si nos interesa aquellas correlaciones descritas anteriormente y que sean superior a 0.2, tenemos la siguiente matriz de correlaciones representadas en la figura (12). Notamos que hay una serie de variables medianamente correlacionadas, y que eventualmente podría incurrir a una multicolinealidad respecto a la predicción de la muerte del paciente. Se observa por ejemplo que *malaise* con $fatigue$ o bien *ascites* con *albumin* presentan una correlación significante (0.6 y 0.5 respectivamente), por lo que un análisis de componentes principales es sugerido de tal forma de reducir las dimensiones del problema.

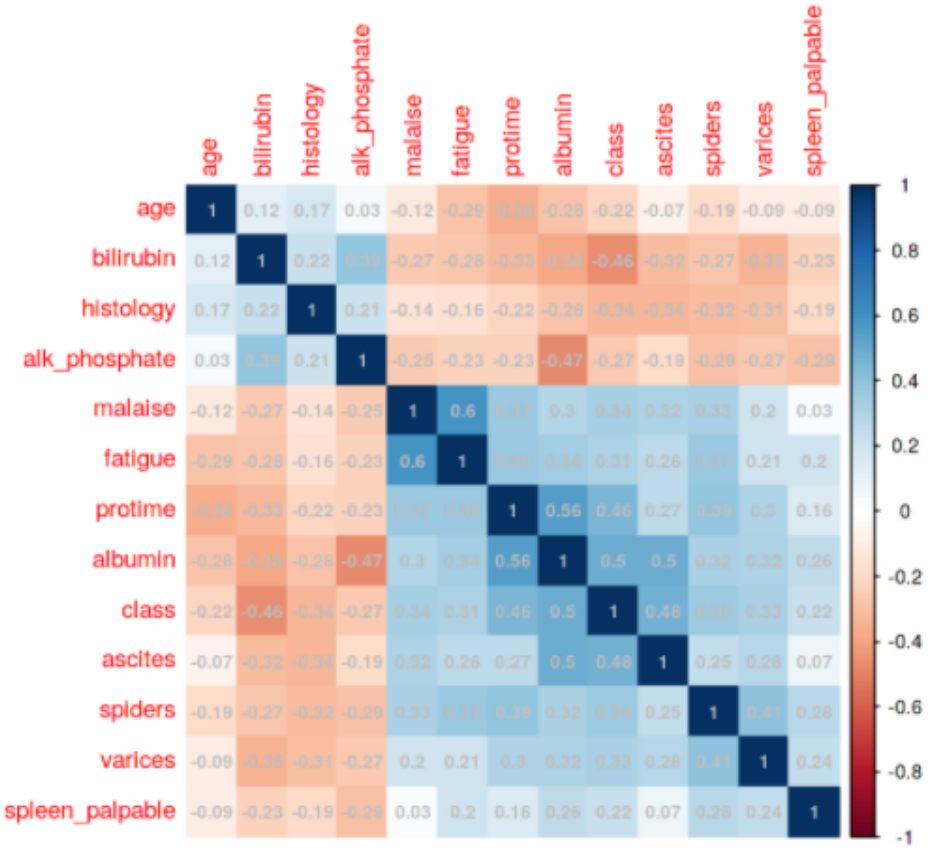

Figura 12: Matriz de correlaciones entre los atributos, filtrados por la correlación entre la variable objetivo y los atributos $> 0{,}2$.

Al aplicar las herramientas de R, se tiene la siguiente figura (13) que muestra los porcentajes de importancia para cada componente. Vemos que PC1 explica un 24.03 % de la varianza total y el resto de los componentes aportan valores menores al 10 %. Así PC1 y PC2 explican un total del 33 % aproximadamente de la varianza total.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.1369 1.31610 1.26091 1.16791 1.07799 1.02999 0.99050
Proportion of Variance 0.2403 0.09116 0.08368 0.07179 0.06116 0.05584 0.05164
Cumulative Proportion  0.2403 0.33149 0.41517 0.48696 0.54812 0.60396 0.65559
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.93621 0.88633 0.87480 0.83399 0.79578 0.76448 0.74170
Proportion of Variance 0.04613 0.04135 0.04028 0.03661 0.03333 0.03076 0.02895
Cumulative Proportion  0.70172 0.74307 0.78335 0.81995 0.85328 0.88404 0.91300
                          PC15    PC16    PC17    PC18   PC19
Standard deviation     0.64756 0.61853 0.58888 0.53468 0.4674
Proportion of Variance 0.02207 0.02014 0.01825 0.01505 0.0115
Cumulative Proportion  0.93507 0.95520 0.97345 0.98850 1.0000
```

Figura 13: Importancia de los componente principales, donde PC1 y PC2 alcanzan un total del 33 % aproximadamente.

En la figura (14), podemos notar que las variables como *albumin*, *protime*, $fatigue$ y *malaise* aportan de forma positiva al componente 1. Por otra parte *bilirubin*, *histology* y $alk - phosphate$ su influencia es mas bien negativa respecto al componente 1. También se puede observar que $liver - big$ y $liver - firm$, contribuyen de forma positiva al componente 2.

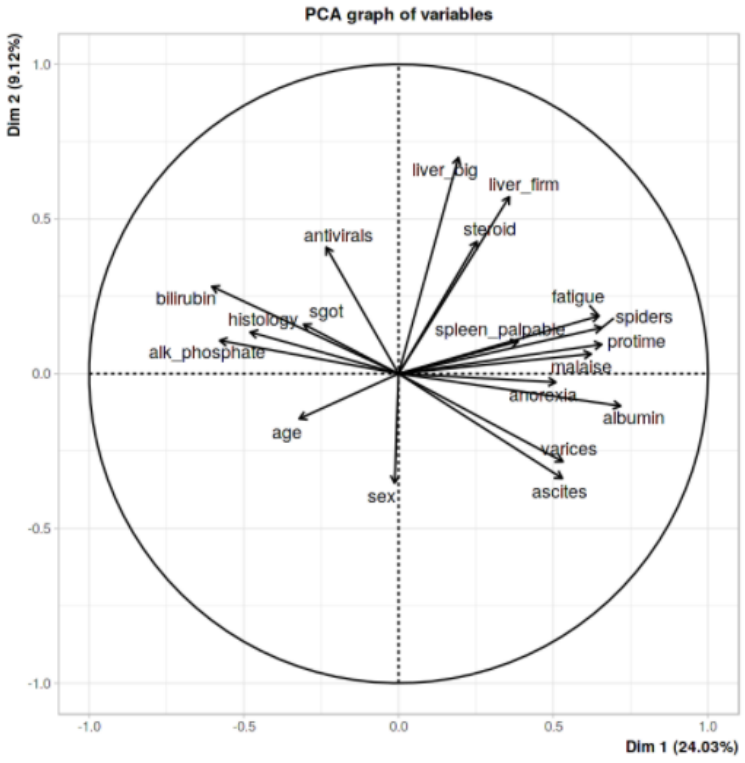

Figura 14: Gráfico de variables de componentes principales.

Se observa en la figura (15) los absolutos de las contribuciones por cada componente, donde podemos notar que para el componente 1, *albumin* tiene un aporte significativo, así como tambien en menor medida como *spiders*, *bilirubin*, *protime*, *fatigue*, *malaise*, entre otros.

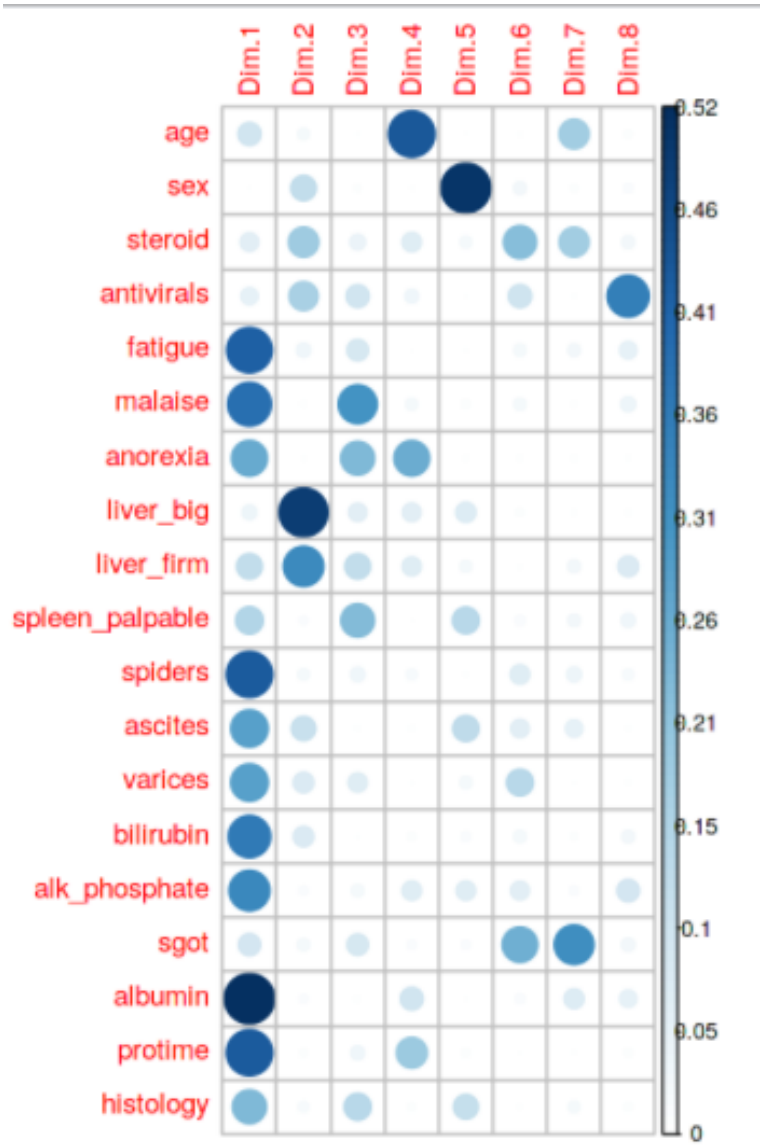

Figura 15: Contribuciones de los atributos por cada componente.

3.1.4. PCA con variables no dicotómicas

Del apartado anterior, surge la necesidad de explorar los atributos no dicotómicos (i.e niveles > 2, para variables de tipo factor), por lo que se considera *age*, *bilirubin*,

$alk - phospate$, $sgot$, $albumin$ y $protime$.

De la figura (16) notamos que el componente 1 alcanza un 40.44 % en cuanto a la explicación de la varianza total, además PC1 y PC2 alcanzan el 59 % aproximadamente. Si descartamos PC5 y PC6, notamos que entre los primeros 4 componentes tendremos una explicación entorno al 84 %.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.5577 1.0473 0.9481 0.7908 0.7711 0.59808
Proportion of Variance 0.4044 0.1828 0.1498 0.1042 0.0991 0.05962
Cumulative Proportion  0.4044 0.5872 0.7370 0.8413 0.9404 1.00000
```

Figura 16: Importancia de los componente principales, de atributos no dicotómicos, donde PC1 y PC2 alcanzan un total del 59 % aproximadamente.

De la figura (17), podemos notar las fuertes contribuciones de $albumin$ y $protime$ de forma negativa al componente 1, en contra parte, $bilirubin$ y $alk - phosphate$, que lo hacen de forma positiva. Notamos ademas que la edad aporta fuertemente al componente 2 de forma negativa. Esto lo complementa de mejor manera la figura (18), donde podemos ver las contribuciones en absoluto.

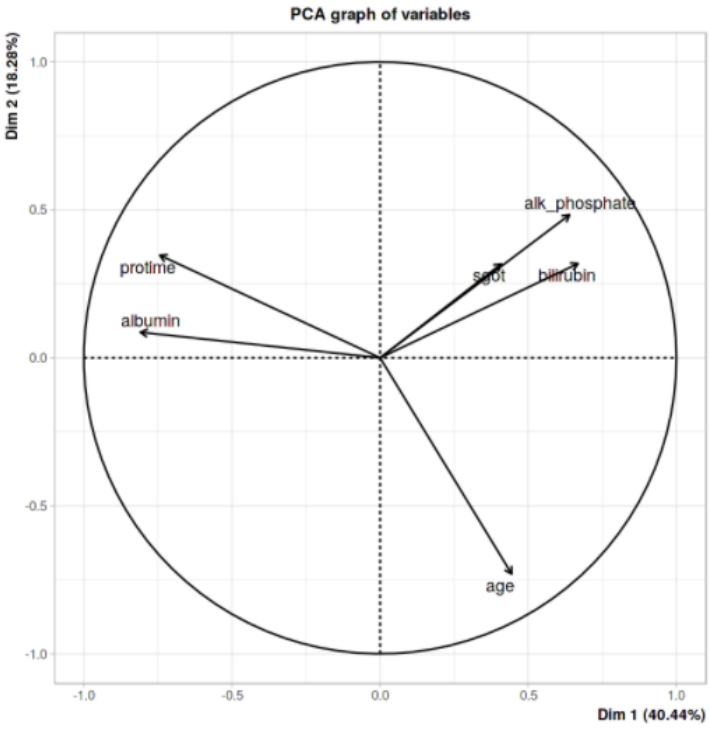

Figura 17: Gráfico de variables de componentes principales, de atributos no dicotómicos.

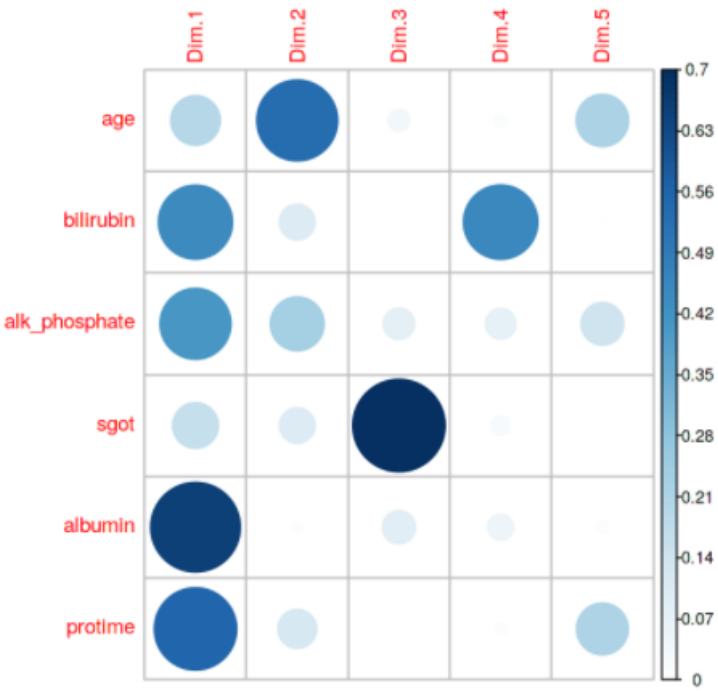

Figura 18: Contribuciones de los atributos no dicotómicos, por cada componente.

3.1.5. Análisis de Factores Múltiples (MFA)

Se considera este análisis para poder contrastarlo con PCA, dado que MFA considera la posibilidad de incorporar variables categóricas y continuas. Para el desarrollo de este análisis se considera particionar la base de datos, acorde a grupos que tengan una cierta relación, ya sea por la naturaleza del problema, por tipo de datos, clasificación sensorial, series de tiempo, u otro tipo de características que los agrupe (10). Para nuestro caso se propuso los siguientes grupos:

1. Descripción del paciente (Grupo 1): *sex* y *age*.

2. Condición del paciente (Grupo 2): *steroid*, *antivirals*, $fatigue$, *malaise*, *anorexia*, *liverbig*, $liverfirm$, *spleenpalpable*, *spiders*, *ascites*, *varices*, *histology*. Correspondientes a variables dicotómicas, y que básicamente responden a la presencia o no de cada uno de estos estados del paciente.

3. Bioquímico (Grupo 3): *bilirubin*, *Alkphosphate*, *sgot*, *albumin*, *protime*. Variables con más niveles (> 30) y/o discretas.

Utilizando esta técnica vemos que de la figura (19) el componente 1 alcanza un nivel de explicación del 22.38 % de la varianza total, alcanzando entorno al 35 % por los dos primeros componentes.

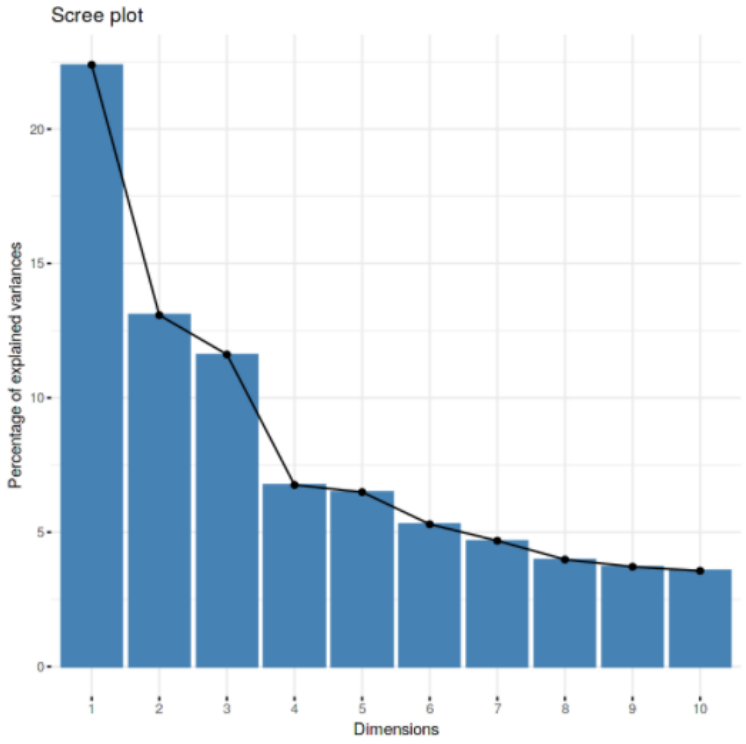

Figura 19: Gráfico de varianzas de los componentes principales utilizando MFA.

Sin embargo, de este análisis de MFA, podemos notar las contribuciones por grupo cada componente, donde se observa que el grupo de Condición y Bioquímico aportan un 40.39 % y 43.38 % respectivamente al componente 1, de acuerdo a la figura (20)

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Descripción	16.21681	90.503447	73.72668	0.9564654	2.909859
Condición	40.39495	7.002394	14.86786	75.5216745	85.522662
Bioquímico	43.38824	2.494159	11.40546	23.5218602	11.567479

Figura 20: Contribuciones de los grupos por cada componente principal utilizando MFA.

Desglosando estos grupos de variables cuantitativas, podemos notar que la *age*, *protime* y *bilirubin*, contribuyen de forma positiva al componente 1, y *sgot* junto a *alk − phosphate* lo hacen fuertemente de forma negativa a este mismo, como se ilustra en la figura (21).

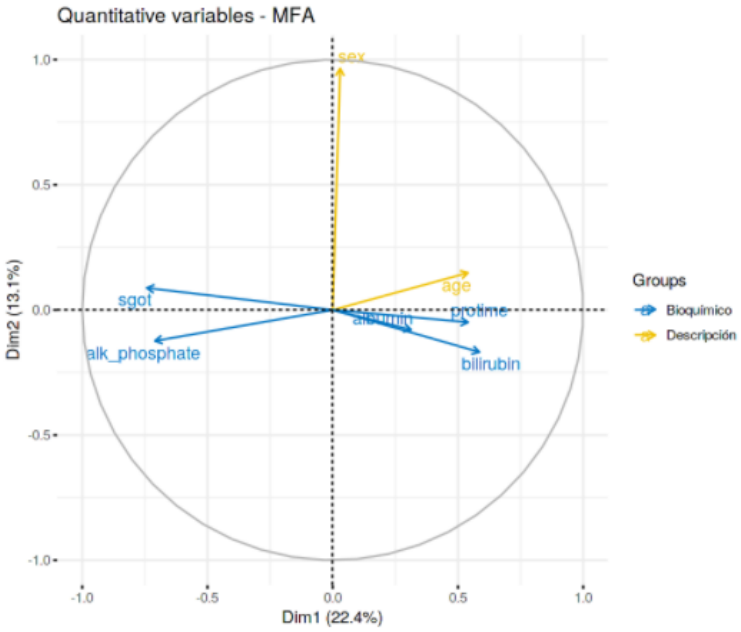

Figura 21: Gráfico de componentes principales, utilizando MFA, de atributos cuantitativos, correspondientes a los grupos Descripción y Bioquímico.

Luego se observan las contribuciones de las variables cualitativas en la figura (22), donde se observa una tendencia en la que las variables de tipo 1 (no presencia de la condición) se agrupan de forma positiva a la componente 1 y por el contrario las de tipo 2 (presencia de la condición), se agrupan de tal forma que contribuyen de forma negativa al mismo componente.

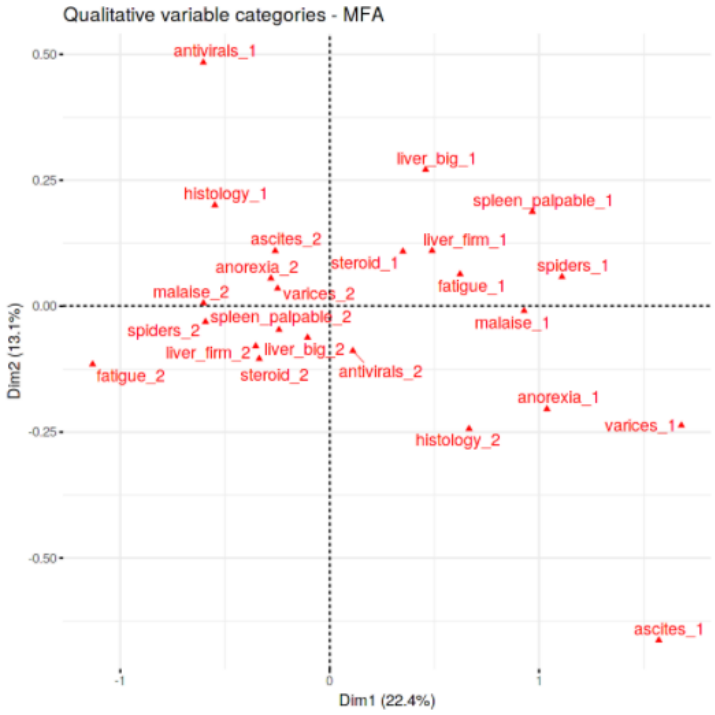

Figura 22: Gráfico de componentes principales utilizando MFA de variables cualitativas o categóricas, correspondiente al grupo Condición.

3.1.6. Test de hipótesis

De acuerdo a los resultados anteriores, vemos que uno los atributos *albumin* que tiene marcada una presencia significativa en los diferentes elementos de PCA, como se puede observar en las figuras (15) y (18), se propone como $H0$ tal que el nivel de albumina no varía en pacientes fallecidos o que sobrevivieron. Para ello se utiliza el test de U de Mann-Whitney, dado que el test t, no cumplen con el requisito de normalidad.

```
        Wilcoxon rank sum test with continuity correction

data:  albuminDeath and albuminLive
W = 638, p-value = 3.758e-09
alternative hypothesis: true location shift is not equal to 0
```

Figura 23: Test de U de Mann-Whitney con $\alpha = 0{,}05$ con $p - value << \alpha$

De la figura (23) vemos que $p - value$ es menos a α, por lo que se rechaza la $H0$ y en consecuencia, la albumina varia significativamente si el paciente ha sobrevivido o ha

fallecido.

Otra variable interesante de observar es la *bilirubin*, donde también se observa en varios elementos presentados anteriormente. Se considera que H0 tal que las medias de la bilirrubinas de los pacientes que fallecieron y sobrevivieron, son iguales. Para este caso, aplicamos test-t con Shapiro-wilk dados en la figura(24), donde se validan las condiciones de normalidad de los datos para ambos grupos de datos y por consiguiente la $H0$ se rechaza, dado que $p - value < \alpha$, según la figura (25). De esto podemos decir que la bilirrubina tiene medias diferentes para pacientes fallecidos y que sobrevivieron con una significancia del 5 %.

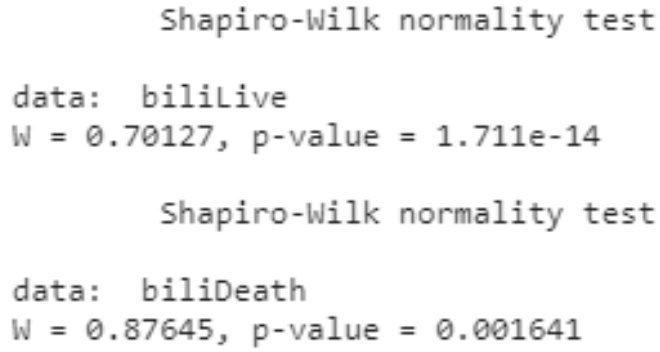

Figura 24: Comprobación de normalidad de los datos con test Shapiro-Wilk con $p - value$ menor que α en ambos casos.

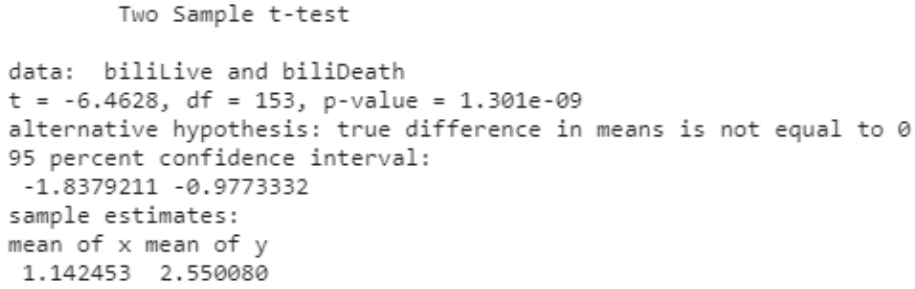

Figura 25: $H0$ se rechaza con Test-t con $p - value$ menor que α

4. Conclusión

La hepatitis es una enfermedad muy presente en todo el mundo, donde existe una amplia gama de documentos científicos entorno a la predicción propuesta en este documento. De acuerdo al análisis estadístico, visualizar cierta multicolinealidad entre los atributos y las correlaciones respecto a la variable objetivo.

Además se pudo identificar una serie de variables que contribuyen a la predicción, incluso se obtuvo que a través de PCA, que los atributos numéricos son bastante significativos y mediante MFA, se logra identificar que el atributo categórico como *ascites* también explica de forma importante al momento reducir la dimensionalidad del problema. Sin embargo, no es concluyente el número de atributos que se puedan reducir.

Finalmente se toman dos test de hipótesis para las contribuciones de la *bilirubin* y *albumin*, donde se obtiene que su variación puede influir sobre si el paciente podría o no sobrevivir ante una hepatitis.

Bibliografía

[1] T.Karthikeyan and P.Thangaraju, *Analysis of Classification Algorithms Applied to Hepatitis Patients*. International Journal of Computer Applications (0975 – 8887), 2013.

[2] J. A. P. Q. y. G. M. A. A. Melina Dennise Medina Gamarra, Francisco Andrés Medina Montoya, *Calidad de vida de en pacientes con trasplante de hígado*. RECIMUNDO, 4(1(Esp), 2020.

[3] O. R.-L. y. R. E. F.-L. Laritza Dayana Potrillé-Rodríguez 1, Maricarmen Prawl-Estévez, *Cambios morfofuncionales del hígado en la cirrosis hepática*. Gaceta Médica Estudiantil, 1(1), 45-56, 2020.

[4] E. K. A. Najla' Fathi Metwally and S. S. Abu-Naser, *Diagnosis of Hepatitis Virus Using Artificial Neural Network*. International Journal of Academic Pedagogical Research (IJAPR), 2018.

[5] B. Elesawy, *Limited reliability of five non-invasive biomarkers in predicting hepatic fibrosis in chronic HCV mono-infected patients opposed to METAVIR scoring*. Pathol. – Res. Pract, 2014.

[6] P. Diaconis and B. Efron, *Computer-Intensive Methods in Statistics*. Scientific American, a division of Nature America, Inc., 1983.

[7] D. Dua and C. Graff, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science*. 2019.

[8] P. L. Mark Saunders and A. Thornhill, *Research Methods for Business Students*. Prentice Hall, 2000.

[9] E. Salaber, "¿te sube la bilirrubina? vigila tu hígado," septiembre 2015.

[10] A. Kassambara, *MFA - Multiple Factor Analysis in R: Essentials, Articles - Principal Component Methods in R: Practical Guide*. 2017.