diciembre, 2023

Informe Trabajo Minería de Datos

Alberto García Martín

Máster en Sistemas Inteligentes
Universidad de Salamanca

Índice

1. Introducción

Este informe documenta el trabajo realizado para la asignatura de Minería de Datos del máster en Sistemas Inteligentes. En este trabajo se ha realizado un estudio sobre un conjunto de datos que recopila distintas características médicas de pacientes con cirrosis. Estos datos fueron recopilados en Estados Unidos por la Clínica Mayo entre 1974 y 1984, aunque fueron donados recientemente, en septiembre de 2023, al repositorio de datos de la Universidad de California en Irvine, bajo el título *Cirrhosis Patient Survival Prediction* [1].

Este conjunto de datos fue creado con el propósito de hacer un estudio de supervivencia de 424 pacientes con colangitis biliar primaria, también conocida como cirrosis biliar primaria, que se sometieron a un ensayo controlado con placebo para probar la eficacia de un tratamiento con penicilamina. De los pacientes iniciales, 312 accedieron a participar en el estudio, proporcionando datos y análisis clínicos exhaustivos. De los 112 pacientes restantes solo se recogieron datos básicos e información sobre su estado de supervivencia, aunque se perdió el contacto con 6 ellos. Por lo que el conjunto de datos incluye información básica de 106 pacientes y datos completos de 312 pacientes.

La cirrosis se produce tras un daño continuado en el hígado, que provoca la formación de tejido cicatricial en el mismo, esta enfermedad suele ser causada por condiciones como el alcoholismo, la hepatitis o la obesidad. Según datos del Ministerio de Sanidad, en España se producen alrededor de 4000 muertes al año por cirrosis [3], siendo la causa de muerte número 12 en España. Por lo tanto, es de gran interés médico poder determinar la eficacia de los tratamientos para esta enfermedad, así como poder predecir la supervivencia de los pacientes con esta enfermedad.

En este trabajo se realizará un estudio de los datos, haciendo un preprocesamiento de los mismos y se aplicarán distintos algoritmos de aprendizaje automático para predecir la supervivencia de los pacientes y entender qué factores influyen en la misma. Para ello se emplearán tanto algoritmos de aprendizaje supervisado como no supervisado, y se compararán los resultados obtenidos con los distintos algoritmos. El objetivo principal es obtener un modelo que permita pronosticar las probabilidades de supervivencia de pacientes con cirrosis a partir de sus datos médicos.

Para el desarrollo del trabajo se ha empleado el lenguaje de programación Python, junto con bibliotecas como Pandas para el procesamiento de los datos, Matplotlib y Seaborn para la visualización de los datos, y Scikit-learn y Mlxtend para la implementación de los métodos de minería de datos.

2. Descripción de los datos

El conjunto de datos con el que se va a trabajar se ha escogido por su gran variedad de características, incluyendo datos numéricos discretos, continuos y categóricos. Contiene 17 características distintas para predecir la supervivencia de los pacientes.

La etiqueta de supervivencia puede tener tres valores distintos: `D`, indica que el paciente murió durante el estudio; `C`, indica un paciente censurado, es decir, que seguía vivo al acabar el estudio; y `CL`, indica que el paciente recibió un trasplante de hígado.

El número total de entradas es de 418, un tamaño adecuado para realizar un análisis de datos, aunque puede que se quede corto para obtener resultados fiables con algunos métodos de minería de datos. Los datos contienen valores perdidos, es decir campos sin ningún valor asignado, caracterizados por el valor `NA`, esto se deberá tener en cuenta a la hora de realizar el preprocesamiento de los datos.

Las columnas del conjunto de datos son las siguientes:

- **ID** (entero): Identificador único del paciente.
- **N_Days** (entero): Número de días desde el registro del paciente en el estudio, hasta la fecha más temprana entre la fecha de muerte del paciente, la fecha de trasplante de hígado o la fecha de finalización del estudio.
- **Status** (categórico): Estado del paciente al finalizar el estudio, puede ser `D` (fallecido), `C` (censurado) o `CL` (censurado por trasplante de hígado).
- **Drug** (categórico): Tipo de medicamento recibido, puede ser `D-penicillamine` (penicilamina) o `Placebo`.
- **Age** (entero): Edad del paciente en días.
- **Sex** (categórico): Sexo del paciente, puede ser `F` (femenino) o `M` (masculino).
- **Ascites** (categórico): Indica si el paciente tiene ascitis, una acumulación de líquido en el abdomen, puede ser `Y` (sí) o `N` (no).
- **Hepatomegaly** (categórico): Indica si el paciente tiene hepatomegalia, un agrandamiento del hígado, puede ser `Y` (sí) o `N` (no).
- **Spiders** (categórico): Indica si el paciente tiene arañas vasculares, una dilatación de los vasos sanguíneos superficiales, puede ser `Y` (sí) o `N` (no).
- **Edema** (categórico): Indica si el paciente tiene edema, una acumulación de líquido en los tejidos, puede ser `N` (sin edema y sin tratamiento diurético para el edema), `S` (con edema y sin tratamiento diurético para el edema, o edema resuelto por tratamiento diurético), `Y` (con edema a pesar del tratamiento diurético).
- **Bilirubin** (real): Nivel de bilirrubina en sangre del paciente, en mg/dl.
- **Cholesterol** (entero): Nivel de colesterol en sangre del paciente, en mg/dl.
- **Albumin** (real): Nivel de albúmina en sangre del paciente, en g/dl.
- **Copper** (entero): Nivel de cobre en orina del paciente, en ug/día.

- **Alk_Phos** (real): Nivel de fosfatasa alcalina en sangre del paciente, en U/L.
- **SGOT** (real): Nivel de transaminasa glutámico-oxalacética en sangre del paciente, en U/ml.
- **Tryglicerides** (entero): Nivel de triglicéridos en sangre del paciente, en mg/dl.
- **Platelets** (entero): Nivel de plaquetas en sangre del paciente, en miles/mm^3.
- **Prothrombin** (real): Tiempo de protrombina del paciente, en segundos.
- **Stage** (categórico): Etapa de la enfermedad del paciente (1, 2, 3 o 4).

La columna `ID` se ignorará en el análisis, ya que no aporta información relevante para el mismo. La columna `N_Days` sí que podría tener interés para realizar un análisis de la esperanza de vida de los pacientes, sobre todo en el caso de los más graves. Sin embargo, este trabajo se centrará en la predicción del estado de supervivencia de los pacientes (columna `Status`), tratando el problema como un problema de clasificación, para el cual las 17 columnas restantes serán las características de entrada.

Analizando la distribución de los datos, se puede observar que el 38.5 % de los pacientes fallecieron durante el estudio, mientras que el 55.5 % de los pacientes fueron censurados y el 6 % restante recibió un trasplante de hígado. Como era esperable, el número de pacientes que recibieron un trasplante de hígado es muy bajo. Además entre los pacientes censurados y los fallecidos, la proporción está desequilibrada. Esta distribución se puede observar en la figura 1.

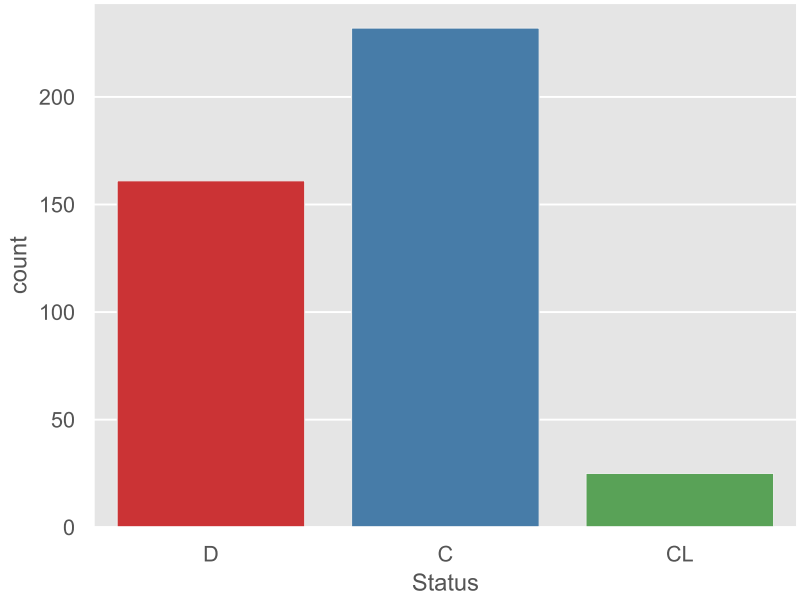

Figura 1: Distribución de los pacientes según su estado de supervivencia.

Continuando con el análisis de distribuciones, para la variable `Drug` el conjunto sí que está balanceado, con un porcentaje del 50 % en ambas clases, esto es de esperar,

ya que el estudio estaba controlado por placebo, por lo que la proporción de pacientes que recibieron penicilamina y los que recibieron placebo debería ser similar. La distribución del resto de características se puede observar en la figura 2.

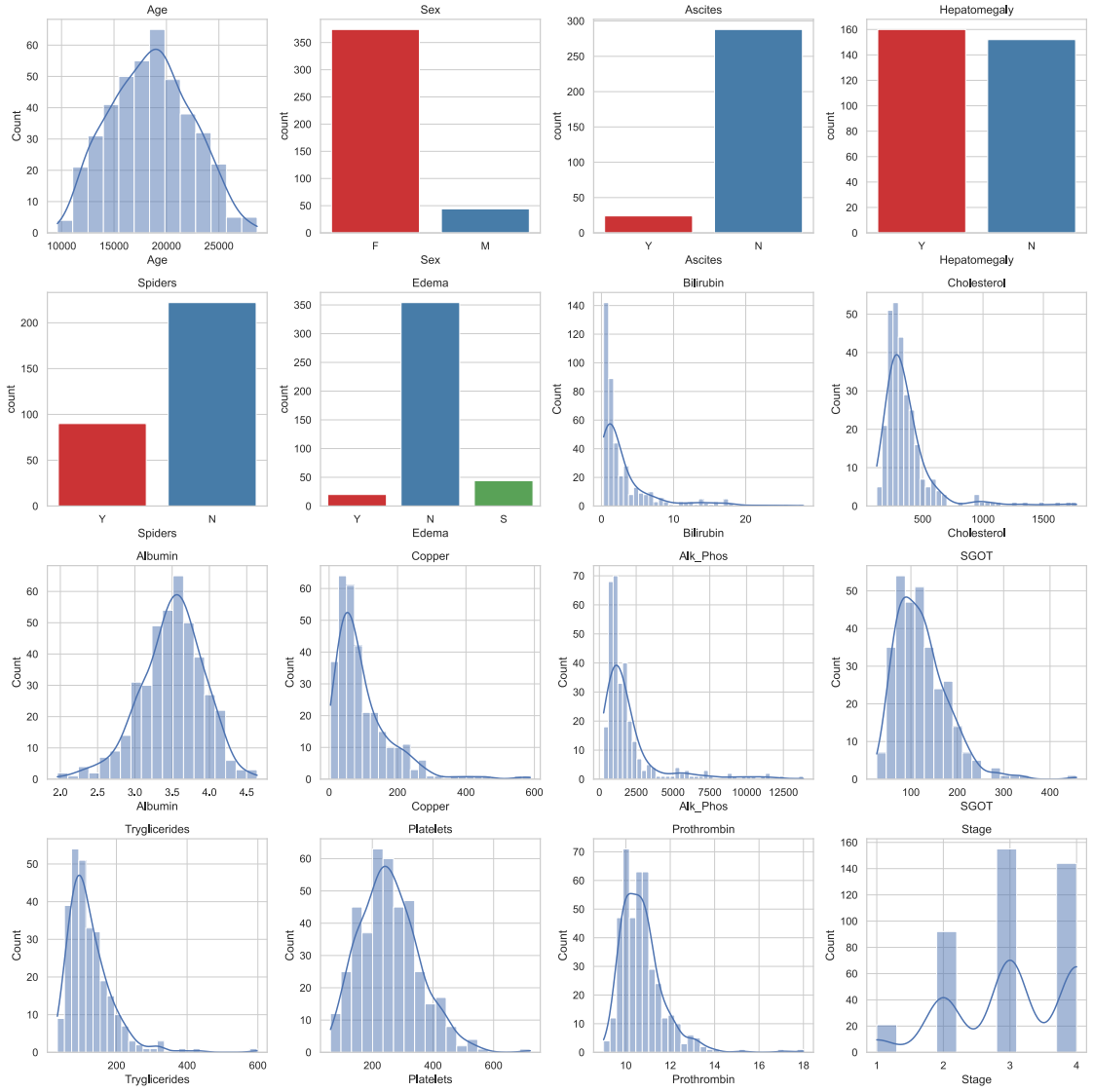

Figura 2: Distribución de las características del conjunto de datos.

La media de edad de los pacientes es de 18533 días, lo que equivale a 50.8 años, con pacientes que van desde los 26 años hasta los 78. Es notable la gran mayoría de pacientes femeninos en el estudio, con un 89.5 % de pacientes femeninos frente a un 10.5 % de pacientes masculinos. Esto puede deberse a que la colangitis biliar primaria es más común en mujeres que en hombres, con una proporción aproximada de 9 mujeres por cada hombre [4].

Analizando la presencia de valores perdidos se puede ver que hay 106 entradas con valores perdidos en la mayoría de las columnas, esto se corresponde con los pacientes que no participaron en el estudio completo. En la tarea de preprocesamiento de los datos habrá que tratar estos valores perdidos.

3. Preprocesamiento de los datos

En esta sección se describirán las tareas de preprocesamiento de los datos que se han realizado para poder aplicar los algoritmos de minería de datos. En primer lugar, se ha eliminado la columna `ID`, ya que no aporta información relevante para el análisis y básicamente se corresponde al número de entrada en el conjunto de datos. También se ha decidido eliminar la columna `N_Days`, ya que este análisis se va a centrar en resolver el problema de clasificación de pacientes en las distintas clases de supervivencia, y no en la predicción de la esperanza de vida de los pacientes, que sería el objetivo de un análisis de regresión.

Posteriormente se han eliminado las entradas que tienen un valor perdido (`NA`) en la columna `Drug`, ya que tienen valores perdidos en la mayoría de las columnas, y no se puede realizar un análisis fiable con ellas, y sustituirlos por valores sintéticos podría alterar los resultados de manera significativa.

El resto de valores perdidos se encuentran en columnas con datos numéricos. Se ha optado por sustituirlos por la mediana de los valores de la columna correspondiente. Debido a que las gráficas de distribución de las características se puede observar que la mayoría de las características tienen una distribución asimétrica, por lo que hay valores atípicos que podrían alterar la media. Por lo tanto, se considera que la mediana es una medida más robusta para sustituir los valores perdidos.

A pesar de que hay valores que se podrían considerar atípicos, no se han eliminado del conjunto de datos, ya que se consideran valores reales asociados a pacientes en una fase muy avanzada de la enfermedad hepática y no un error de medición. Por lo tanto estos valores podrían resultar útiles en la aplicación de los algoritmos de aprendizaje automático.

Para utilizar las variables categóricas en distintos métodos de minería de datos, como el de máquinas de vectores de soporte, es necesario convertir las cadenas de texto en enteros. Se ha aplicado una codificación *one-hot* para convertir cada categoría a un entero distinto. La codificación resultante para cada característica categórica es la siguiente:

- **Status**: `C` $\rightarrow$ 0, `CL` $\rightarrow$ 1, `D` $\rightarrow$ 2.
- **Drug**: `D-penicillamine` $\rightarrow$ 0, `Placebo` $\rightarrow$ 1.
- **Sex**: `F` $\rightarrow$ 0, `M` $\rightarrow$ 1.
- **Ascites**: `N` $\rightarrow$ 0, `Y` $\rightarrow$ 1.
- **Hepatomegaly**: `N` $\rightarrow$ 0, `Y` $\rightarrow$ 1.
- **Spiders**: `N` $\rightarrow$ 0, `Y` $\rightarrow$ 1.
- **Edema**: `N` $\rightarrow$ 0, `S` $\rightarrow$ 1, `Y` $\rightarrow$ 2.

Se ha decidido discretizar la variable `Age` en 4 intervalos, con el objetivo de convertir las variables numéricas en categorías de edad representativas del grupo de pacientes, reduciendo el sesgo hacia ciertas edades. Los intervalos escogidos tienen la misma amplitud, de 13 años, resultando en los siguientes intervalos: [26, 39), con 57 pacientes; [39, 52), con 124 pacientes; [52, 65), con 104 pacientes; y [65, 78], con 27 pacientes.

Para no incluir características irrelevantes en el análisis, se ha realizado un análisis de correlación entre las características y la etiqueta de supervivencia. Como función de evaluación se ha utilizado la prueba chi al cuadrado para medir la correlación, al ser adecuada para problemas de clasificación. Los resultados obtenidos se muestran en la tabla 1. La característica `Drug` es la que tiene una correlación más baja con la etiqueta de supervivencia, siendo la única con un valor menor que 1. Por lo tanto, se ha decidido eliminar esta característica del conjunto de datos. Esto coincide con el estudio original realizado por los autores del conjunto de datos, que determinaron que la penicilamina no tenía ningún efecto en la supervivencia de los pacientes [2].

Característica	**Puntuación**
Drug	0.11404
Albumin	1.52179
Prothrombin	5.14472
Sex	7.90465
Spiders	12.0237
Age	13.5700
Stage	14.3536
Hepatomegaly	17.3710
Ascites	31.0933
Edema	45.2538
Platelets	352.410
Bilirubin	374.972
Tryglicerides	407.266
SGOT	711.359
Cholesterol	1567.98
Copper	3571.65
Alk_Phos	39375.3

Tabla 1: Correlación entre las características y la etiqueta de supervivencia.

Los pacientes que han recibido un trasplante de hígado se han eliminado del conjunto de datos, ya que han recibido un tratamiento clínico considerablemente distinto al resto de pacientes. El objetivo de este análisis es entender los factores que afectan la supervivencia de los pacientes con cirrosis con un tratamiento estándar, por lo que los pacientes que han recibido un trasplante de hígado no son representativos de este grupo. La eliminación de estas entradas es asumible, ya que solo hay 19 pacientes que han recibido un trasplante de hígado, relativamente pocos en comparación con las otras categorías.

Con la eliminación de las entradas con etiqueta `CL` el problema ha pasado a ser un problema de clasificación binaria. Para compensar el desequilibrio de clases, se ha hecho un remuestreo de la clase minoritaria, añadiendo 43 entradas. Además se ha cambiado la etiqueta de la clase correspondiente a los pacientes fallecidos de `2` a `1`.

Tras realizar esta transformación de los datos, el conjunto de datos resultante tiene 337 entradas y 16 características, con una distribución uniforme de las etiquetas de supervivencia. Este conjunto de datos se considera limpio y listo para usarlo en los algoritmos de minería de datos.

4. Aplicación de algoritmos no supervisados

En esta sección se aplicarán distintos algoritmos de aprendizaje no supervisado para realizar un análisis exploratorio de los datos. En primer lugar, se aplicará el algoritmo k-medias para realizar una agrupación de los pacientes en distintos grupos, y posteriormente se aplicará el algoritmo apriori para realizar un análisis de las reglas de asociación entre las características de los pacientes.

4.1. *Clustering* con k-medias

Para determinar el número de grupos en los que se agruparán los pacientes se ha utilizado el método del codo, que consiste en aplicar el algoritmo **k-medias** con un número de grupos creciente, del 1 al 10 en este caso, y calcular la suma de las distancias al cuadrado de cada punto al centroide de su grupo. Posteriormente se representa gráficamente la suma de las distancias al cuadrado en función del número de grupos. El número de grupos óptimo es el punto en el que la suma de las distancias al cuadrado deja de disminuir de manera significativa, formando un punto de inflexión o "codo" en la gráfica. En este caso, el número de grupos seleccionado es 2 (figura 3).

Tras aplicar el algoritmo k-medias con dos grupos, se ha obtenido un grupo con 299 registros y otro con 37, las medias obtenidas para cada característica en cada grupo se pueden observar en la tabla 2. Los valores medios para todas las características son muy similares en ambos grupos, excepto para la característica `Alk_Phos`, que tiene una diferencia de 7000. Esto puede deberse a que el algoritmo k-medias empleado se basa en la distancia euclídea, y la característica `Alk_Phos` tiene valores muy altos en comparación con el resto de características, por lo que tiene un peso mayor en el cálculo de las distancias.

Como se ve en el gráfico de dispersión de la figura 4, los grupos no se corresponden con las clases originales, habiendo una mezcla de pacientes de las dos clases en ambos grupos. Mientras tanto, si se representan con respecto a su valor de `Alk_Phos` como se puede ver en la figura 5, se puede observar que los grupos se han formado en función de este valor, con un grupo con valores bajos y otro con valores altos. Esto confirma la hipótesis de que la característica `Alk_Phos` tiene un peso mayor en el cálculo de las distancias, y por lo tanto, en la formación de los grupos.

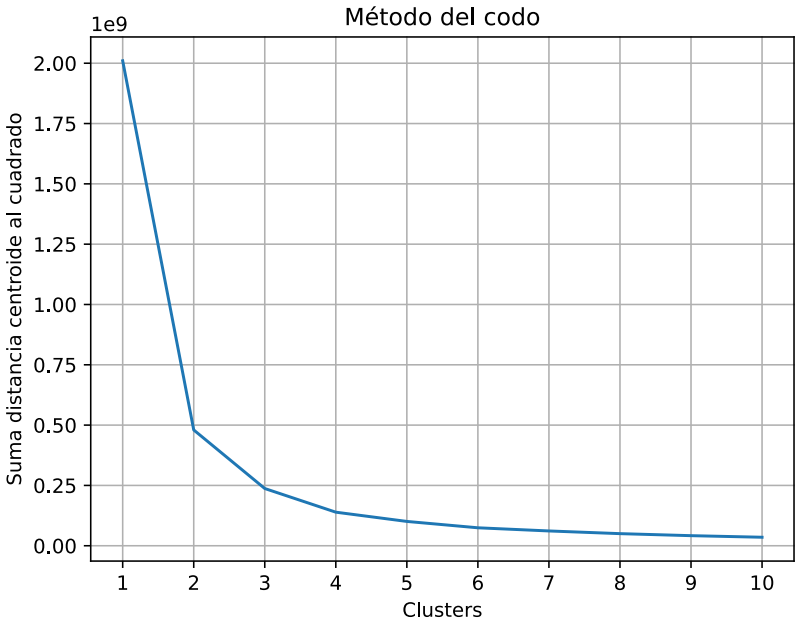

Figura 3: Método del codo para determinar el número de grupos.

Cluster	Num	Status	Age	Sex	Ascites	Hepat.
0	299	0.47	1.39	0.12	0.08	0.52
1	37	0.73	1.30	0.14	0.08	0.65
Cluster	**Spiders**	**Edema**	**Bilir.**	**Choles.**	**Albumin**	**Copper**
0	0.29	0.23	3.63	369.80	3.51	98.31
1	0.41	0.35	3.96	366.93	3.35	133.22
Cluster	**APL**	**SGOT**	**Tryglic.**	**Plat.**	**Proth.**	**Stage**
0	1438.97	124.01	123.30	255.40	10.77	2.07
1	8255.96	128.91	139.22	267.38	10.94	1.97

Tabla 2: Medias para cada grupo obtenido con el algoritmo k-medias, con 2 grupos.

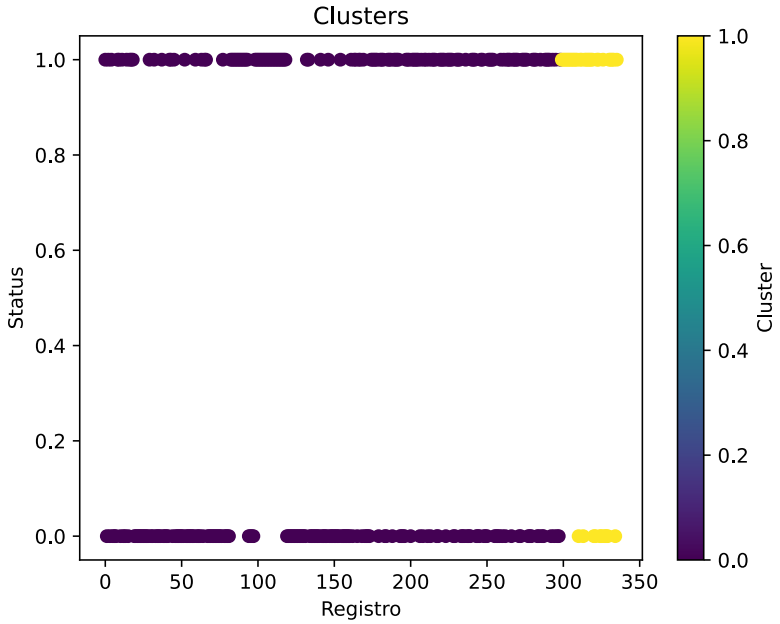

Figura 4: Gráfico de dispersión de los 2 grupos, con respecto a la etiqueta.

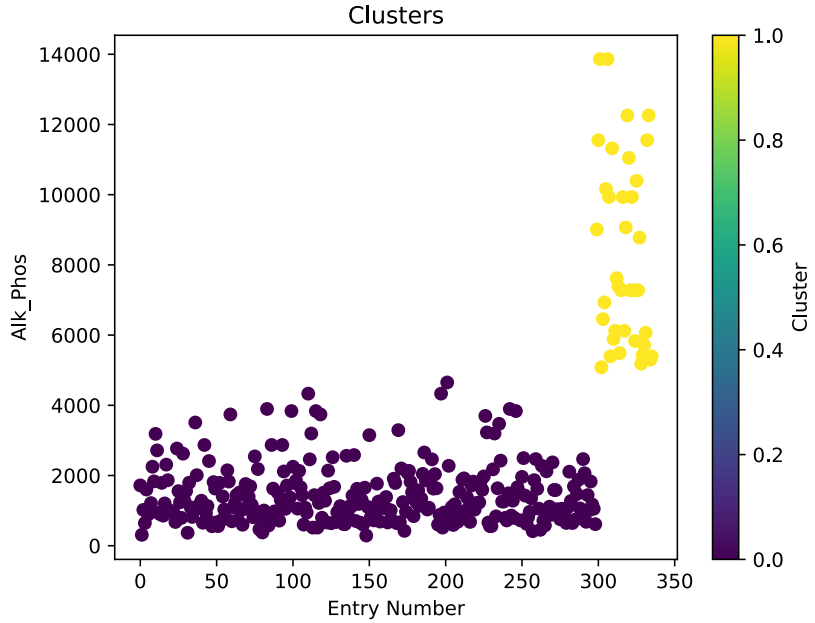

Figura 5: Gráfico de dispersión de los 2 grupos, con respecto a `Alk_Phos`.

A continuación se han normalizado las características numéricas, para que todas tengan una media de 0 y una desviación estándar de 1. Esta normalización es útil para que el algoritmo k-medias no esté sesgado hacia las características con valores numéricos más altos, asegurando que todas las características contribuyan equitativamente al análisis. Volviendo a aplicar el método del codo, se puede ver en la figura 6 que ya no hay un punto de inflexión tan claro, se ha decidido usar un número de grupos de 5.

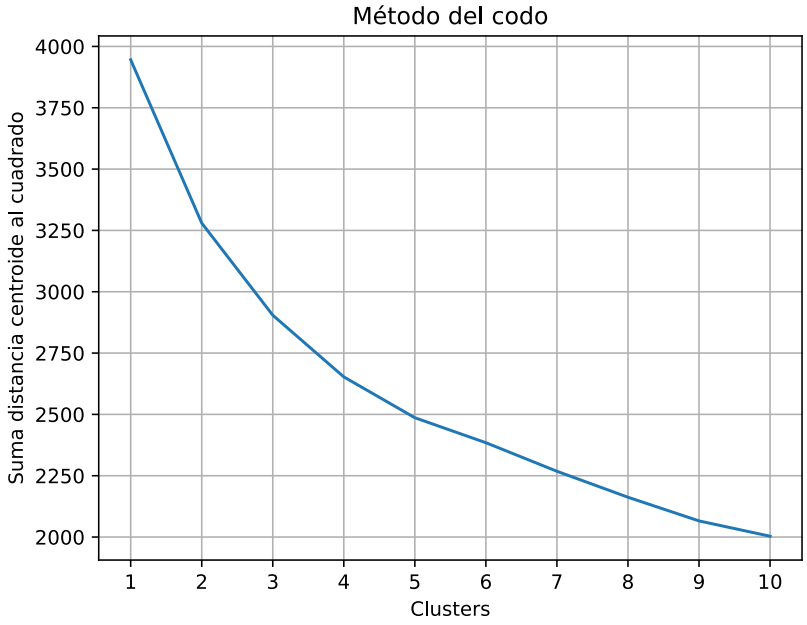

Figura 6: Método del codo con los datos normalizados.

Al aplicar el algoritmo k-medias con 5 grupos, se han obtenido los grupos con las medias que se pueden observar en la tabla 3. En este caso destaca el cluster número 2, cuyos registros pertenecen en su totalidad a pacientes fallecidos, como se puede observar en el gráfico de dispersión de la figura 7. Este grupo se corresponde con los pacientes con valores altos de `Bilirubin`, `Copper`, `Tryglicerides` y `Prothrombin`, y valores bajos de `Albumin`.

Por otra parte, el grupo 3 en su mayoría está compuesto por pacientes que sobrevidieron al estudio, con un 86 % de pacientes censurados y un 14 % de pacientes fallecidos. Este grupo se corresponde con pacientes con valores bajos de `Bilirubin`, `Copper`, `Alk_Pho`, `SGOT` y `Prothrombin`, y valores altos de `Albumin`. Además es el grupo con el menor número de pacientes con `Hepatomegaly` y `Spiders`.

Del resto de grupos no se han podido extraer conclusiones claras, ya que tienen valores medios similares en la mayoría de las características.

Cluster	Num	Status	Age	Sex	Ascites	Hepat.
0	104	0.81	1.71	0.21	0.15	0.81
1	30	0.77	0.73	0.23	0.00	0.67
2	24	1.00	1.63	0.00	0.46	0.83
3	159	0.14	1.26	0.07	0.01	0.26
4	19	0.74	1.21	0.11	0.00	0.68
Cluster	**Spiders**	**Edema**	**Bilir.**	**Choles.**	**Albumin**	**Copper**
0	0.47	0.43	-0.09	-0.30	-0.57	0.31
1	0.27	0.00	0.73	2.10	0.26	0.69
2	0.75	1.13	2.97	0.52	-0.97	0.98
3	0.13	0.06	-0.50	-0.26	0.48	-0.52
4	0.26	0.05	-0.26	-0.14	-0.09	0.36
Cluster	**ALP**	**SGOT**	**Trygli.**	**Plat.**	**Proth.**	**Stage**
0	-0.17	0.13	-0.34	-0.63	0.48	2.77
1	0.19	1.09	0.82	0.43	-0.17	1.97
2	0.18	1.02	1.58	-0.20	1.17	2.54
3	-0.35	-0.44	-0.22	0.27	-0.48	1.55
4	3.37	-0.07	0.39	0.75	0.18	1.95

Tabla 3: Medias para cada grupo obtenido con el algoritmo k-medias, con 5 grupos.

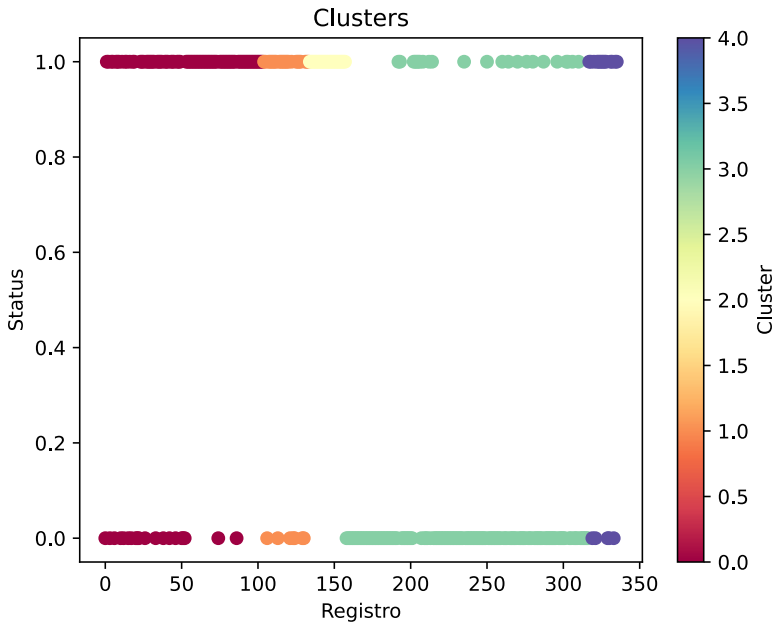

Figura 7: Gráfico de dispersión de los 5 grupos, con respecto a la etiqueta.

Para la aplicación del algoritmo **apriori** se ha decidido discretizar todas las características continuas, para poder identificar fácilmente reglas de asociación entre las características, revirtiendo los cambios realizados para el algoritmo k-medias. Para ello se ha realizado una búsqueda de los rangos normales para cada valor de análisis clínico, y se ha discretizado la característica en tres valores, que indican si el valor está por debajo, dentro o por encima del rango normal (o en el rango normal, ligeramente por encima, y muy por encima en el caso del colesterol o la bilirrubina). Los rangos normales para las características se pueden observar en la tabla 4. De estos rangos cabe destacar que ningún paciente tiene un nivel de albumina por encima del rango normal.

Característica	**0**	**1**	**2**
Bilirubin (mg/dL)	≤ 1	(1 - 2.5)	≥ 2.5
Cholesterol (mg/dL)	≤ 200	(200 - 240)	≥ 240
Albumin (g/dL)	≤ 3.4	(3.4 - 5.4)	≥ 5.4
Copper (ug/day)	≤ 10	(10 - 30)	≥ 30
Alk_Phos (U/L)	≤ 440	(440 - 1470)	≥ 1470
SGOT (U/mL)	≤ 80	(80 - 450)	≥ 450
Tryglicerides (mg/dL)	≤ 150	(150 - 199)	≥ 199
Platelets (miles/mm^3)	≤ 150	(150 - 450)	≥ 450
Prothrombin (seconds)	≤ 10	(10 - 13)	≥ 13

Tabla 4: Rangos de discretización para las características continuas.

4.2. Reglas de asociación con apriori

Para aplicar el algoritmo apriori se ha utilizao un valor mínimo de soporte de 0.2, con el objetivo de filtrar las reglas con un soporte muy bajo que no aportan información relevante, al no cumplirse para un número significativo de pacientes. Posteriormente se han ordenado las reglas obtenidas primero por su confianza y posteriormente por su *lift*, para obtener las reglas más relevantes. La confianza indica la probabilidad de que se cumpla el consecuente dado el antecedente, mientras que el *lift* indica la correlación entre el antecedente y el consecuente, siendo 1 una correlación neutra, y valores mayores que 1 una correlación positiva. De todas las reglas obtenidas se han extraído las 3 primeras, que se muestran en la tabla 5.

La primera regla indica que los pacientes sin ascitis, con el colesterol muy alto, con valores normales de transaminasa glutámico-oxalacética, con niveles altos de cobre en orina y sin hepatomegalia tienen un nivel normal de plaquetas. Esta regla no resulta muy interesante, ya que contiene muchos términos en el antecedente para determinar un valor normal de plaquetas.

La segunda regla indica que los pacientes sin edema y con niveles muy altos de bilirrubina tienen niveles altos de cobre en orina. Esta regla sí que puede ser de interés, ya que muestra una correlación que podría ser de utilidad para el diagnóstico de la enfermedad.

Por último, la tercera regla obtenida indica que los pacientes con un tiempo de protrombina bajo no tienen ascitis. Esta regla también puede ser de interés, al mostrar una correlación entre una característica clínica y la ausencia de una enfermedad.

Regla	**Soporte**	**Confianza**	**Lift**
`Ascites_0, Cholesterol_2, SGOT_1, Copper_2, Hepatomegaly_0` → `Platelets_1`	0.22	1.00	1.20
`Edema_0, Bilirubin_2` → `Copper_2`	0.26	1.00	1.15
`Prothrombin_0` → `Ascites_0`	0.26	1.00	1.09

Tabla 5: Reglas de asociación obtenidas con el algoritmo apriori.

Posteriormente se ha vuelto a aplicar el algoritmo apriori de la misma forma, pero esta vez filtrando las reglas para obtener solo aquellas con un consecuente que corresponda a la etiqueta de supervivencia. Las dos reglas más relevantes para cada valor de la clase `Status` se muestran en la tabla 6.

La primera regla muestra que los pacientes femeninos con colesterol muy alto, con hepatomegalia, con valores normales de transaminasa glutámico-oxalacética y con niveles muy altos de bilirrubina tienen una alta probabilidad de fallecer. La segunda regla es idéntica, pero elimina el atributo de transaminasa del antecedente. Estas reglas muestran una correlación alta entre la presencia de colesterol muy alto, hepatomegalia y bilirrubina muy alta con la muerte de los pacientes.

La tercera regla muestra que los pacientes femeninos sin ascitis, con valores normales de bilirrubina y fosfatasa alcalina tienen una alta probabilidad de sobrevivir. La cuarta regla es idéntica, pero elimina el atributo del sexo del antecedente. Estas reglas muestran una correlación alta entre la ausencia de ascitis y valores normales de bilirrubina y fosfatasa alcalina con la supervivencia de los pacientes.

Regla	**Soporte**	**Confianza**	**Lift**
`Sex_0, Cholesterol_2, Hepatomegaly_1, SGOT_1, Bilirubin_2` → `Status_1`	0.21	0.91	1.82
`Sex_0, Cholesterol_2, Hepatomegaly_1, Bilirubin_2` → `Status_1`	0.21	0.90	1.80
`Sex_0, Ascites_0, Bilirubin_0, Alk_Phos_1` → `Status_0`	0.21	0.90	1.79
`Ascites_0, Bilirubin_0, Alk_Phos_1` → `Status_0`	0.22	0.89	1.78

Tabla 6: Reglas de asociación de la etiqueta obtenidas con el algoritmo apriori.

Se puede considerar que la aplicación del algoritmo apriori ha aportado información relevante y desconocida anteriormente sobre las correlaciones entre características clínicas y la supervivencia de los pacientes. Los valores de confianza de las reglas son altos, lo que indica que la proporción de pacientes que cumplen el antecedente y el consecuente es alta. Además, los valores de *lift* son superiores a 1, indicando que las

reglas aparecen con más frecuencia de lo esperado en condiciones de independencia del antecedente y el consecuente, por lo que son reglas relevantes.

5. Aplicación de algoritmos supervisados

En esta sección se aplicarán distintos algoritmos de aprendizaje supervisado para realizar un análisis predictivo de los datos, intentando resolver un problema de clasificación binaria para predecir si un paciente sobrevivirá o no al estudio. En primer lugar se aplicará un algoritmo individual de clasificación, el de máquinas de vectores de soporte, y posteriormente se aplicará un algoritmo multiclasificador, el de bosques aleatorios. Los resultados obtenidos con cada algoritmo se discutirán después de describir la metodología seguida para la aplicación de cada uno de ellos, comparando las distintas métricas de evaluación.

Para validar los resultados obtenidos con los algoritmos de clasificación se utilizará una validación cruzada con 20 particiones, con el objetivo de obtener una estimación más robusta del rendimiento de los algoritmos, comparada con una única partición de entrenamiento y validación. Las particiones seleccionadas se han tomado barajando los datos antes, con una semilla constante, y se ha mantenido la proporción de clases en cada partición, para evitar perder el equilibrio entre las clases.

5.1. Máquinas de vectores de soporte

El algoritmo de **máquinas de vectores de soporte** (SVM) se encarga de encontrar un hiperplano que separe los datos en dos clases, maximizando la distancia entre los puntos más cercanos de cada clase. Para ello se utiliza una función de kernel que transforma los datos a un espacio de mayor dimensión, donde es más fácil encontrar un hiperplano que separe las clases. También es destacable el parámetro de regularización C, que controla la penalización por clasificar erróneamente un punto. Otro parámetro que se tratará es el de γ, que define el coeficiente de la función de kernel.

Se comienza la implementación de este algoritmo empleando los parámetros por defecto de la herramienta usada y sin aplicar ninguna transformación al conjunto de datos. La exactitud media obtenida durante la validación cruzada es de 66.3 %, con una desviación estándar de 11.8 %, esta métrica indica el porcentaje de predicciones correctamente realizadas por el algoritmo. A pesar de que el resultado obtenido es mejor que el de un clasificador aleatorio, no es un resultado muy bueno, y se considera que hay margen de mejora.

Para intentar mejorar el resultado obtenido se ha decidido aplicar una transformación de los datos, estandarizando las características numéricas como se hizo en el algoritmo de k-medias. Aplicar esta estandarización es recomendable para el algoritmo SVM, ya que intenta maximizar la distancia entre puntos de distintas clases, y si una característica tiene valores numéricos muy altos, esta característica

tendrá un peso mayor. Tras aplicar esta transformación, manteniendo el resto de parámetros iguales, se obtiene un valor de exactitud media del 82.2 %, con una desviación estándar de 10.5 %. Este resultado es mucho mejor que el obtenido sin aplicar la transformación, en concreto se ha mejorado en 15.9 puntos porcentuales, por lo que se considera que la estandarización de los datos es una mejora significativa. La matriz de confusión obtenida con esta configuración se puede observar en la tabla 7.

Clase	**Predicción Fallecido**	**Predicción Censurado**
Fallecido	136	32
Censurado	28	140

Tabla 7: Matriz de confusión del algoritmo SVM, con estandarización de datos.

Los falsos positivos y los falsos negativos obtenidos son muy similares, por lo que se puede considerar que el algoritmo no está sesgado hacia ninguna clase. Sin embargo para este caso es más importante reducir los falsos negativos, al ser un problema médico donde una predicción errónea puede tener consecuencias graves para el paciente. Por lo tanto, se ha decidido ajustar el peso de las clases, para que el algoritmo penalice más los falsos negativos que los falsos positivos. La clase 0 (censurado) mantiene un peso de 1, mientras que la clase 1 (fallecido) pasa a tener un peso de 1.5. Tras aplicar este cambio, manteniendo el resto de parámetros, se obtiene un valor de exactitud media del 83.3 %, con una desviación estándar de 11.9 %. Este resultado es ligeramente mejor que el obtenido anteriormente. Además, como se ve en la tabla 8, se ha reducido el número de falsos negativos, pasando de 28 a 15. Por lo tanto, se considera que este cambio es una mejora significativa en la calidad de las predicciones.

Clase	**Predicción Fallecido**	**Predicción Censurado**
Fallecido	127	41
Censurado	15	153

Tabla 8: Matriz de confusión del algoritmo SVM, con ajuste de pesos de clase.

Para intentar mejorar aún más los resultados del modelo, se ha decidido ajustar los parámetros del algoritmo SVM. Para ello se ha utilizado una búsqueda en cuadrícula, probando distintos valores de los parámetros C, kernel y γ. Los valores de C probados han sido 0.1, 1, 10 y 100, los kernels probados han sido lineal, polinómico, radial y sigmoide, y los valores de γ probados han sido el de defecto y 1/16. Tras realizar la búsqueda usando la exactitud como métrica de puntuación, los parámetros devueltos son iguales a los de defecto, con un valor de C de 1 y un kernel radial. Por lo que no se ha conseguido mejorar el resultado obtenido.

Las métricas de evaluación más relevantes para este modelo de SVM de clasificación binaria se pueden observar en la tabla 9. El área bajo la curva ROC (AUC) obtenida es del 86.9 %, esta métrica evalúa el rendimiento del modelo, siendo 1 el valor máximo y 0.5 el valor correspondiente a un clasificador aleatorio. La precisión

es del 80.2 %, esta métrica indica la probabilidad de que una predicción positiva sea correcta. El *recall* es del 91.0 %, esta métrica indica la probabilidad de que una instancia positiva sea clasificada correctamente. Por último, el valor de F1 es del 84.9 %, esta métrica es la media armónica entre precisión y *recall*. Todas estas métricas son relativamente altas, por lo que se considera que el modelo obtenido es bueno.

Métrica	**Valor (%)**	**Desviación Estándar**
Exactitud	83.3	± 11.9
AUC	86.9	± 9.2
F1	84.9	± 9.9
Precisión	80.2	± 12.9
Recall	91.0	± 8.4

Tabla 9: Resultados de las métricas de evaluación en porcentajes con su desviación estándar.

5.2. Bosques aleatorios

A continuación se aplicará el algoritmo de **bosques aleatorios** (Random Forest), un algoritmo multiclasificador que combina la predicción de varios árboles de decisión mediante el método de *bagging*, para mejorar la calidad de sus predicciones. Para trabajar con este algoritmo se han revertido los cambios realizados sobre el conjunto de datos para el algoritmo SVM.

No se intentará estandarizar los valores continuos, ya que este algoritmo está basado en árboles de decisión, por lo que no emplea ninguna medida de distancia y no es sensible a valores de gran valor numérico en una característica. Los árboles de decisiones toman decisiones en función de umbrales que dividen las distintas características, por eso no es necesario estandarizar los datos.

Inicialmente se ha aplicado el algoritmo con los parámetros por defecto de la herramienta usada, con un valor de semilla constante para poder obtener resultados comparables entre sí. La exactitud media obtenida durante la validación cruzada es de 84.8 %, con una desviación estándar de 7.8 %. Este resultado de inicio ya es mejor que el obtenido al final con el algoritmo SVM, por lo que se considera que este algoritmo es más adecuado para este problema. Sin embargo, como se puede ver en la tabla 10, el número de falsos negativos es de 22, superior al obtenido con el algoritmo SVM. No se intentará ajustar el peso de las clases, ya que *Random Forest* permite que los árboles de decisión crezcan hasta que todas las hojas sean puras, por lo que ajustar el peso de las clases no tiene un efecto significativo en el resultado.

Para intentar optimizar el resultado obtenido se ha decidido ajustar los parámetros del algoritmo, realizando una búsqueda en cuadrícula sobre los parámetros que se ven en la tabla 11. Los mejores parámetros descubiertos han devuelto una exactitud del 85.1 %, con una desviación estándar de 9.4 % y sin una mejora en el número de falsos

Clase	Predicción Fallecido	Predicción Censurado
Fallecido	139	29
Censurado	22	146

Tabla 10: Matriz de confusión del algoritmo Random Forest, con parámetros por defecto.

negativos. Este resultado es ligeramente mejor que el obtenido con los parámetros por defecto, pero la diferencia es muy inferior a la desviación estándar obtenida en ambos resultados, por lo que la mejora se considera no significativa y se mantienen los parámetros por defecto.

Parámetro	Valores Probados
n_estimators	10, 50, 100, 200, 500
max_features	None, sqrt, log2
max_depth	None, 3, 5, 10, 20, 30
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrap	True, False

Tabla 11: Parámetros probados en la búsqueda en cuadrícula.

Por último, se ha intentado mejorar la calidad de las predicciones aplicando una transformación de los datos, discretizando las características continuas como se hizo para el algoritmo apriori. La efectividad de este método en el caso del algoritmo *Random Forest* depende de lo bien que se haga la partición de los datos, ya que es imprescindible realizar una buena división para no perder información de ninguna de las características. Al discretizar los valores se obtiene exactamente la misma exactitud y matriz de confusión que con los parámetros por defecto, por lo que se considera que esta transformación no es útil para este algoritmo. Esto se debe a que no se tiene un conocimiento experto sobre el dominio médico de las características, por lo que no se gana ninguna información útil al discretizar los valores según los rangos normales.

Las métricas de evaluación más relevantes para este modelo se pueden ver en la tabla 12, junto con las obtenidas previamente para el modelo de SVM. El modelo de *Random Forest* obtiene un mejor valor que SVM en todas las métricas excepto en el *Recall*, donde obtiene un valor ligeramente inferior. Esto se debe a que tiene un mayor número de falsos negativos, por lo tanto, a pesar de que el modelo de *Random Forest* es muy bueno, se puede considerar el uso del modelo de SVM para este problema, ya que su menor número de falsos negativos puede suponer la diferencia entre la vida y la muerte de un paciente.

Métrica	Valor (%)	stddev
Exactitud	84.8	± 7.8
AUC	91.2	± 7.1
F1	85.1	± 7.9
Precisión	85.0	± 10.8
Recall	87.2	± 13.0

Tabla 12: Métricas de RF.

Métrica	Valor (%)	stddev
Exactitud	83.3	± 11.9
AUC	86.9	± 9.2
F1	84.9	± 9.9
Precisión	80.2	± 12.9
Recall	91.0	± 8.4

Tabla 13: Métricas de SVM.

El algoritmo *Random Forest* también permite ver la importancia que ha asignado a las distintas características para realizar las predicciones. En la tabla 14 se puede observar que las características más importantes son `Bilirubin`, `Copper` y `Prothrombin`. Estos resultados tienen sentido ya que la bilirrubina es un indicador del correcto funcionamiento del hígado, y aparece múltiples veces en las reglas de asociación obtenidas con el algoritmo apriori.

Característica	Importancia
Bilirubin	0.20
Copper	0.14
Prothrombin	0.12
SGOT	0.09
Alk_Phos	0.08
Albumin	0.08
Platelets	0.06
Cholesterol	0.06
Tryglicerides	0.05
Stage	0.03
Age	0.03
Hepatomegaly	0.02
Edema	0.02
Sex	0.01
Ascites	0.01
Spiders	0.01

Tabla 14: Importancia de las características.

Para finalizar, en la figura 8 se muestra un ejemplo de un árbol de decidiones aleatorio obtenido del modelo de *Random Forest*. Este árbol se ha obtenido con los parámetros por defecto, y se puede observar que es un árbol muy profundo, con 14 niveles.

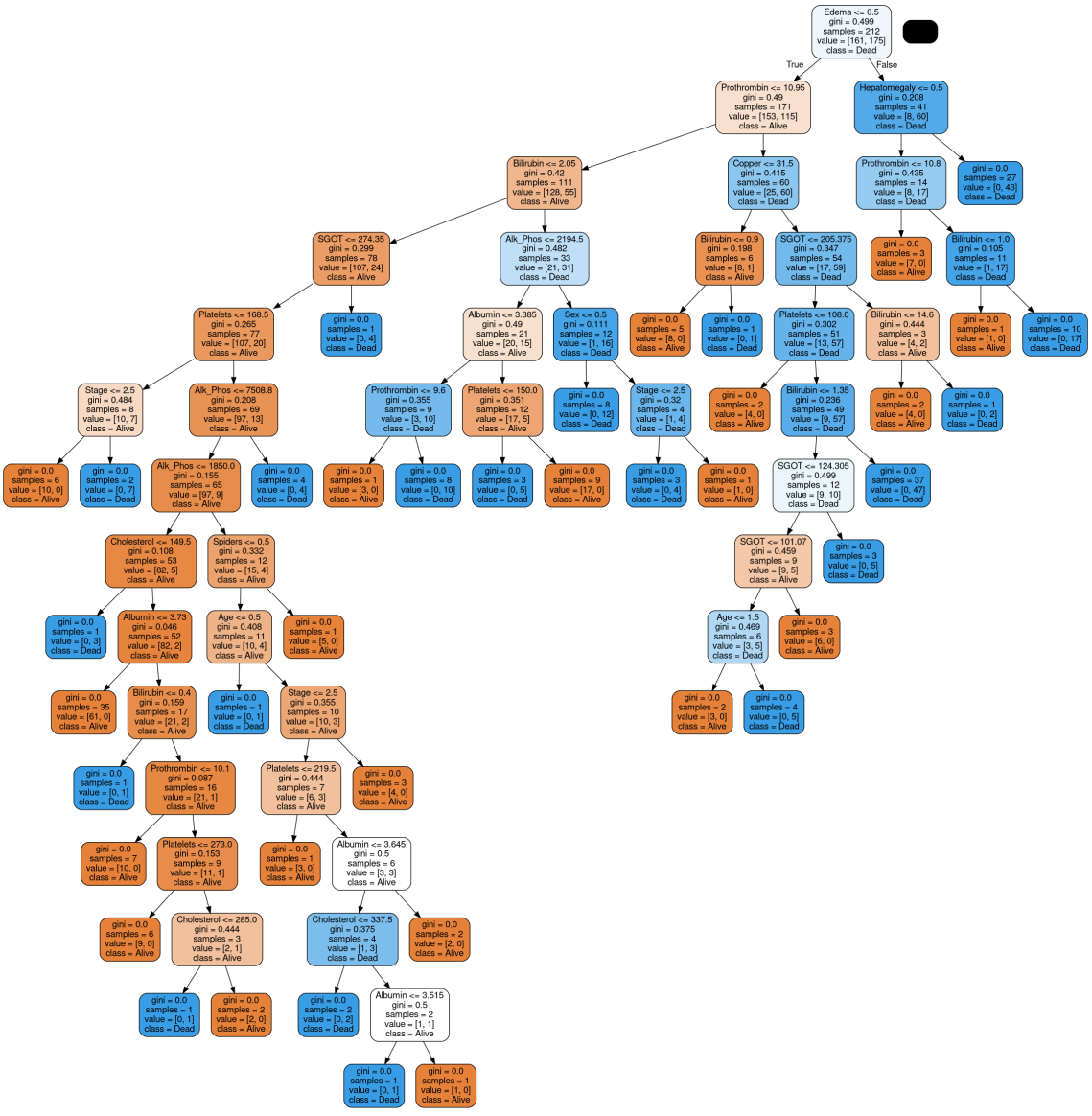

Figura 8: Árbol de decisión aleatorio obtenido del modelo de Random Forest.

6. Conclusiones

En este trabajo se ha realizado un análisis exploratorio de los datos, aplicando distintos algoritmos de agrupamiento y asociación para obtener información relevante sobre los datos. Se ha aplicado el algoritmo k-medias para agrupar los datos en distintos grupos, y se ha obtenido información relevante sobre 2 de los 5 grupos obtenidos. También se ha aplicado el algoritmo apriori para obtener reglas de asociación entre las características clínicas y la supervivencia de los pacientes. Se han obtenido reglas que muestran una correlación entre distintas características clínicas y la supervivencia de los pacientes.

También se han aplicado distintos algoritmos de aprendizaje supervisado para realizar un análisis predictivo de los datos, intentando resolver un problema de clasificación binaria para predecir si un paciente sobrevivirá o no al estudio. Se ha aplicado el algoritmo de máquinas de vectores de soporte, y posteriormente se ha aplicado un algoritmo multiclasificador, el de bosques aleatorios. Se han obtenido resultados relativamente buenos con ambos algoritmos, teniendo cada uno ventajas y desventajas. El algoritmo SVM obtiene un menor número de falsos negativos, sin embargo, el algoritmo *Random Forest* obtiene un mejor valor en todas las métricas de evaluación, excepto en el *Recall*.

En conclusión, se considera que se han cumplido los objetivos de este trabajo, al obtener dos modelos que permiten pronosticar las probabilidades de supervivencia de pacientes con cirrosis biliar primaria, obteniendo también información sobre la importancia de las distintas características clínicas, entre las que no se incluye la penicilamina, ya que no se ha demostrado que tenga correlación alguna con la supervivencia de los pacientes.

Referencias

[1] E. Dickson, P. Grambsch, T. Fleming, L. Fisher y A. Langworthy. Cirrhosis Patient Survival Prediction. UCI Machine Learning Repository, 2023. DOI: 10.13026/C2F305.

[2] E. Rolland Dickson, Thomas R. Fleming, Russell H. Wiesner, William P. Baldus, C. Richard Fleming, Jurgen Ludwig y John T. McCall. Trial of Penicillamine in Advanced Primary Biliary Cirrhosis. *New England Journal of Medicine*, 312(16):1011-1015, 1985. DOI: 10.1056/NEJM198504183121602. PMID: 3885033.

[3] Ministerio de Sanidad. Patrones de mortalidad en España, 2019, Madrid, 2022. NIPO: 133-20-003-1.

[4] Albert Parés, Agustín Albillos, Raúl-J. Andrade, Marina Berenguer, Javier Crespo, Manuel Romero-Gómez, Mercè Vergara, Belén Vendrell y Alicia Gil. Colangitis biliar primaria en España. Resultados de un estudio Delphi sobre su epidemiología, diagnóstico, seguimiento y tratamiento. es. *Revista Española de Enfermedades Digestivas*, 110:641-649, octubre de 2018. ISSN: 1130-0108.