FACULTAD DE CIENCIAS

SOCIALES Y JURÍDICAS DE ELCHE

TRABAJO FIN DE GRADO

COMO AFECTA LA REDUCCIÓN DE ATRIBUTOS EN LOS ARBOLES DE CLASIFACIÓN CON PACIENTES DE HEPATITIS

***Alumno**: Francisco José Richarte Rodríguez*

***Tutor**: Alejandro Rabasa Dolado y Miriam Esteve Campello*

4º ESTADÍSTICA EMPRESARIAL

Índice de Contenido

Índice de Tablas

Índice de Figuras

1.RESUMEN

El seguimiento de los pacientes con cualquier enfermedad de la que se quiera extraer información cada vez es más habitual, debido a dos razones, numerosos avances en las investigaciones de las enfermedades y numerosos profesionales que en esto últimos años se están formando en la ciencia de datos (extracción, procesamiento, análisis y predicción).

Esta investigación tiene como objetivo principal averiguar si una selección de atributos previa a la construcción del modelo afecta negativamente o positivamente al desarrollo del modelo, tiempo de ejecución, compresibilidad de los resultados etc. Además de hacer un estudio completo de los datos que se presentan para utilizar todas las técnicas de Minería de datos (factorización y Selección de Atributos) e incluso intentaremos predecir el desenlacé de los pacientes (Arboles de decisión).

En conclusión, haremos un estudio completo de los pacientes con hepatitis obtenido de la Universidad de Carnegie-Mellon en noviembre de 1988, pero siempre respondiendo la pregunta de, ¿qué mejora obtenemos al incluir una selección de atributos previa a la construcción del modelo?

Palabras clave: Enfermedad, técnicas de Minería de datos, Factorización, Selección Atributos, Arboles de decisión

2. INTRODUCCIÓN Y OBJETIVOS

2.1 Introducción al TFG

En la actualidad tanto hospitales como universidades o centros de investigación han unido fuerzas para entender mediante los datos recogidos el posible final del paciente dependiendo de alguna sintomatología durante la enfermedad. En nuestro caso la Hepatitis es una enfermedad que se cobra anualmente alrededor de un millón de vidas, por ello muchos centros de investigación centra su investigación en esta enfermedad.

La Hepatitis es una enfermedad que afecta al Hígado, principalmente ocurre que este se inflama impidiendo con ello el buen funcionamiento de este, recordemos que junto con el corazón el Hígado es uno de los órganos más importantes para el correcto funcionamiento del ser humano, es un órgano vital.

Las técnicas de Minería de datos permiten depurar los datos que recibimos por parte de los hospitales, con estas herramientas conseguimos clasificar y predecir el posible desenlace que tendrá el paciente, así conseguimos más información sobre que situaciones son las más importante, cuales de ellas nos proporcionan una mala señal sobre el desenlace de la persona.

Esta investigación tiene como objetivo aplicar las técnicas que hemos nombrado anteriormente (Factorización, Selección de Atributos y Árbol de clasificación) sobre los datos obtenidos de los pacientes en la Universidad Carnegie-Mellon en Yugoslavia a fecha de noviembre de 1988, con ello obtendremos un Ranking de variables, de las variables más importantes para predecir si un paciente Muere o vive.

Además de aplicar estas técnicas consiguiéremos comparar que precisión conseguimos con el modelo completo y con uno más reducido al haber utilizado una técnica de selección de atributos, además podremos comparar su rendimiento, pero más importante aún su entendimiento cuando observemos la figura del árbol de decisión. En esta investigación se ha obtenido un matriz de confusión con el modelo reducido (Seleccionando atributos) de un 84,61% de precisión.

2.2 Objetivos del TFG

Los objetivos principales para este trabajo son tres:

- Demostrar los beneficios de utilizar una selección de atributos para mejorar el modelo.
- Conseguir Predecir el posible desenlace de nuevos pacientes o que factores son críticos para cada desenlace.
- Explorar y aprender en profundidad las técnicas nombradas anteriormente de una forma práctica en un entorno real.

2.3 Objetivos Personales

Los objetivos personales que me planteo conseguir con este estudio son:

- Aplicar los conocimientos y técnicas aprendidas en el Grado de Estadística Empresarial y conseguir una visión práctica de estas técnicas.

- Profundizar más en las técnicas que hemos mencionado anteriormente y ponerlas en práctica en un ejemplo de la vida real.

3. ESTADO DE LA CUESTIÓN Y MARCO TEÓRICO

3.1 Métodos de Factorización

A menudo necesitaremos que algunas variables continuas tengan que ser asignadas a etiquetas previamente establecidas para así poder tratar los datos como variables discretas, a este proceso de transformación de las variables se le denomina factorización o discretización, una situación común aparece cuando tenemos una variable que se refiere a la edad del individuo, el peso o la altura estas variables son numéricas, pero nos interesa obtener información de los sub-rangos(etiquetas) que hemos comentado anteriormente, gracias a esta transformación obtendremos si queremos una visión descriptiva de los datos que hemos discretizado mucho más clara que si no hubiésemos aplicado esta transformación y muchos de los algoritmos que se utilizan para clasificar o predecir requieren que las variables del modelo sean categóricas u ordinales, además conseguiremos una mayor optimización del modelo. Existen numerosas formas de clasificar los métodos de discretización (División vs Fusión, Global vs Local, Estático vs. Dinámico y Supervisado vs No Supervisado) aunque nos centraremos solo en desarrollar el último grupo. (Dougherty, Kohavi, & Sahami, 1995)

La principal diferencia entre estos dos grupos es que los métodos no supervisados discretizan la variable sin tener en cuenta la etiqueta en cambio los supervisados utilizan la información de esta etiqueta y entonces comienzan a discretizar la variable. Además los métodos no supervisados dividen la variable utilizando la distribución de los valores, en cambio los supervisados incluyen la anchura del intervalo, anchura de la frecuencia, la relación entre la variable factorizada y su clasificación, nosotros utilizaremos el primero de ellos, los algoritmos no supervisados, en concreto utilizaremos unos de los algoritmos más utilizados en estos casos, K-means este método al referirse sólo sobre una variable análisis de clustering de k-means "unidimensional". Este método fue propuesto por primera vez por Stuart Lloyd en 1957 aunque esté no se publicó hasta 1982. (Revista de investigación Industrial Data, 2021)

Este método tiene como objetivo dividir n observaciones en k grupos, y que estas observaciones pertenezcan al grupo con la media más cercana al centroide del grupo (menor distancia euclídea al cuadrado) buscaremos que la distancia intra-cluster sean pequeñas y las distancias inter-cluster sean grandes. Este modelo se puede describir así y tienes dos pasos:

1. Paso de asignación: Asignar cada observación al grupo más cercano:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\},$$

Fig. 1. Paso de Asignación K-means.
Fuente: (NC State University)

Donde cada Xp se asigna exactamente a cada Si(t)

2. Paso de actualización: recalcula esos grupos

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Fig. 2. Paso de Actualización K-means.
Fuente: (NC State University)

El algoritmo ha convergido cuando las asignaciones tras la "actualización" no cambian.

3.2 Métodos de Selección de Atributos

En muchas ocasiones nos enfrentaremos a conjuntos de datos de la vida real y estos en la mayoría de los casos tienen datos redundantes ya que la mayoría de veces se guarda todo aunque la información de unas variables pueda sustituir por otra, para ello intentaremos seleccionar las "importantes", este método muchas veces se confunde con los métodos de reducción de las dimensiones, estos métodos su objetivo es transformar las variables de tal forma que una variable resuma n variables y las represente en una nueva dimensión, las técnicas más utilizas son (Análisis de componentes principales y Análisis factorial) no obstante nosotros no utilizaremos la reducción de dimensiones si no la selección de atributos que a diferencia del anterior método este no modifica las variables simplemente se "eligen". Seguramente pensemos que la forma idónea para construir un modelo es utilizar sus así tenemos toda la información, este planteamiento tiene una serie de problemas (rendimiento, costes, variables con la misma información, sobreajuste del modelo etc.) lo que se pretende con la selección de atributos es obtener unos datos que se consigan unos resultados similares, si no mejores, en comparación con los datos originales, en el modelo. (Gonzalez, aprendeIA, 2020)

Los beneficios de realizar una selección de atributos previamente al modelado son los siguientes:

- **Mayor Precisión**: Eliminar atributos que pueden ser irrelevantes o redundantes: con ello conseguiremos no confundir al modelo que clasifica.
- **Optimizar el tiempo y los recursos:** Al tener menos variables el modelo conseguirá clasificar antes y utilizar menos recursos de nuestro ordenador.

Para reducir la dimensionalidad hay diferentes métodos en esta imagen podemos ver varios de ellos:

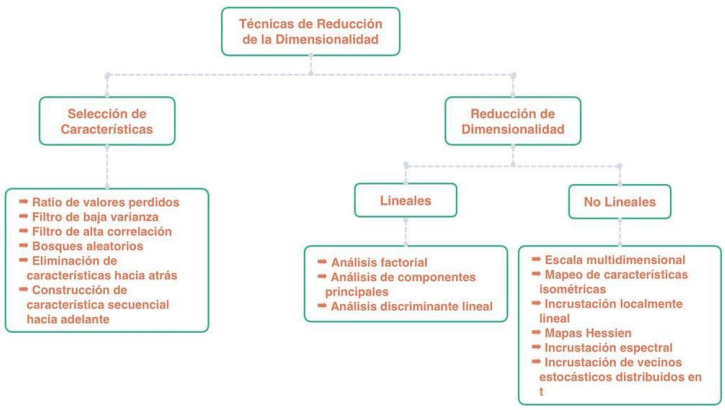

Fig. 3. Clasificación de las diferentes técnicas de reducción de la dimensionalidad. Fuente: (González, aprendeIA, 2020)

Esta selección característica se puede hacer con diferentes técnicas:

- **Selección Univariante:** este método utiliza técnicas estadísticas para seleccionar que variables tienen una mayor correlación con la variable que estamos estudiando. (Gonzalez, aprendeIA, 2020)
- **Eliminar características de baja variación**: Eliminamos todas las características cuya varianza no alcance un umbral predeterminado que a menudo suele ser cero que significaría que todas las características son iguales para diferentes valores de la variable objetivo. (Gonzalez, aprendeIA, 2020)
- **Eliminar características altamente correlacionadas**: Cuando tenemos características colineales pueden causar un sobreajuste, por ello podemos eliminar una de ellas que no perderemos mucha información, la cuestión es cual eliminar y es característica que tenga menos correlación con el objetivo. (Gonzalez, aprendeIA, 2020)
- **Métodos de filtro:** Estos se utilizan generalmente como un paso en el preprocesamiento de datos, las características se clasifican según unos puntajes estadísticos respecto a la variable objetivo. Según las características y la predicción de la variable objetivo usaremos (Gonzalez, aprendeIA, 2019):

↓Características/Predicción →	Continuo	Categórico
Continuo	Correlación de Pearson	LDA
Categórico	Anova	Chi-cuadrado

Fig. 4. Tipo de medidas de correlación. Fuente: (González, aprendeIA, 2019)

- **Métodos de Envoltura:** A diferencia del anterior este utiliza Algoritmos de Machine Learning y este utiliza su rendimiento como criterio de selección, es decir, busca la características que más se adecua al algoritmo pero también que este sea optimo, tratamos de usar subconjunto de características y entrenar el modelo, basándonos en los resultados de ese modelo decimos si agregar o eliminar características del subconjunto, se puede reducir a un método de búsqueda por ello es tan importante lo óptimo que sea el modelo ya que computacionalmente es muy caro(selección hacia delante, Eliminación hacia atrás, Eliminación de características recursivas..)
- **Métodos Integrados:** Combina los dos anteriores incorporando algoritmos que tiene sus propios métodos de selección, los más utilizados son los métodos de regresión LASSSO Y RIDGE. (Gonzalez, aprendeIA, 2019)

En este proyecto usaremos el paquete llamado Boruta disponible en R que utiliza un método de envoltura para seleccionar características.

El algoritmo de Boruta comienza creando copias aleatorias de las características para agregar aleatoriedad a estas características se denominan características de sombra, entrena un Random Forest con este conjunto y aplica una medida de importancia como Mean Decrease Acurracy y evalúa cada característica, en cada iteración comprueba si una característica real tiene una mayor importancia. Finalmente se detiene cuando las características se confirman o se rechazan. Boruta encuentra todas las características que son fuertemente o débilmente relevantes para la variable objetivo, esto en un entorno médico es adecuado ya que podemos observar que variables están conectadas con la variable objetivo y cuáles no.

Definitivamente seleccionaremos las características de (Protime, Albumin, Histology, Ascites, Anorexia y Varices) para crear nuestro modelo para predecir resultados.

3.3 Métodos de Clasificación

En el marco de la ciencia de datos este proceso se realiza en dos pasos un paso de aprendizaje y un paso de predicción. En el primero desarrollaremos el modelo en base a nuestros datos de entrenamiento y en el siguiente paso lo utilizaremos para predecir introduciéndole unos datos. En el primer paso deberemos separa nuestro conjunto de datos en dos subconjuntos uno de entrenamiento y otro de test, el subconjunto de entrenamiento lo utilizaremos para construir el modelo y el test para poder validarlo al igual que para saber la precisión que obtenemos con el modelo que hemos formado.

Existen diversos métodos, tanto por la rama de algoritmos supervisados como la de no supervisados, nos centraremos en estos últimos, existen números algoritmos de clasificación, pero solo hablaremos de los árboles de decisión (Decision Tree) a diferencia de otros algoritmos este sirve para dar solución a problemas de regresión y clasificación, nos centraremos en esto últimos el objetivo es predecir la variable objetivo dependiendo de unos parámetros.

Hay numerosos arboles de decisiones (ID3,C4.5,CART,CHAID,MARS), pero nombraremos solo los dos primeros, el ID3 (Quinlan,1986) es el primero que se publicó, este utiliza la métrica de ganancia de información que mide la diferencia de la entropía antes de la división y la entropía media después de la división.

Pero no utilizaremos esta “versión” de los árboles de decisión puesto que solo trabaja con variables categóricas y nuestro conjunto de datos tiene tanto variables categóricas como continuas, su sucesor el C4.5 (Quinlan,1993) sí que puede lidiar con valores tanto categóricos como continuos además utiliza la métrica de relación de ganancia, la relación de ganancia es una mejora de la anterior métrica ya que tiene en cuenta el número de ramas en el que se va dividir teniendo en cuenta la información intrínseca de una división.

Los árboles de decisión clasifican los ejemplos ordenándolos hacia abajo donde nuestro conjunto de entrenamiento esta en la raíz hasta un nodo terminal (no se subdivide) o un nodo de decisión que se subdivide.

Como vemos es un método muy completo y que nos muestra con una representación muy sencilla los tipos de nodos.

Una vez hemos creado el modelo con los datos entrenamiento, tendremos que validar que nuestro modelo sea correcto y no este sobre ajustando el modelo, esto sería hacer una rama con cada caso y así obtener un 100% de precisión.

La precisión o "Acurracy" es la métrica de calidad del modelo en ella se refleja registros bien clasificados frente al total, esta métrica toma valores entre 0 y 1 (donde 1 significa que total de registros están correctamente clasificados). Para ello utilizaremos el subconjunto de test.

Para entender esto utilizaremos un ejemplo visual para esto utilizaremos la matriz de confusión para mostrar valores bien y mal registrados, que tiene un aspecto al similar a la que aparece a continuación

Tabla 1 : Estructura de una matriz de confusión
Fuente: Elaboración propia

	+	-
+	200	7
-	8	100

En este ejemplo concreto tenemos 315 registros en total, donde 300 se clasificaron de manera correcta (Suma diagonal:200+100) y 15 que se clasificaron mal, dentro de ellos observamos que 8 han sido falsos negativos (registros que eran + y el modelo ha clasificado como -) y 7 falsos positivos (registrados como – y el modelo ha clasificado como +). En este ejemplo obtendríamos un acurracy de 300/315=0.95(95%), no obstante, aunque siempre nos guiaremos por esta medida para escoger uno u otro, por regla general siempre se busca que el error se repara de manera equitativa entre falsos negativos y falsos positivos para que no todo el error sea de un tipo, pero esto depende ya que algunos casos podremos permitirnos muchos falsos positivos pero muy pocos o ningún falso negativo o viceversa, dependiendo el ámbito donde estemos trabajando podemos permitirnos un tipo de error u otro.

4.HIPOTESIS DE PARTIDA

4.1 Modelos Predictivos precisos y Fáciles de interpretar

De los diferentes modelos disponibles utilizaremos el algoritmo de CART (*Classification and Regression Trees)* que está en la librería MachineLearning en R como RPART, con este algoritmo conseguiremos mucha precisión, la versatilidad que nos proporciona los parámetros en R-Studio además como hemos nombrado anteriormente lo compresible que es entender los resultados que arroja el árbol de decisión visualmente.

4.2 Utilizar la selección de atributos como optimizador del árbol de decisión

En este TFG se pretende comprobar si la selección de atributos mejora en conste computacional y en resultados más precisos si la selección de unas características es mejor o no, utilizaremos el método Boruta (apartado 3.3)

5.METODOLOGIA

5.1 LOS DATOS

Este conjunto de datos pertenece a la Universidad Carnegie-Mellon en Yugoslavia a continuación se muestra la naturaleza de los distintos atributos que componen este conjunto de datos:

Tabla 2 : Variables en el conjunto de datos el cual vamos hacer el estudio
Fuente: Elaboración propia

Atributo	Descripción	Tipo de datos
Class	Resultado del paciente	Binario: 1-muere, 2-sobrevive
Age	Edad del paciente	Numérico: mínimo 7, máximo 78
Sex	Sexo del paciente	Binario: 1-hombre, 2-mujer
Steroid	¿Ha tomado esteroides?	Binario: 1-no, 2-sí
Antivirals	¿Ha tomado Antivirales?	Binario: 1-no, 2-sí
Fatigue	¿Ha padecido Fatiga?	Binario: 1-no, 2-sí
Malaise	¿Ha padecido Malestar?	Binario: 1-no, 2-sí
Anorexia	¿Ha padecido Anorexia?	Binario: 1-no, 2-sí
Liver big	¿Ha tenido el síntoma del hígado "Grande"?	Binario: 1-no, 2-sí
Liver firm	¿Ha tenido el síntoma del hígado "pequeño y duro"?	Binario: 1-no, 2-sí
Spleen Palpable	¿Ha tenido el síntoma del Bazo Palpable?	Binario: 1-no, 2-sí
Spiders	¿Ha tenido vasos sanguíneos inflamados en forma de araña?	Binario: 1-no, 2-sí
Ascites	¿Ha tenido acumulación de Líquidos en el cuerpo?	Binario: 1-no, 2-sí
Varices	¿Ha tenido Varices?	Binario: 1-no, 2-sí
Bilirubin	Cantidad de Bilirrubina	Numérico: mínimo 0.3, máximo 8
Alk phosphate	Cantidad de Fosfata alcalina	Numérico: mínimo 26, máximo 295
Sgot	Cantidad de Aspartato aminotransferasa	Numérico: mínimo 14, máximo 648
Albunim	Cantidad de Albunimia	Numérico: mínimo 2.1, máximo 6,4
Protime	Cantidad de Protombina	Numérico: mínimo 0, máximo 100
Histology	¿Ha tenido Histologia?	Binario: 1-no, 2-sí

El estudio tiene como objetivo comprobar cómo afecta la selección de atributos a nuestra predicción además de ver todo el proceso desde el preprocesado, descriptivos y posterior predicción.

Para llevar a cabo este estudio utilizaremos el software Rstudio, este programa ofrece numerosas librerías de Machine Learning además utilizaremos la Liberia "Plotly" para poder representar las variables en gráficos, he decidido utilizar este la librería debido a que cuando se utiliza para aplicar en un web los gráficos se convierten en iterativos, es decir, colocas el cursor encima y te dice los datos además de la facilidad que tiene para

representar con apenas dos líneas de código. Contamos con un hardware cuyo nombre es DESKTOK-LBRMQNF, con un procesador Intel(R) Core (TM) i5-9300H CPU @ 2.40GHz, memoria RAM de 16,0GB (15,9BG usable) y cuyo sistema operativo es de 64 bits, procesador basado en una arquitectura x64, también dispone de dos discos duros uno solido de 256 GB y un disco duro mecánico de 1 TB

5.2 Descriptivos

A continuación, veremos las distintas variables relacionadas con la variable que deseamos predecir, en busca de una posible relación, patrón etc con ello podremos hacernos una idea de cómo se distribuyen los datos y si tenemos que hacer alguna tarea de preprocesado como eliminar blancos, sustituirlos por la media por una predicción etc.

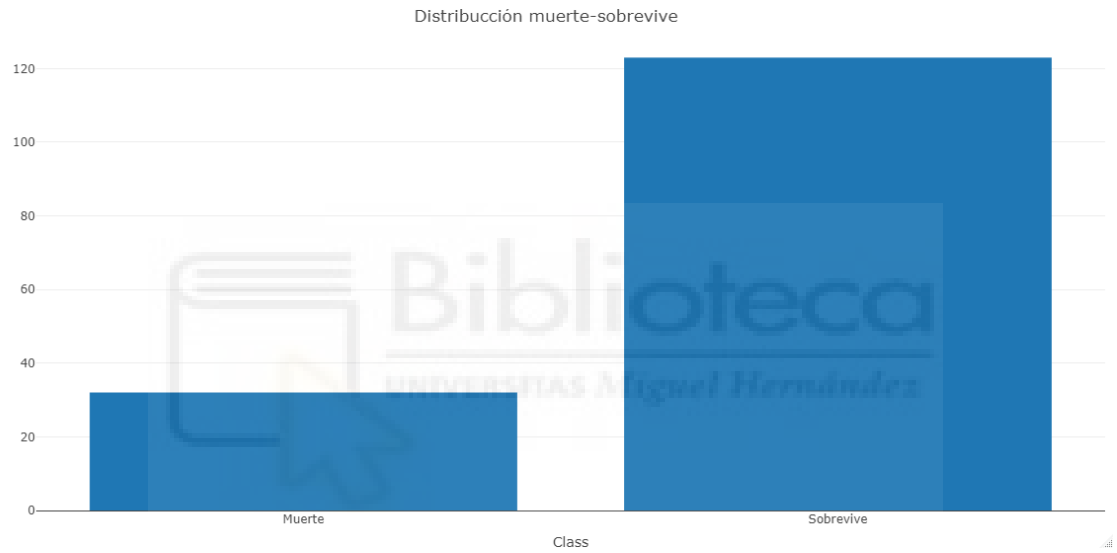

Fig. 5. Distribución de la variable a predecir muerte-sobrevive
Fuente: Elaboración Propia

Se observa en la figura 5 que 32(20.64%) de los pacientes acaban falleciendo mientras el 123(79.35%) sobrevive, por lo que la mayoría de nuestra muestra sobrevive a la enfermedad.

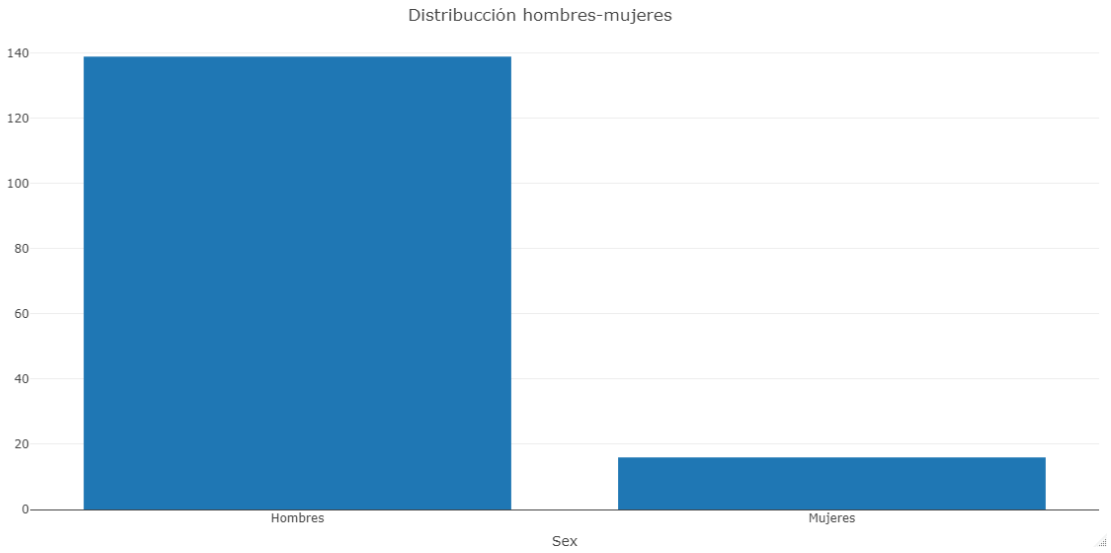

Fig. 6. Distribución hombres-mujeres en la muestra
Fuente: Elaboración propia

Observamos la figura 6, 139(89.67%) hombres y 16(10.32%) mujeres, tenemos muy pocas mujeres en nuestra muestra, esto podría significar que la variable sexo no es importante para la variable a predecir.

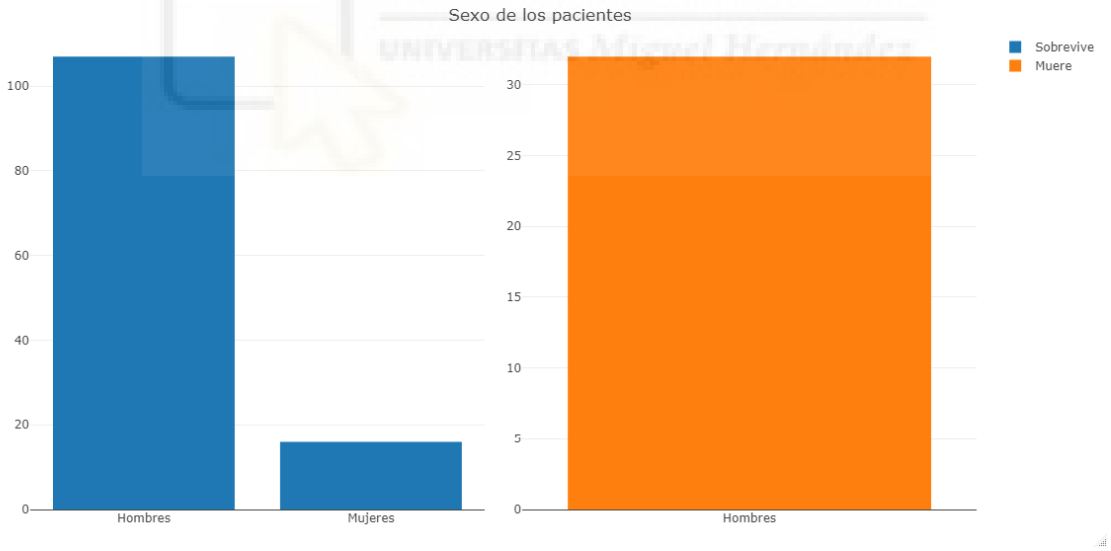

Fig. 7. Distribución hombres-mujeres o no relaciona con la variable a predecir
Fuente: Elaboración propia

Podemos observar en la figura 7, que no tenemos ninguna mujer en esta muestra que muera de esta enfermedad, 32 de los de los 139 hombres mueren lo que representa un 23.03% sobre los hombres y 20.64% sobre la muestra completa.

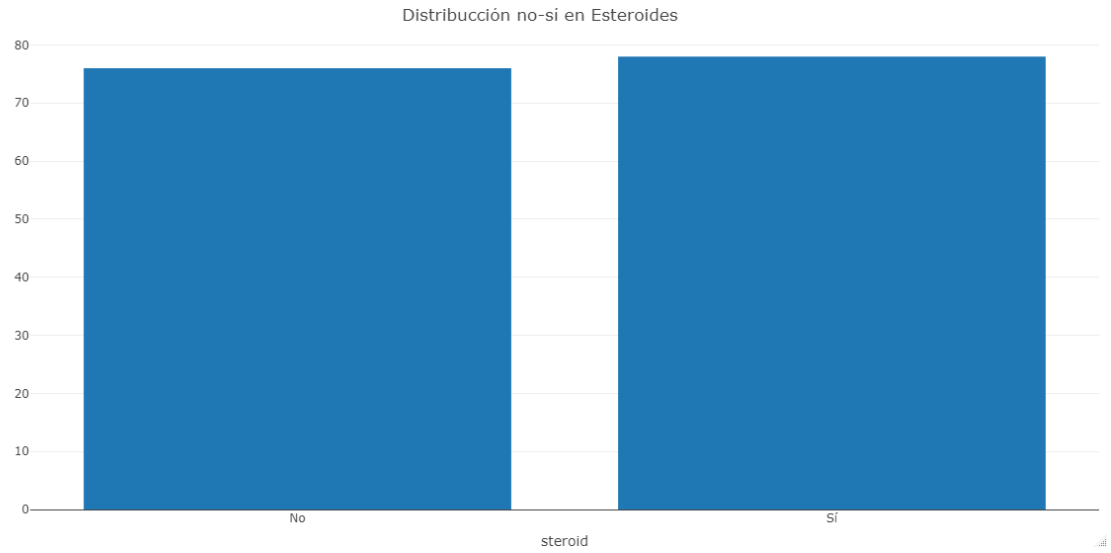

Fig. 8. Distribución del consumo de esteroides
Fuente: Elaboración propia

Si observamos la figura 8, el consumo de esteroides puede desencadenar en una hepatitis y que esta a su vez sea más grave. La proporción de Personas que no han tomado esteroides y los que si han tomado es muy semejante con 49,03%(76) y 51,97%(78) respectivamente.

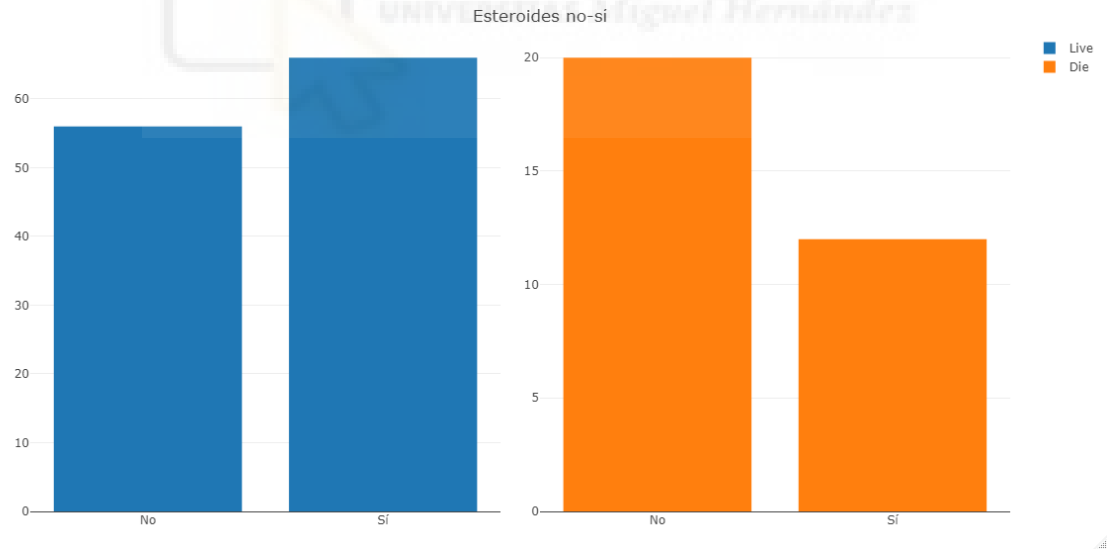

Fig. 9. Distribución del Consumo de esteroides o no relaciona con la variable a predecir
Fuente: Elaboración propia

Si observamos la figura 9 podemos ver una población igual repartida entre vivos y muertos que hallan consumido esteroides y que no por lo que al igual que la anterior no parece que esta variable a simple vista guarde una relación con la variable a predecir.

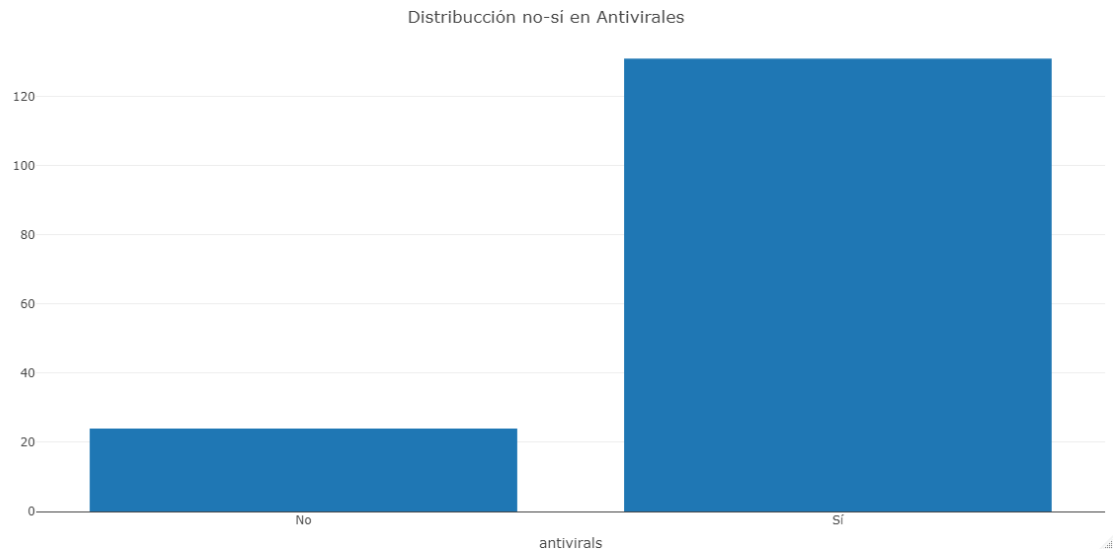

Fig. 10. Distribución del Consumo de Antivirales
Fuente: Elaboración propia

Normalmente los antivirales se administran para reducir la parte inflamatoria y los niveles altos del Virus.

Podemos comprobar en la figura 10 que 24(15,48%) de los pacientes no tomaron antivirales y que 131(84,51%) sí han tomado.

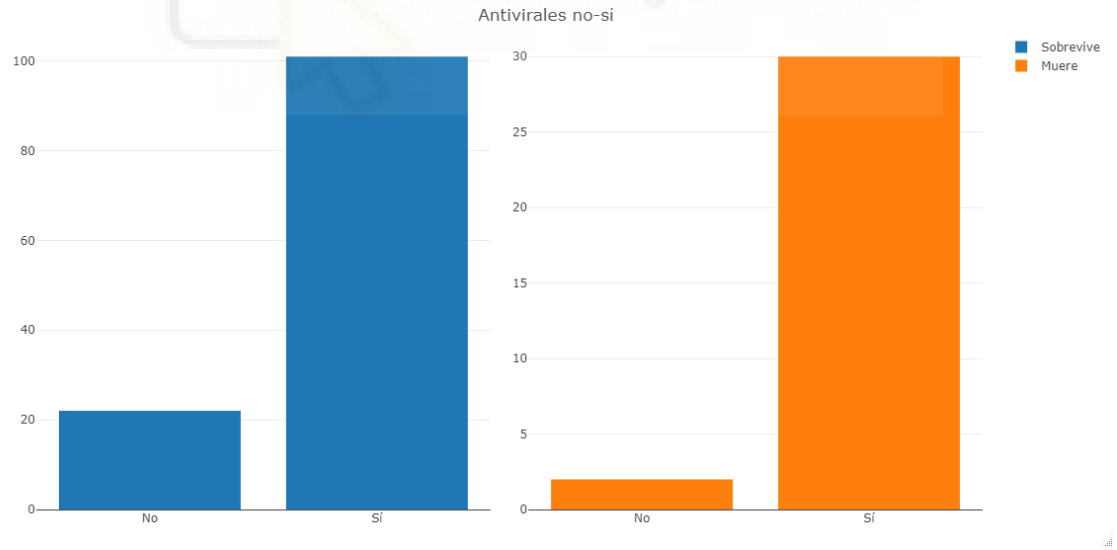

Fig. 11. Distribución del Consumo de antivirales o no relaciona con la variable a predecir
Fuente: Elaboración propia

Observamos en la figura 11 que 30 de los 131(22.90%) que sí tomaron mueren y que 2 de los 24(0.08%) mueren, al haber un número tan pequeño de pacientes que no han tomado antivirales y han muerto parece que esta variable no guarde relación con la variable a predecir.

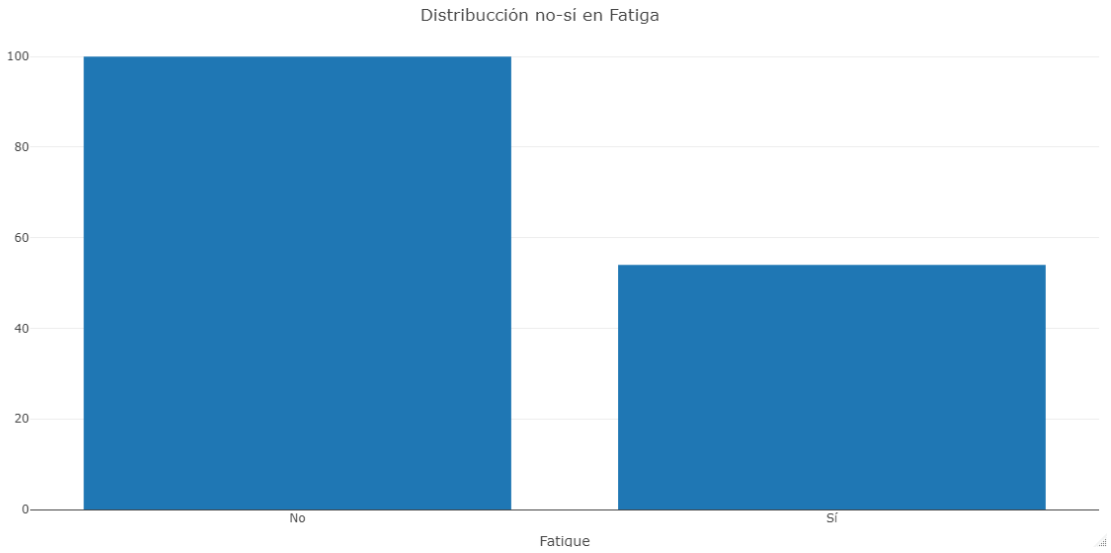

Fig. 12. Distribución si ha tenido Fatiga o no
Fuente: Elaboración propia

La fatiga suele ser unos de los síntomas más frecuentes en los pacientes de hepatitis producida por la misma enfermedad, aunque no suele ser un síntoma determinante debido a que se asocia a enfermedades que producen inflamación.

Se observa en la figura 12, que 100(64,51%) pacientes no padecen síntomas de Fatiga y el 55(35,49%) sí que la padecen.

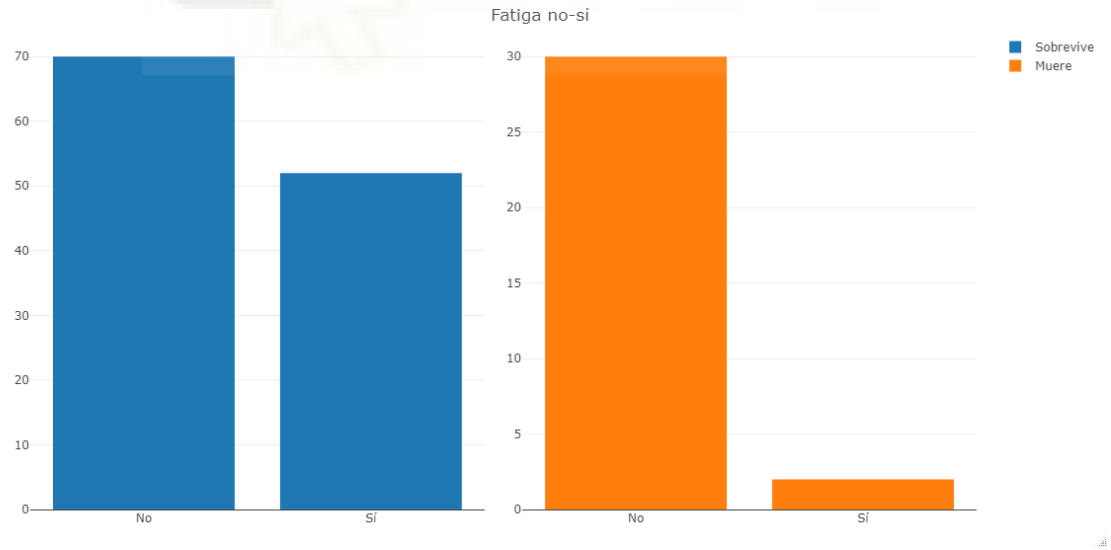

Fig. 13. Distribución de si tenía Fatiga o no relaciona con la variable a predecir
Fuente: Elaboración propia

Por lo que observamos en la figura 13 tener fatiga no parece un síntoma muy importante ya que los vivos sí que muestran este síntoma 52(94.54%), pero los muertos poco de ellos la sufren 3(0.05%).

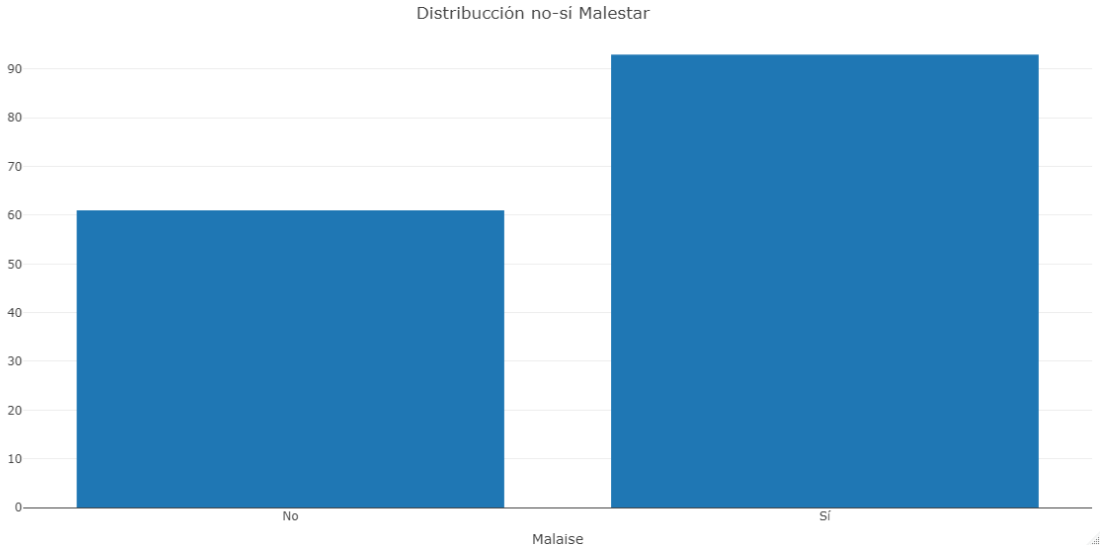

Fig. 14. Distribución de si tenía Malestar o no
Fuente: Elaboración propia

Se observa en la figura 14 que 61(39,35%) pacientes no padecen síntomas de Malestar y el 94(60.65%) sí que la padecen.

Esta enfermad comúnmente suele venir acompañada de grandes malestares debido a la inflamación del hígado que produce unos dolores a la altura de las costillas.

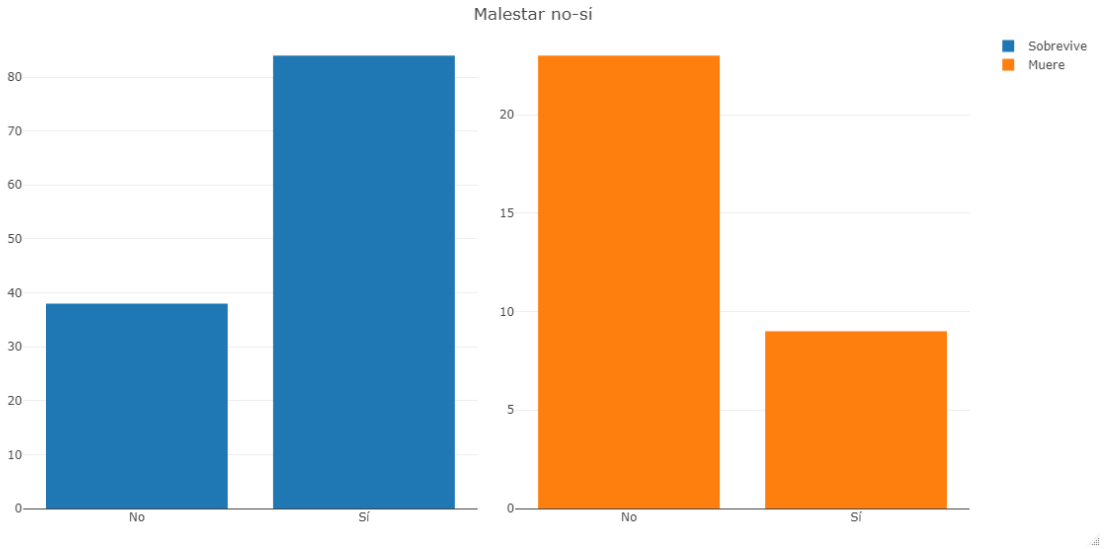

Fig. 15. Distribución de si tenía Malestar o no relaciona con la variable a predecir
Fuente: Elaboración propia

Observamos en la figura 15 que muy pocos de los que tenían malestar mueren 9(9.57%) en cambio el número de muertos que no tienen molestias es mayor 23(37.70%).

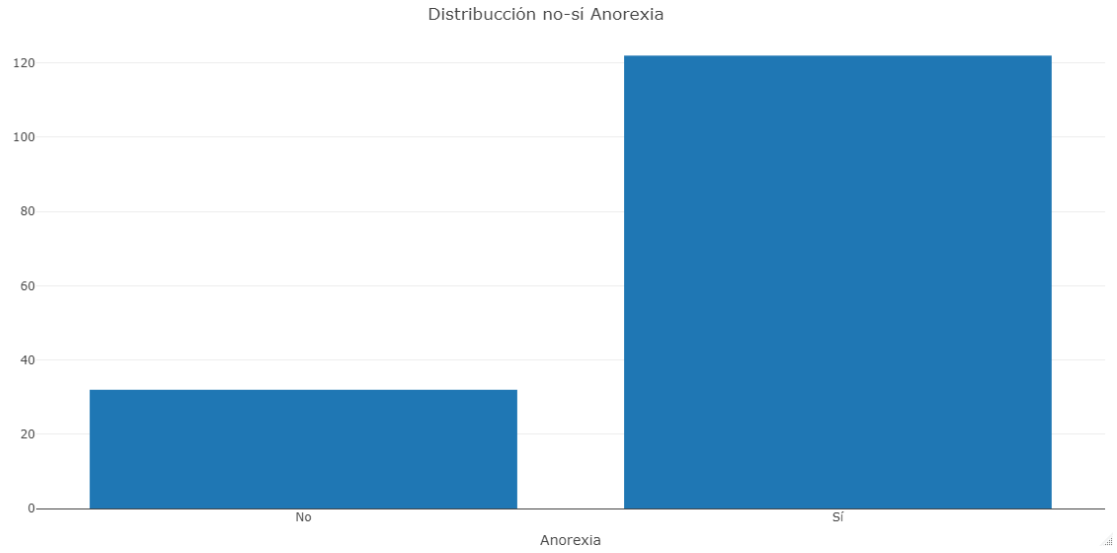

Fig. 16. Distribución de si tenía Anorexia o no
Fuente: Elaboración propia

En algunas ocasiones encontraremos paciente con anorexia que debido a ella sufren un fallo hepático, aunque aún no se saben qué relación guardan estos dos fenómenos.

Se observa en la figura 16 que 32(20.64%) no han sufrido esta enfermedad y que el 123(79,36%) sí que la han padecido.

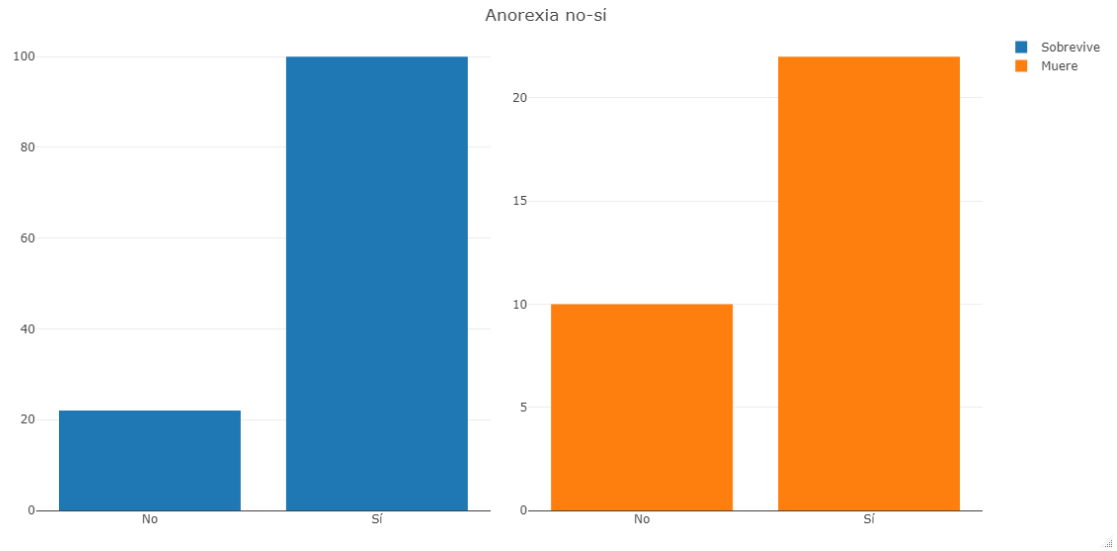

Fig. 17. Distribución de si tenía Anorexia o no relaciona con la variable a predecir
Fuente: Elaboración propia

Por lo que observamos en la figura 17, la enfermedad de la Anorexia puede no estar relacionada con la hepatitis ya que solo 22 de los 122 que han sufrido Anorexia han fallecido esto representa 18.03%

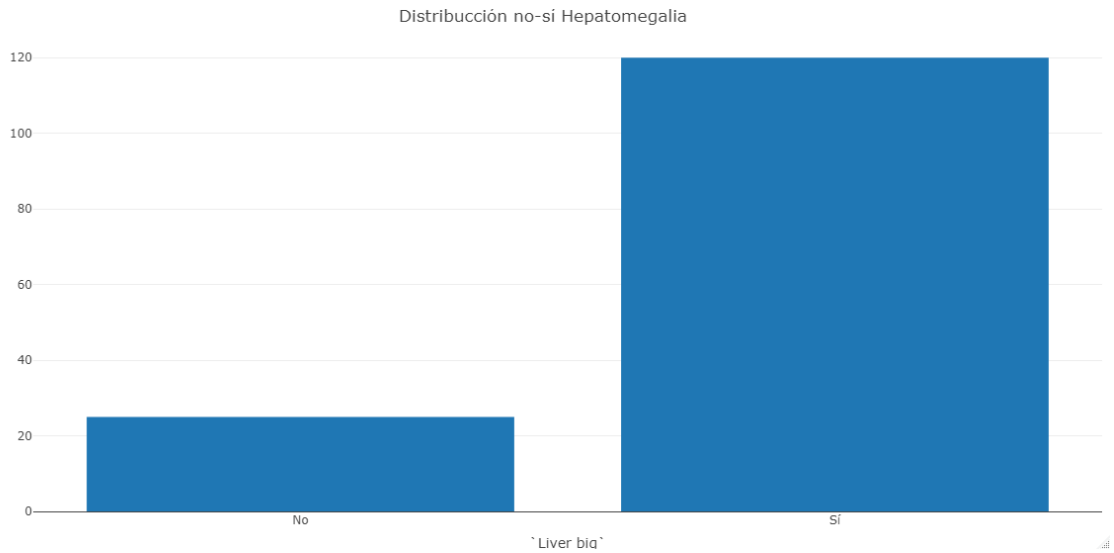

Fig. 18. Distribución de si tenía Hepatomegalia o no
Fuente: Elaboración propia

Se observa en la figura 18 que 25(16,12%) no ha sufrido este síntoma y que 120(82.75%) sí que la han padecido.

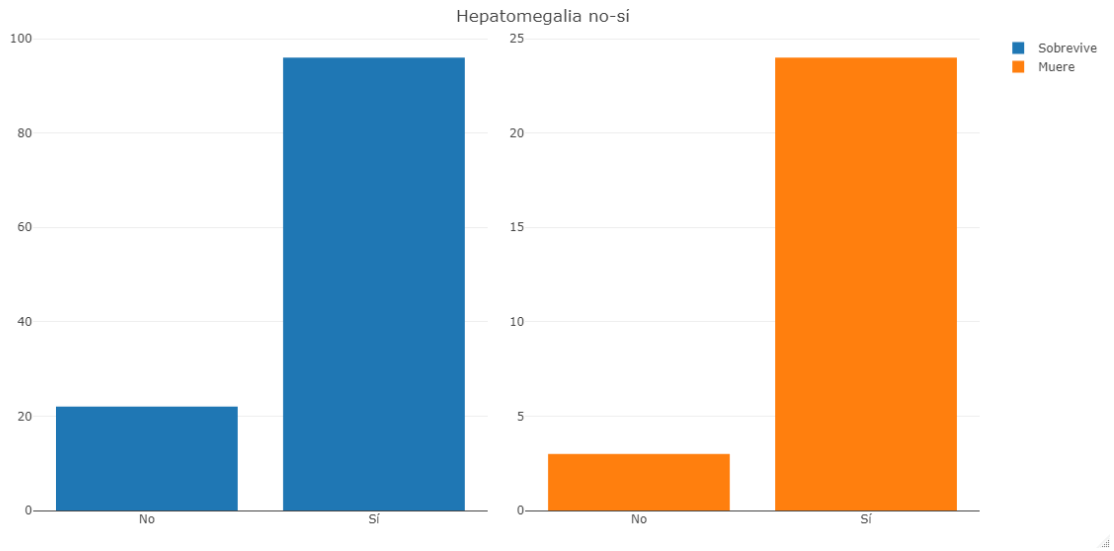

Fig. 19. Distribución de si tenía Hepatomegalia o no relaciona con la variable a predecir
Fuente: Elaboración propia

Observamos en la figura 19 que los pacientes que no fallecieron 24 de 120(20%) sí padecieron Hepatomegalia (Agrandamiento del Hígado producido por la inflamación) y que 3 de 25(12%) no la sufrieron, pero fallecieron.

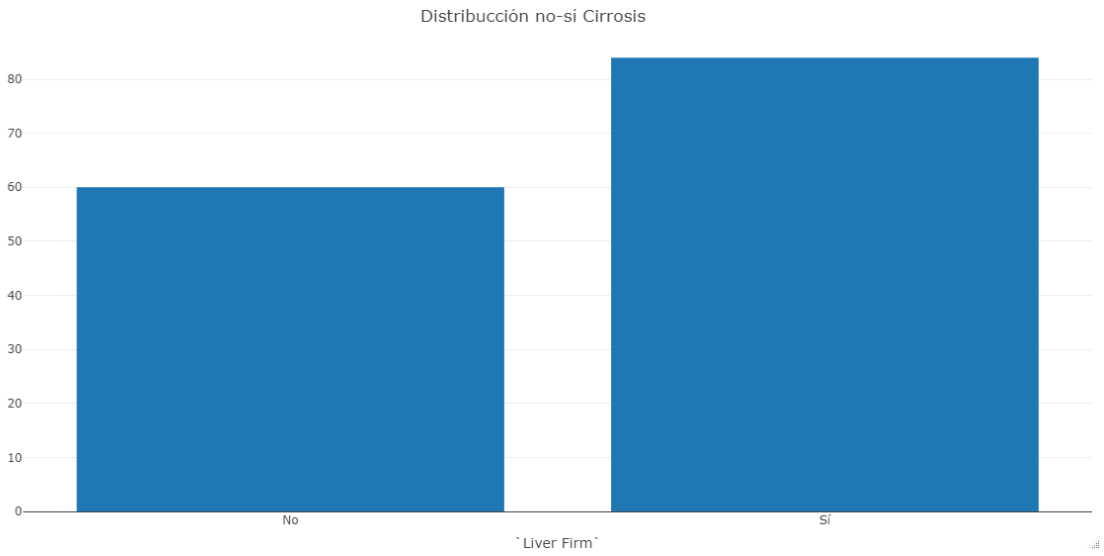

Fig. 20. Distribución de si tenía Cirrosis o no
Fuente: Elaboración propia

Se observa en la figura 20 que un 60(38,70%) no ha sufrido este síntoma y que 95(61,29%) sí que la han padecido.

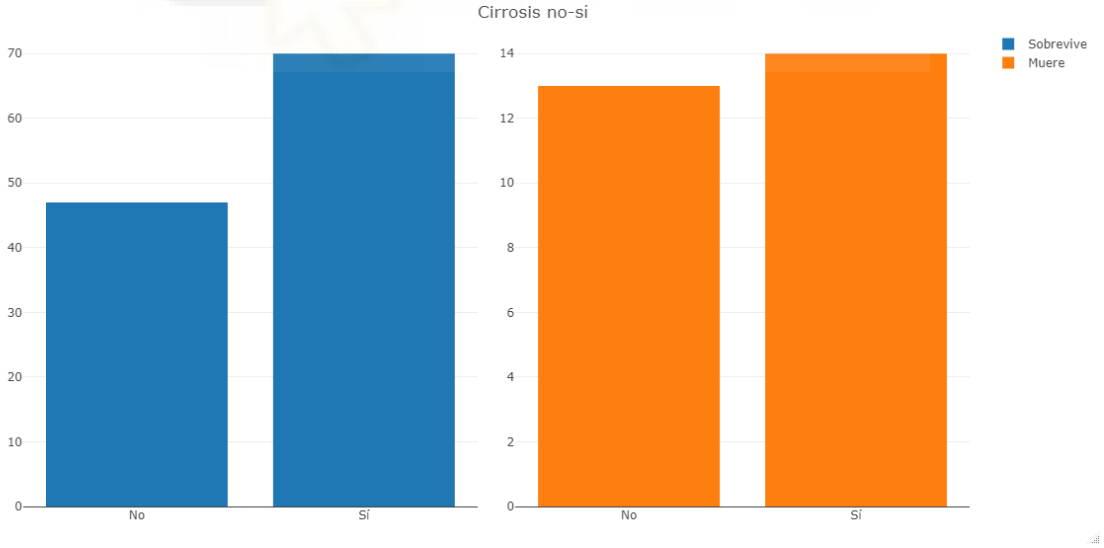

Fig. 21. Distribución de si tenía Cirrosis o no relaciona con la variable a predecir
Fuente: Elaboración propia

Se observa en la figura 21 que los Pacientes que no sufrieron este síntoma la mayoría de ellos sobrevivió 13 de 60(26.67%) y que de los que si sufrieron este síntoma pocos fallecieron 14 de 95(14.73%), no parece que esta variable guarde relación con la variable a predecir.

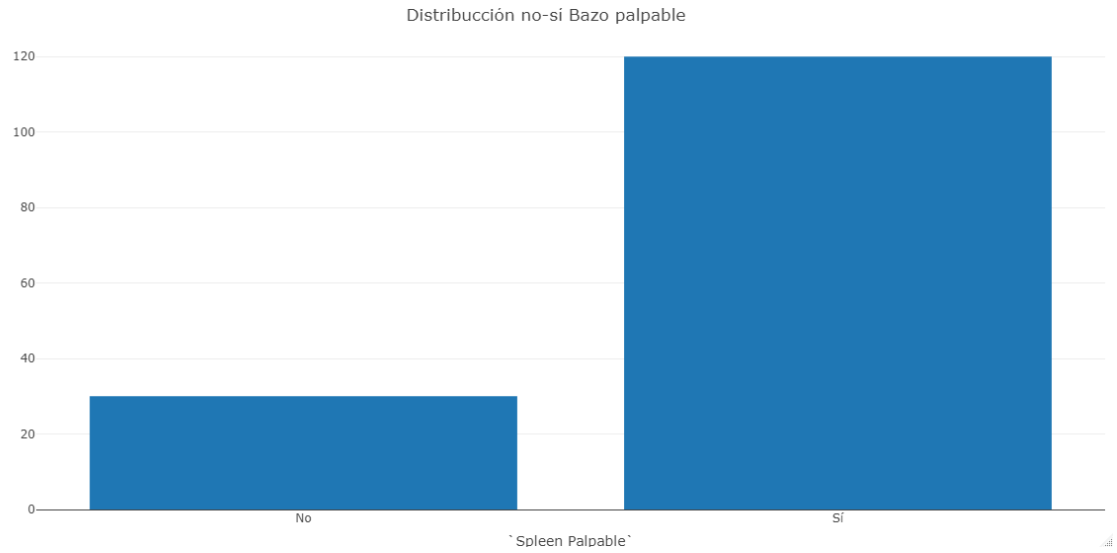

Fig. 22. Distribución de si tenía Bazo palpable o no
Fuente: Elaboración propia

En la figura 22 se observa que 30(19,35%) pacientes no ha sufrido este síntoma y que 125(80,65%) sí que la han padecido.

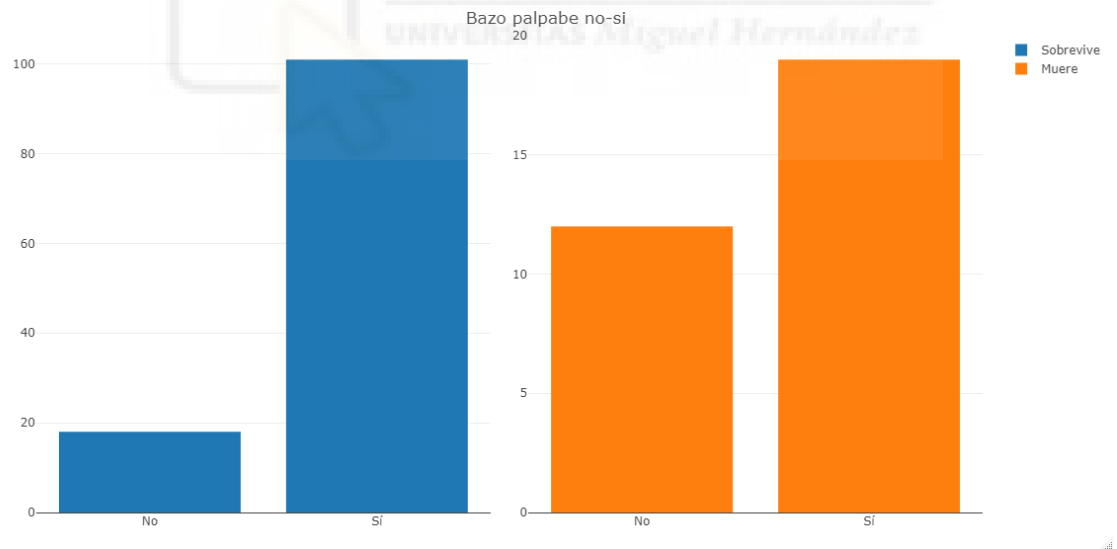

Fig. 23. Distribución de si tenía Bazo palpable o no relaciona con la variable a predecir
Fuente: Elaboración propia

Se observa en la figura 23 que de los pacientes que fallecieron un gran número de los que no tenía esta sintomatología fallecieron 12 de 23(52.17%) y que muy pocos que sí que la tenían fallecieron 19 de 125(15.2) el gran número de pacientes sin presentar esta sintomatología que fallecen puede afectar a la variable a predecir.

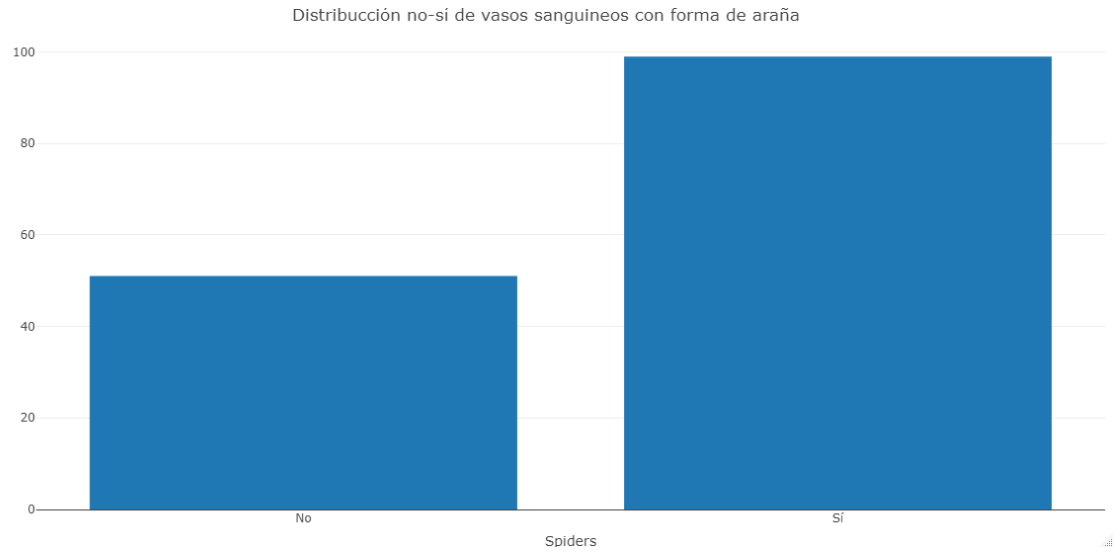

Fig. 24. Distribución de si tenía vasos sanguíneos hinchados en forma de araña en la piel o no
Fuente: Elaboración propia

Este síntoma se relaciona con algunos problemas relacionados con el hígado no simplemente la hepatitis si no también enfermedades como el Cáncer.

Se observa en la figura 24 que 51(34%) no ha sufrido este síntoma y que 99(66%) sí que la ha padecido.

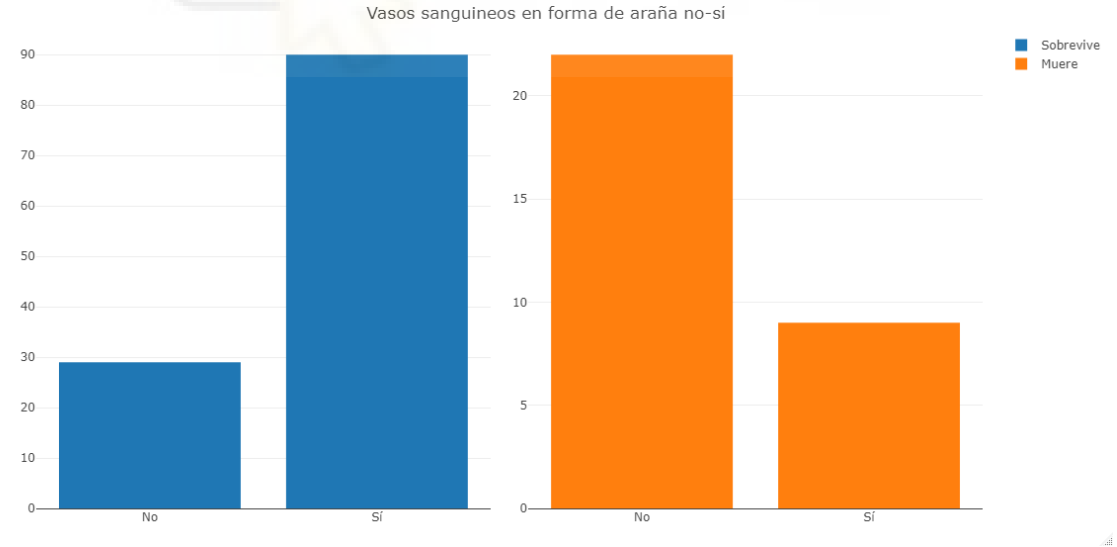

Fig. 25. Distribución de si tenía vasos sanguíneos hinchados en forma de araña en la piel o no relaciona con la variable a predecir
Fuente: Elaboración propia

Se observa en la figura 25 que muchos de los pacientes que no padecieron este síntoma fallecieron 22 de 51(43.13%) y que muy pocos de los que sí padecieron esta sintomatología fallecieron solo el 8.65%.

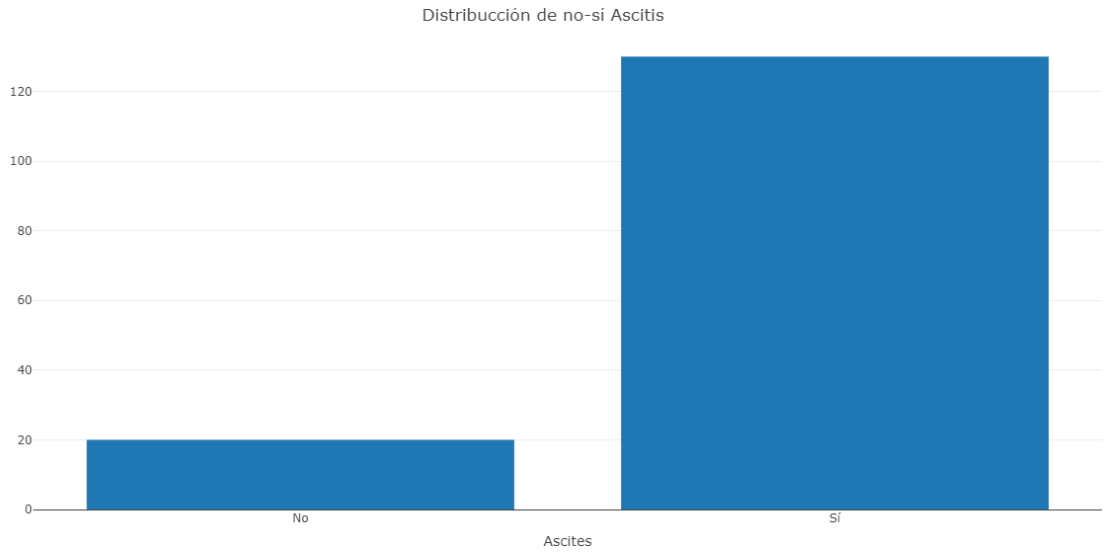

Fig. 26. Distribución de si tenía acumulación anormal de líquidos en el abdomen o no
Fuente: Elaboración propia

La Ascitis es la retención de líquidos en todas las partes del cuerpo, aunque se suele dar en la zona abdominal, esta enfermedad puede llevar a numerosas complicaciones.

Se observa en la figura 26 que 20(12,90%) no han sufrido este síntoma y que 135(87,09%) sí que la ha padecido.

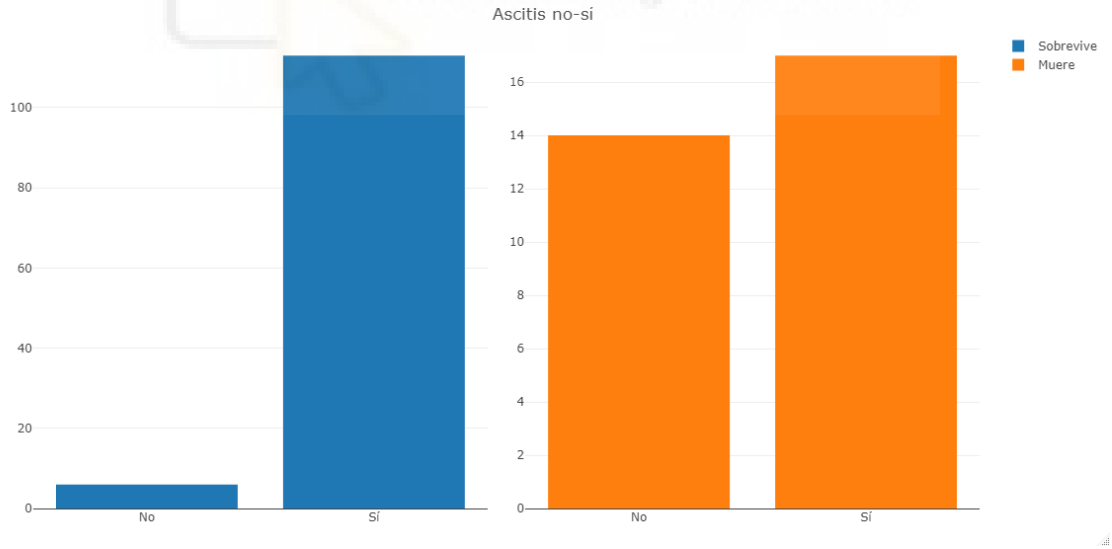

Fig. 27. Distribución de si tenía acumulación anormal de líquidos en el abdomen o no relaciona con la variable a predecir
Fuente: Elaboración propia

Si observamos la figura 27 como pasa con las anteriores sintomatologías muchos pacientes que no tienen el síntoma acaban falleciendo en este caso 14 de 20(70%) y muy pocos que si que la padecen mueren 17 de 135(12.59%).

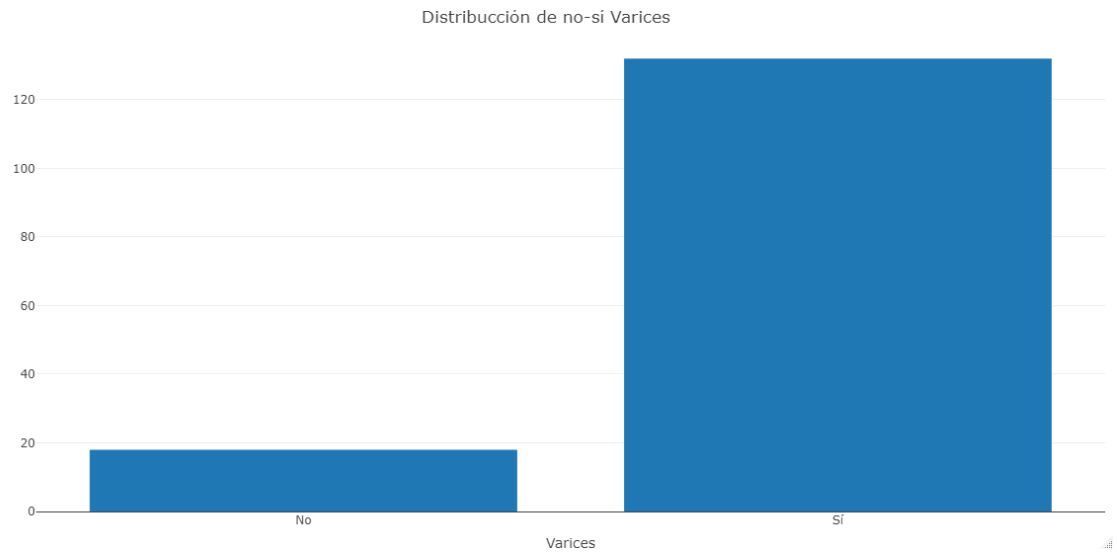

Fig. 28. Distribución de si tenía varices o no
Fuente: Elaboración propia

Estas Varices se refieren a unas venas dilatadas en el tubo que conecta la garganta con el estómago, también llamadas varices esofágicas.

Se observa en la figura 28 que 18(11,61%) no ha sufrido este síntoma y que el 137(88.38%) sí que la han padecido.

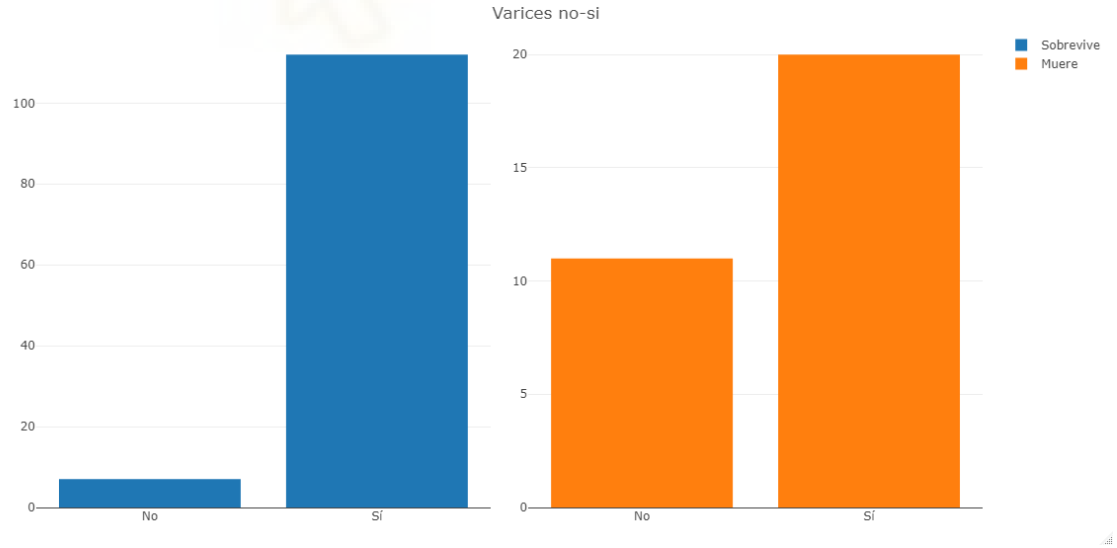

Fig. 29. Distribución de si tenía varices o no relaciona con la variable a predecir
Fuente: Elaboración propia

Se observa en la figura 29 que de los pacientes que fallecieron muchos de ellos no tenían varices 11 de 32(34.3%) muy poca proporción que si tenían esta sintomatología fallecieron 20 de 137(14.59%).

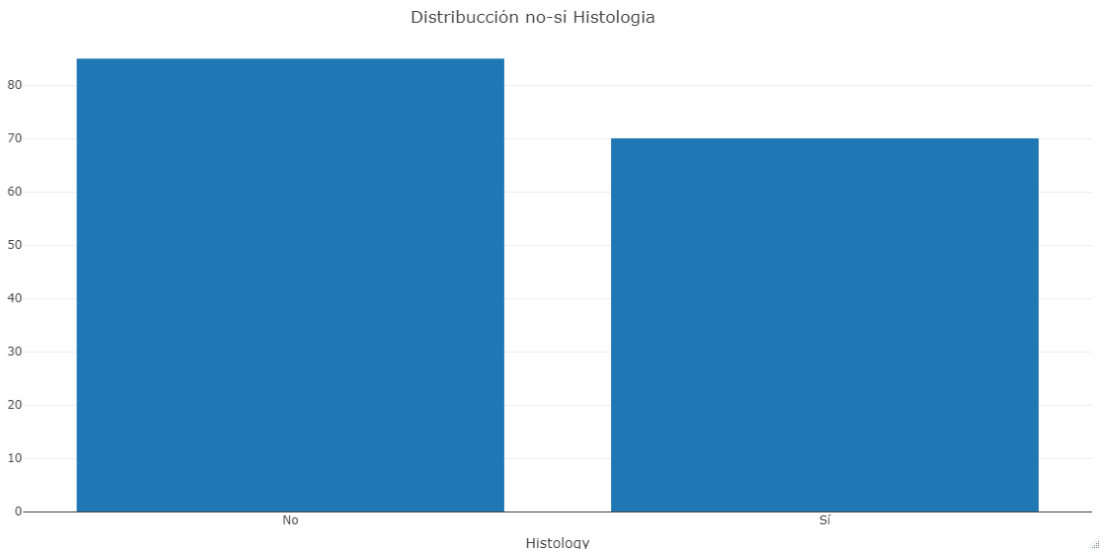

Fig. 30. Distribución de si tenía Histología o no
Fuente: Elaboración propia

Se observa en la figura 30 que 85(54,83%) no la ha sufrido en cambio un 70(45,16%) sí que la han sufrido.

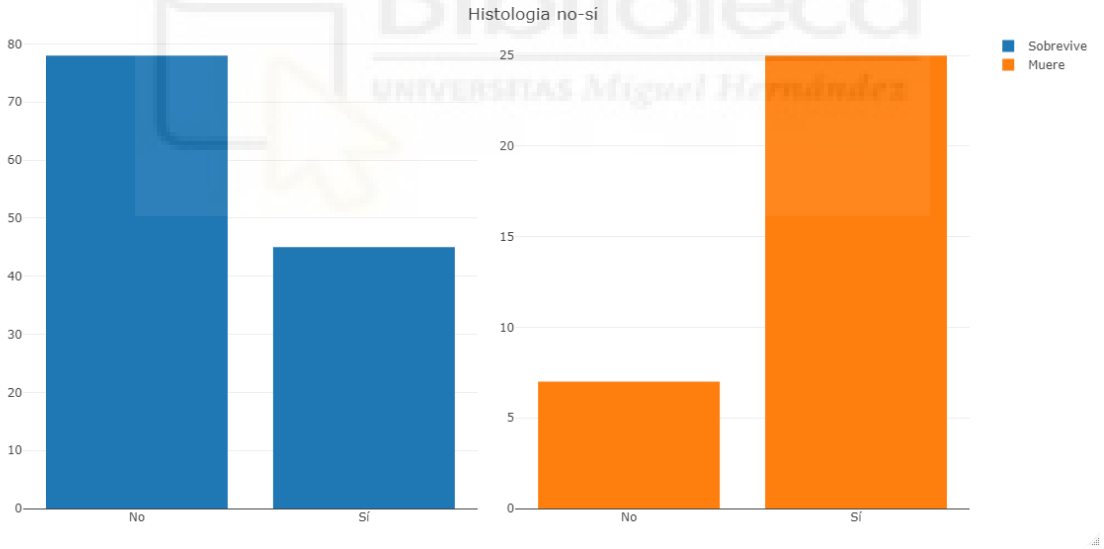

Fig. 31. Distribución de si tenía Histología o no relaciona con la variable a predecir
Fuente: Elaboración propia

Se observa en la figura 31 que de los que si tuvieron histología muchos fallecieron 25 de 70(35.71%) y que muy pocos de los que no la tuvieron fallecieron 7 de 85(8.2%)

A continuación, cambiaremos de tipo de gráfico ya que las siguientes variables son continuas a diferencia de las anteriores que eran categóricas, hemos elegido el grafico de bigote puesto que nos representa muy bien cómo se concentran los datos además de proporcionarnos directamente métricas muy interesantes como la media, median y diferentes Cuartiles

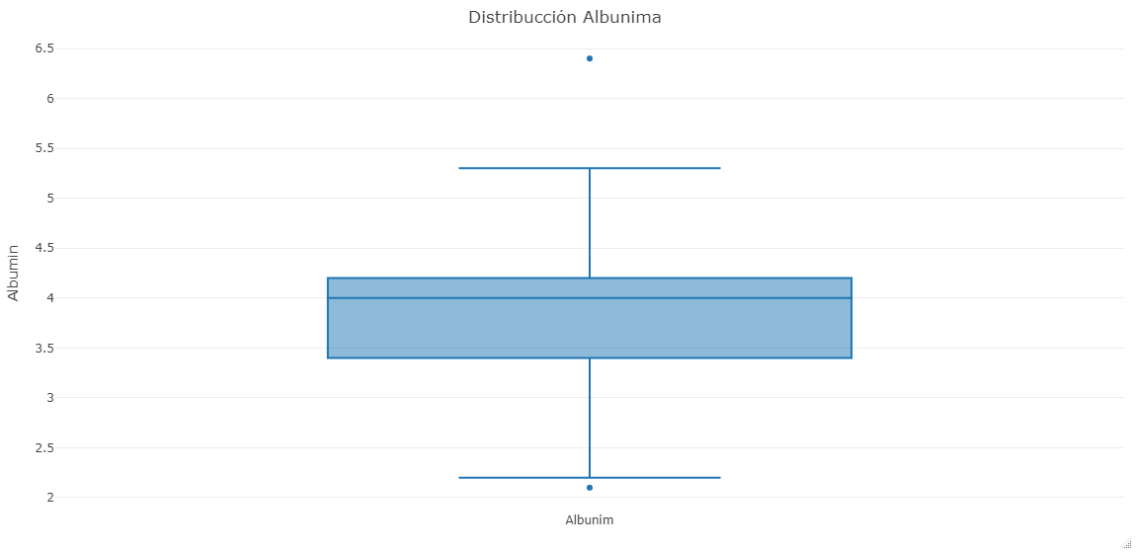

Fig. 32. Distribución de si tenía Albumina o no
Fuente: Elaboración propia

El dato inferior es de 2,10 y el máximo de 6,4 la mayor parte de los datos se concentran en 3 y 4,5 siendo la media 3,8 y la mediana 4.

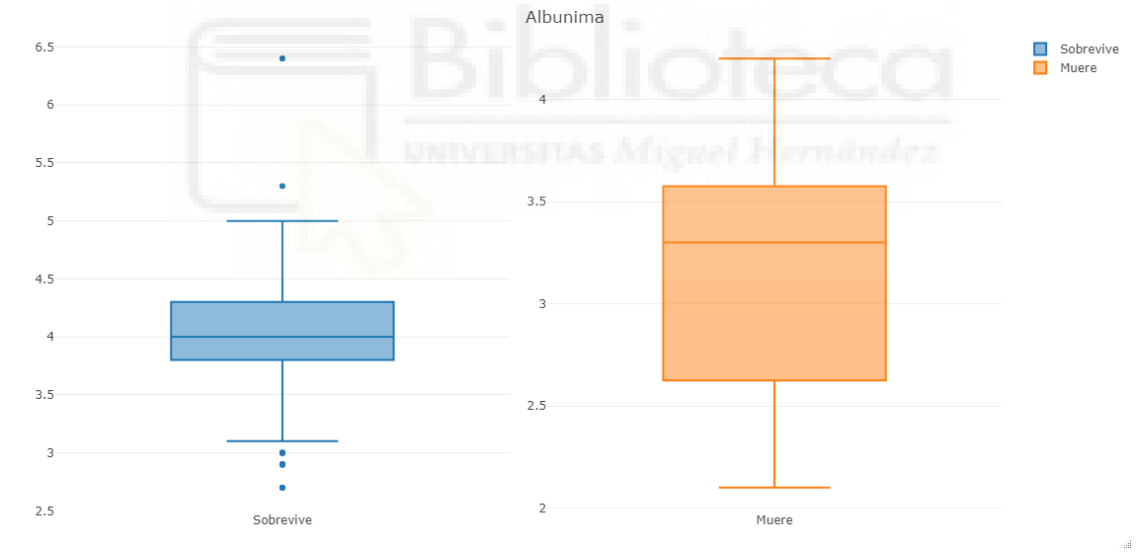

Fig. 33. Distribución de los valores de Albúmina relaciona con la variable a predecir
Fuente: Elaboración propia

Vemos que en pacientes que sobreviven los valores de esta proteína son por lo general más alto que cuando fallecen, ya que cuando fallecen los valores por lo general son más bajos.

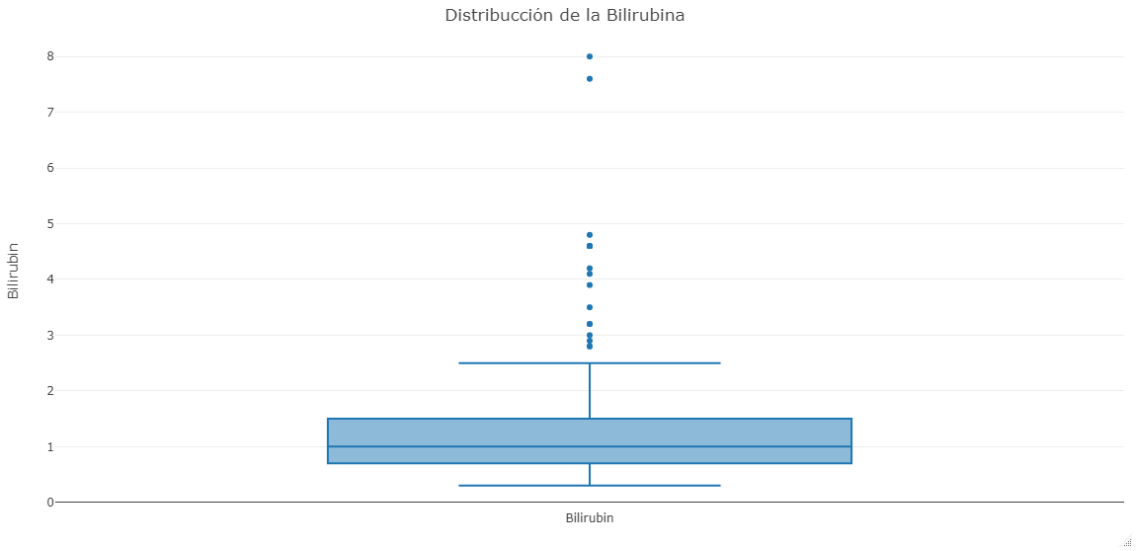

Fig. 34. Distribución de los valores de la Bilirrubina
Fuente: Elaboración propia

La Bilirrubina es un pigmento de color amarillento que se produce de la degradación de la hemoglobina de los glóbulos rojos, cuando se produce un problema en el hígado normalmente se produce una acumulación de este pigmento produciendo un color amarillento en la piel.

Al contrario que en la anterior variable aquí no tenemos una concentración de los valores, por lo que parecen que sean mucho más dispersos.

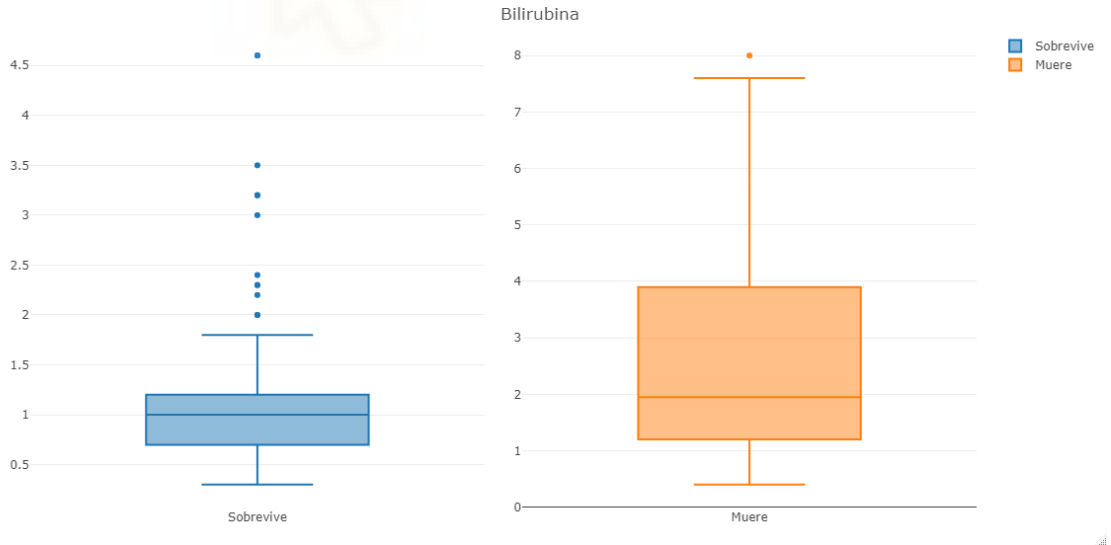

Fig. 35. Distribución de los valores de la Bilirrubina relaciona con la variable a predecir
Fuente: Elaboración propia

Viendo la imagen podemos atrevernos a decir que una elevada concentración de este pigmento puede significar la muerte del paciente cosa que por otro lado parece lógica, como hemos comentado antes una concentración de este pigmento suele significar algún fallo en el hígado.

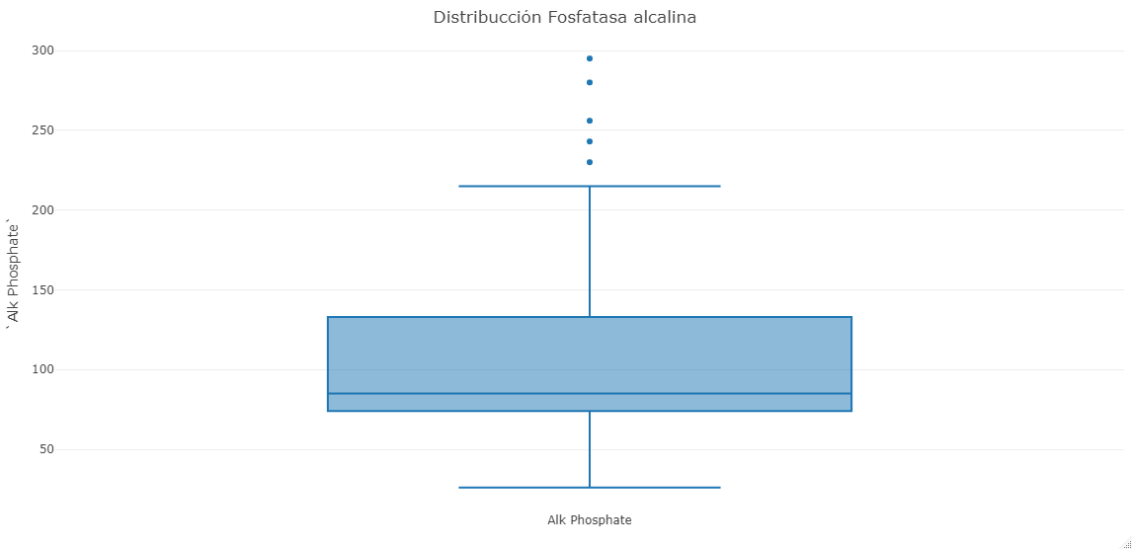

Fig. 36. Distribución de los valores de la Fosfatasa Alcalina
Fuente: Elaboración propia

La fosfatasa alcalina es una encima, responsable de eliminar grupos de fosfatos en algunas moléculas, se suele medir ya que algunas hepatitis como la hepatitis C hace que este valor aumente y si aumenta significaría que hay un bloqueo en el flujo del tracto biliar.

Los valores están concentrados entre 80 y 150 siendo la media de 105,33 y la mediana de 85 el valor mínimo es 26 y el máximo 295.

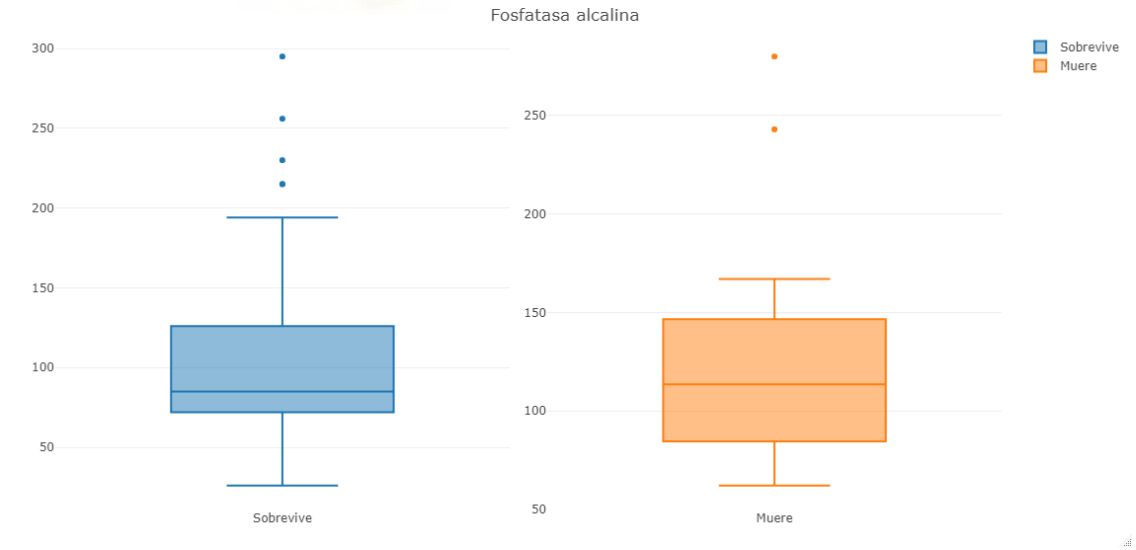

Fig. 37. Distribución de los valores de la Fosfatasa alcalina relaciona con la variable a predecir
Fuente: Elaboración propia

Tanto en pacientes que sobreviven como en los que no sobreviven vemos valores semejantes salvo algunos valores que vemos más altos en los pacientes que sobreviven.

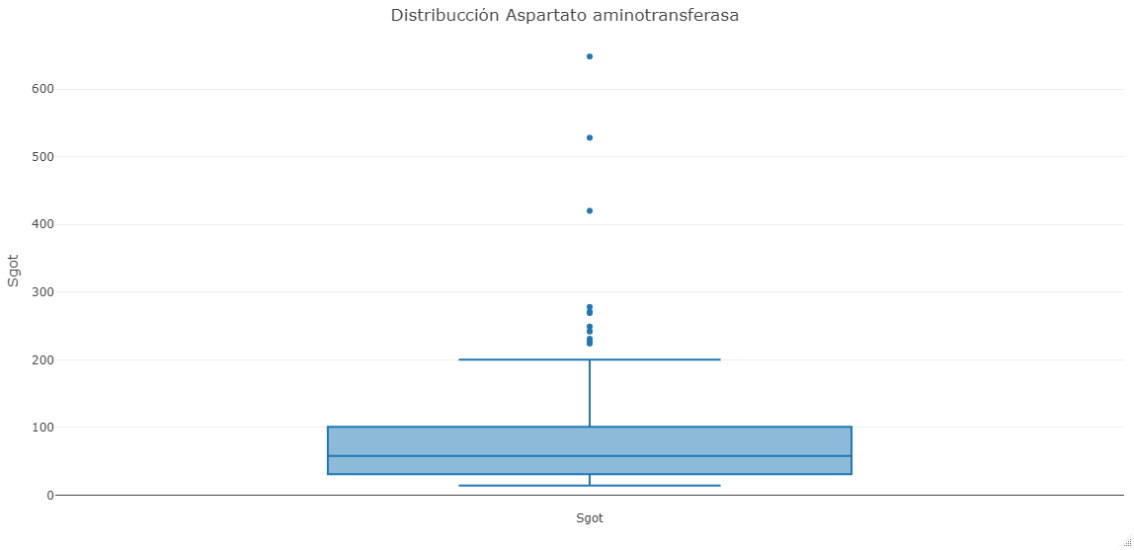

Fig. 38. Distribución de los valores Aspartato aminotransferasa
Fuente: Elaboración propia

El aspartato aminotransferasa es una encima que está en los tejidos como el corazón y el hígado, por lo general cuando la concentración de esta sustancia aparece, es por un daño en el hígado.

Observamos que los valores son muy dispersos están concentrados entre 35 y 100 siendo la media 85,89 y la mediana 58 el valor mínimo es 14 y el máximo 648, observamos que los valores son muy dispersos

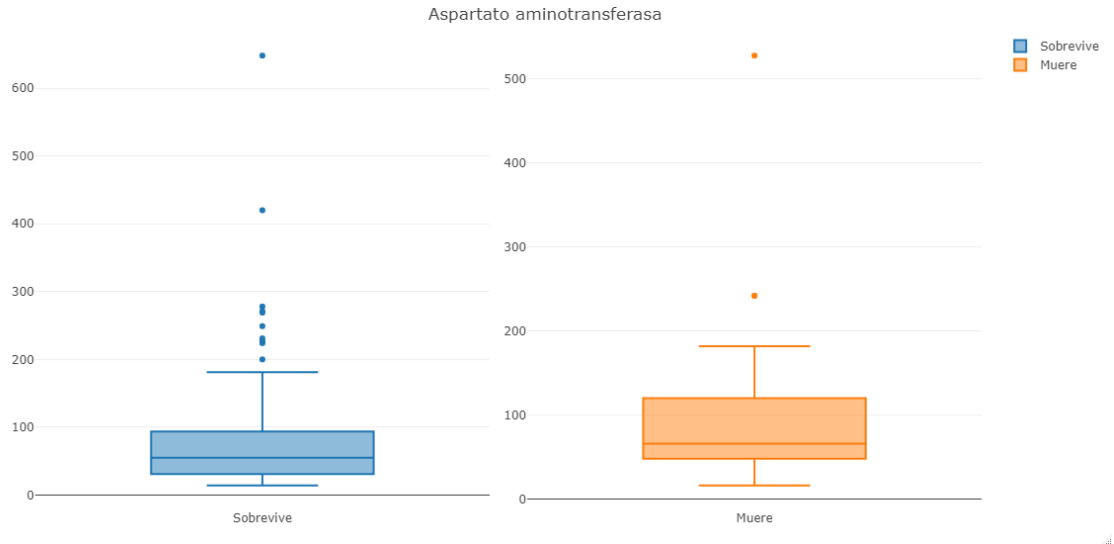

Fig. 39. Distribución de los valores del Aspartato aminotransferasa relaciona con la variable a predecir
Fuente: Elaboración propia

Podemos comprobar que los valores en lo pacientes que no han muerto por lo general son más altos que los que sí, pero no lo vemos de una forma muy significativa por lo que no nos atreveríamos a tomar una decisión basándonos en este gráfico.

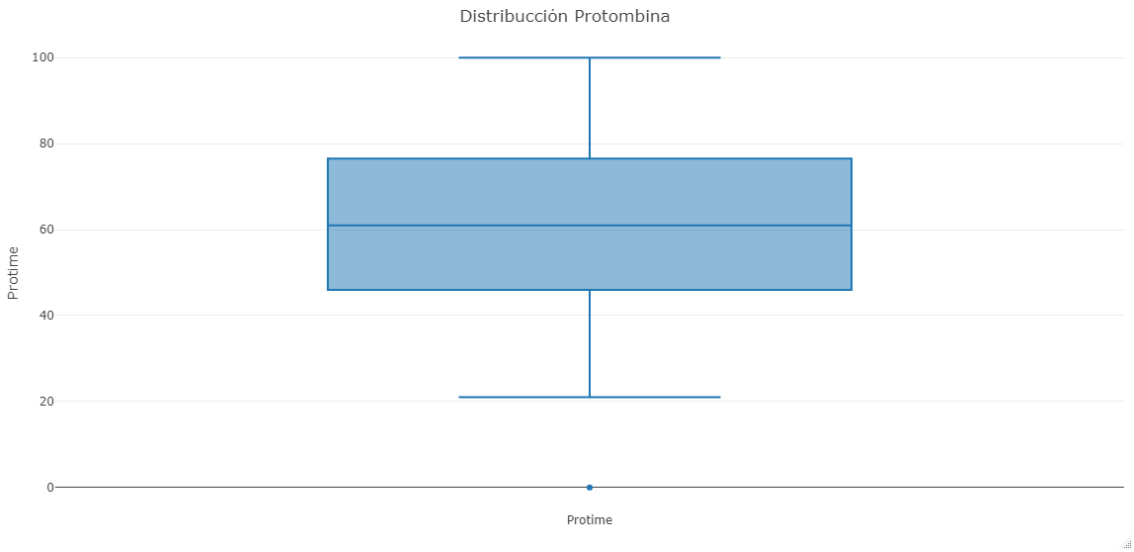

Fig. 40. Distribución de los valores de Protombina
Fuente: Elaboración propia

La Protrombina es una proteína que se produce en el hígado, es muy importante esta proteína puesto que es una de las protagonistas en la correcta coagulación de la sangre

La mayoría de los datos se concentra entre 45 y 80 con una media 61,85 y una mediana 61 un mínimo de 0 y un máximo de 100.

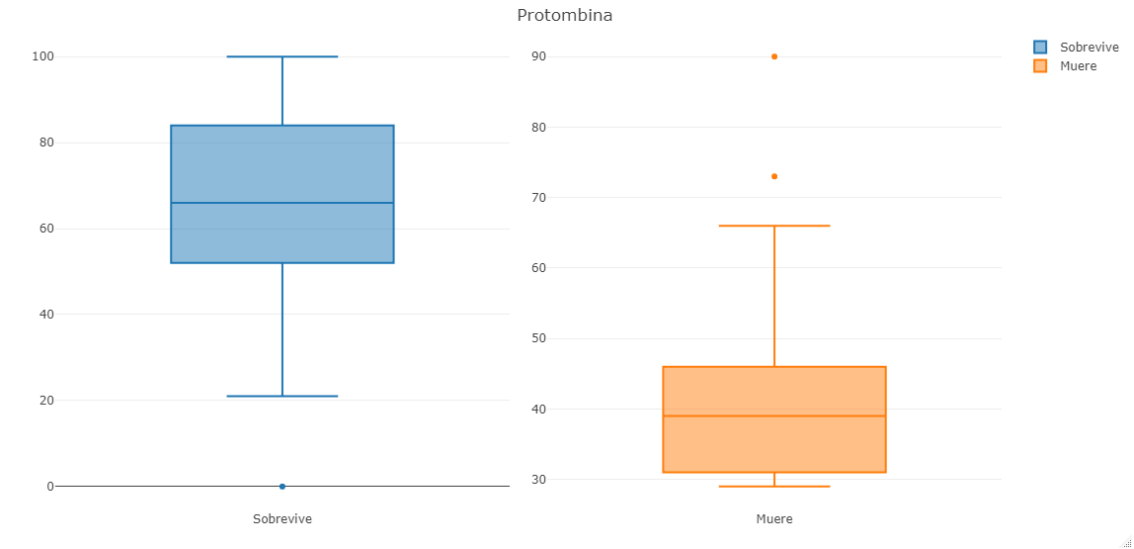

Fig. 41. Distribución de los valores de Protrombina relaciona con la variable a predecir
Fuente: Elaboración propia

Podemos observar muy fácilmente que en los pacientes que mueren tiene una concentración de esta sustancia mucho más baja que en la de los vivos, por lo que sí que parece que esta variable vaya a ser muy importante pare predecir el posible desenlace.

A continuación, observaremos la variable que posteriormente pasaremos a discretizar

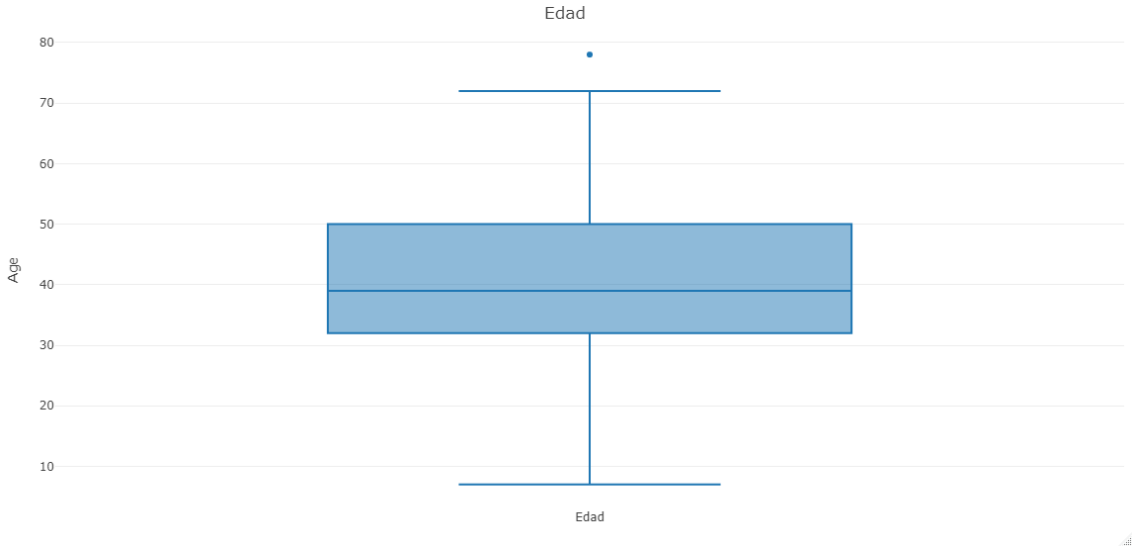

Fig. 42. Distribución de los valores de edad
Fuente: Elaboración propia

Podemos observar que la mayoría de observaciones se concentra entre los valores de 30 y 50 con una media 41,2 y una mediana de 39 un valor mínimo de 7 y un valor máximo de 78.

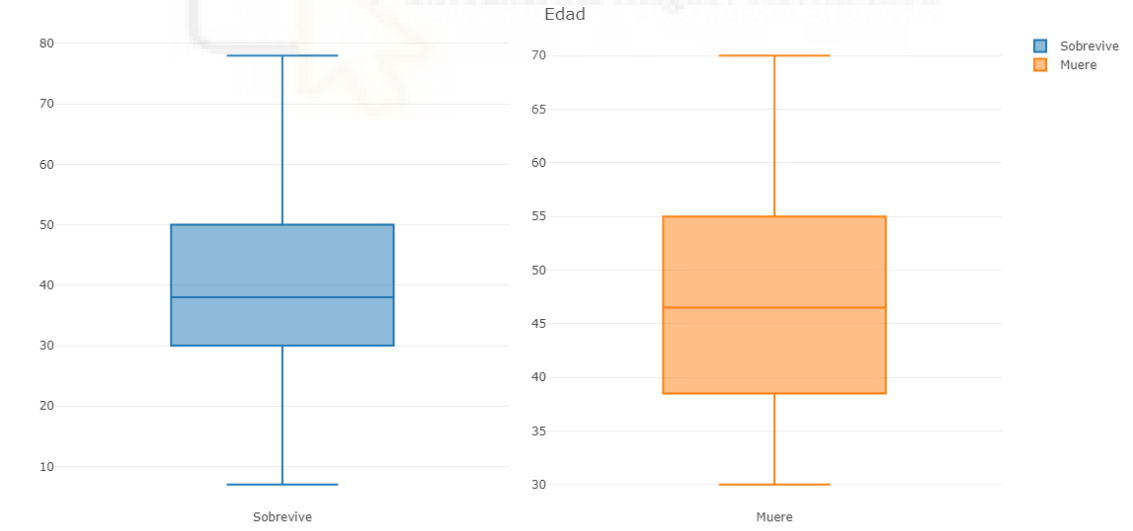

Fig. 43. Distribución de los valores de Edad relaciona con la variable a predecir
Fuente: Elaboración propia

Podemos observar que los pacientes que no sobreviven a la enfermedad y los que sí sobreviven a la enfermedad su edad es bastante parecida, por lo que no parece que la edad pueda ser una variable que determine el resultado final.

No obstante, anteriormente hemos representado la variable edad como una variable continua, pero para sacar información de los gráficos que estamos utilizando necesitamos discretizar esta variable de la forma que hemos explicado en el apartado 3.1

5.3 Factorización

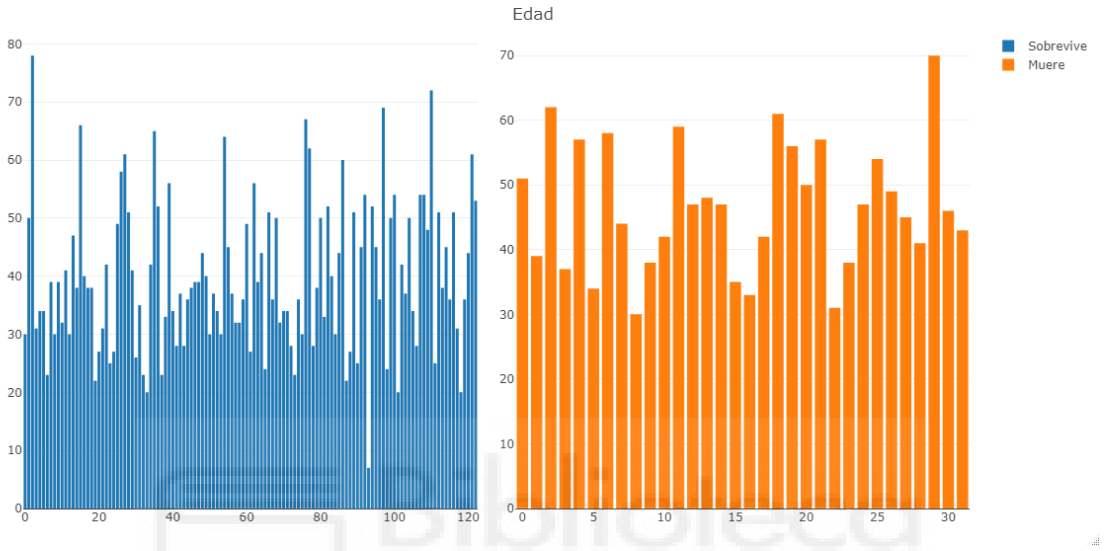

Fig. 44. Distribución de la edad (sin factorizar) relaciona con la variable a predecir
Fuente: Elaboración propia

Podemos observar en el gráfico que es un gráfico muy difícil de interpretar e intentar sacar una conclusión de él, por ello lo más apropiado es factorizar esta variable, para ello utilizaremos la factorización Kmeans que hemos explicado anteriormente, consiguiendo así que todos nuestros datos se concentren en 3 grupos (7 a 33 años, 33 a 48 años y 47 a 78 años)

Para ello hemos utilizado la función “discretize” de la Liberia “Arules”, he utlizado esta librería y no otra, porque tiene la versatilidad de elegir entre diferentes tipos de discretización en ella podemos seleccionar "interval","frecuency","cluster" y "fixed" he utilizado cluster que es la que hace referencia al método que hemos nombrando anteriormente de K-means.

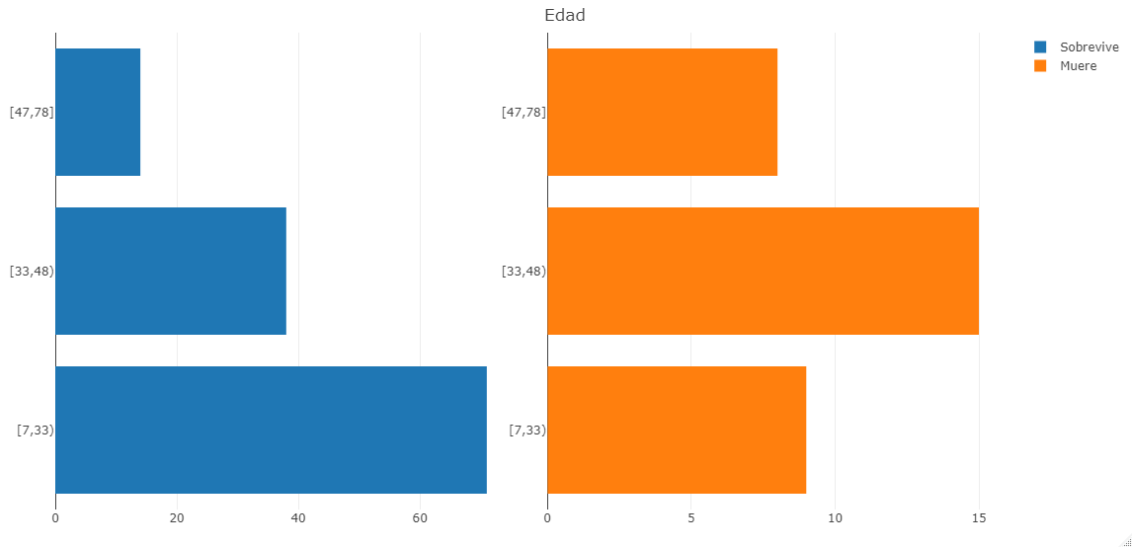

Fig. 45 . Distribución de la edad (Factorizada mediante el método de K-means con la variable a predecir
Fuente: Elaboración propia

Ahora observamos un gráfico mucho más entendible para poder sacar información de él, una conclusión que podemos sacar es que los pacientes de 7 a 33 años mueren en menor proporción que el resto de clases y que nuestra mayor mortalidad se estable en el grupo de 33 a 48 años, no parece que la edad sea una variable de importancia para determinar el desenlace.

5.4 Rankings Variables

Como hemos explicado anteriormente vamos a proceder a seleccionar las variables más importantes mediante el algoritmo Boruta este nos proporcionara las variables que guardan mayor relación con la variable de vivo o muerto, con el objetivo de optimizar el modelo para no utilizar todas además de conseguir una mejor visual del árbol de decisión, aunque este los veremos en el siguiente punto

Para ello seleccionamos la función Boruta del paquete Boruta, seleccionaremos este paquete porque utiliza el método de envoltura como hemos nombrado anteriormente pero además lo utilizaremos por los resultados tan claros que arroja el algoritmo proporcionando cuales son “confirmed” son los atributos que son mejores que las sombras, después nos proporciona las variables “Tentative” esto es que el algoritmo con consigo tomar una decisión, y “Rejected” que son las variables rechazadas.

Tabla 3 : Resultado de las variables con su “decisión” sin discretizar la edad
Fuente: Elaboración propia

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
Protime	11.856.229.987.162	120.645.065.945.019	869.724.610.739.256	144.801.726.357.421	1	Confirmed
Albumin	659.384.799.147.803	674.616.649.085.127	373.596.188.684.163	88.539.461.535.562	0.96969696969697	Confirmed
Histology	618.821.753.082.263	616.347.202.194.781	35.584.966.402.701	9.306.402.934.178	0.96969696969697	Confirmed
Ascites	547.564.479.237.168	549.601.179.500.238	279.647.347.556.711	905.460.036.637.101	0.939393939393939	Confirmed
Varices	398.254.335.850.875	392.829.714.714.506	152.348.738.573.159	658.081.307.767.576	0.787878787878788	Confirmed
Anorexia	387.994.672.070.919	388.998.819.769.705	182.247.097.790.733	598.257.910.753.061	0.787878787878788	Confirmed
`Alk Phosphate`	32.371.693.542.875	340.122.835.422.709	-0.00443310248121266	536.092.799.640.581	0.575757575757576	Tentative
`Liver big`	233.537.219.119.129	237.460.044.649.642	-0.12359630197882	422.359.363.917.859	0.414141414141414	Tentative
Spiders	232.247.020.123.855	22.533.217.513.123	-0.0221844386896633	590.372.208.278.407	0.383838383838384	Tentative
Bilirubin	165.984.470.047.417	161.354.002.376.451	-0.53145844208696	410.749.358.952.739	0.0707070707070707	Rejected
Malaise	152.678.654.138.983	157.735.726.208.454	-128.728.543.221.352	354.297.675.098.296	0.111111111111111	Rejected
Sex	0.77249788722165	0.70873700674957	-0.439552341932405	283.396.726.949.927	0	Rejected
Fatigue	0.537967359369352	0.855855157651482	-119.973.136.767.564	231.139.399.604.465	0	Rejected
Age	0.22709069469277	0.157994231618235	-172.237.239.669.406	235.500.430.793.085	0	Rejected
steroid	0.0518008897919485	-0.0804812786160139	-18.637.342.752.531	184.760.641.214.393	0	Rejected
`Liver Firm`	0.0120703350657991	-0.0210531566284422	-110.066.809.137.473	0.988475155828346	0	Rejected
Sgot	-0.104473309867279	-0.123948323230931	-199.499.660.141.443	139.920.241.324.221	0	Rejected
`Spleen Palpable`	-0.493452416598996	-0.429162831741796	-190.228.968.510.723	203.041.161.192.141	0	Rejected
antivirals	-0.901119642929607	-0.856272119185367	-193.519.089.557.596	0.553085285292799	0	Rejected

Una vez hemos lanzado el algoritmo observamos que las variables seleccionadas son (Protime, Albumin, Histology, Ascites, Anorexia, Varices) ahora comprobamos si la discretización que hemos explicado anteriormente afecta a la selección de atributos.

Tabla 4: Resultado de las variables con su "decisión" discretizada la edad
Fuente: Elaboración propia

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
Protime	108.278.051.844.103	107.797.511.157.437	804.313.379.149.802	133.836.443.355.363	0.98989898989899	Confirmed
Albumin	618.129.637.306.282	620.881.118.053.642	431.133.399.257.188	898.270.458.472.855	0.939393939393939	Confirmed
Histology	559.768.450.276.839	563.197.737.977.283	325.636.144.040.359	852.456.576.281.683	0.94949494949495	Confirmed
Ascites	551.226.358.214.341	553.287.949.633.881	332.105.737.415.196	745.129.550.901.914	0.919191919191919	Confirmed
Anorexia	390.023.710.726.292	395.823.495.202.122	173.688.957.562.379	586.605.556.580.446	0.747474747474748	Confirmed
Varices	338.845.129.182.584	343.385.512.946.645	102.883.413.697.985	547.886.129.140.037	0.646464646464647	Confirmed
`Alk Phosphate`	22.683.675.686.117	230.716.277.632.314	-0.12954293206403	465.867.952.432.623	0.424242424242424	Tentative
Bilirubin	2.077.300.672.942	21.334.354.189.504	-196.894.204.710.888	492.036.895.205.414	0.444444444444444	Tentative
Sex	178.204.991.836.975	178.442.682.770.408	-0.710848664671381	364.214.845.678.248	0.212121212121212	Rejected
`Liver big`	174.869.670.537.558	172.991.260.518.615	-0.209542842785368	376.035.460.783.561	0.252525252525253	Rejected
Spiders	16.269.253.391.525	171.130.009.950.766	-131.810.964.739.006	425.617.034.472.606	0.242424242424242	Rejected
Malaise	13.890.146.453.223	14.844.548.305.803	-167.523.368.182.284	358.559.626.819.341	0.171717171717172	Rejected
Age	0.714696631047347	0.499506522092481	-171.049.599.739.158	279.845.770.678.826	0.0707070707070707	Rejected
Fatigue	0.547569590260567	0.802796019766833	-0.898989963233654	171.441.630.965.777	0	Rejected
Sgot	0.433243115215937	0.225589059346793	-0.559054077265141	183.693.220.901.369	0	Rejected
steroid	-0.00615929892035737	0.221959308869547	-152.958.607.273.665	146.926.770.098.807	0	Rejected
`Spleen Palpable`	-0.121449574245897	-0.0203393579791367	-147.929.737.098.363	182.553.364.168.984	0	Rejected
`Liver Firm`	-0.370322428965544	-0.376477943347367	-218.072.343.028.648	0.812278939086232	0	Rejected
antivirals	-0.721421774277871	-0.795947901500307	-151.483.089.938.157	0.845584153102861	0	Rejected

Como observamos en la tabla la factorización de la variable edad, ha servido al algoritmo para descartarla.

Ahora probaremos a hacer una segunda exploración con las variables "Tentative" para comprobar si finalmente las podemos elegir o las descartamos, para ello utilizaremos la función TentativeRoughFix que se encuentra propiamente dentro de la librería de Boruta.

Tabla 5 : Resultado de la segunda exploración
Fuente: Elaboración propia

attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
Protime	111.047.146.438.987	109.890.025.519.201	859.012.672.391.324	140.798.960.303.497	1	Confirmed
Albumin	620.295.102.729.343	626.672.513.946.589	371.598.294.324.106	922.422.295.174.306	0.97979797979798	Confirmed
Histology	565.501.741.819.997	565.925.031.705.645	262.654.205.608.434	866.173.013.632.242	0.95959595959596	Confirmed
Ascites	52.711.675.421.195	520.165.937.925.635	271.118.539.326.574	766.627.441.304.723	0.95959595959596	Confirmed
Anorexia	366.203.271.087.926	371.760.326.075.328	130.611.005.535.844	566.642.515.666.428	0.777777777777778	Confirmed
Varices	330.957.044.470.387	343.533.732.512.463	-0.00363114153327594	659.154.251.844.045	0.696969696969697	Confirmed
`Alk Phosphate`	236.976.988.223.402	234.263.131.288.392	-0.626734809220637	515.392.335.173.165	0.484848484848485	Rejected
Bilirubin	188.362.121.008.404	176.658.373.809.077	-0.717158712124335	441.214.650.493.319	0.434343434343434	Rejected
`Liver big`	187.258.093.394.251	188.457.949.593.016	-15.692.596.381.493	40.415.521.592.745	0.404040404040404	Rejected
Spiders	16.789.053.953.318	167.056.749.487.083	-160.959.173.865.315	425.705.769.959.672	0.171717171717172	Rejected
Sex	153.520.850.003.522	179.596.096.592.538	0.0146057071490384	329.424.858.313.994	0.0505050505050505	Rejected
Malaise	0.912624637107028	102.142.275.740.801	-109.160.461.220.389	250.007.772.247.053	0.0505050505050505	Rejected
Age	0.910421828549151	0.966445533181959	-13.641.223.757.823	322.288.368.588.477	0.0505050505050505	Rejected
Fatigue	0.641575150067759	0.464717316179677	-0.189380054736933	342.669.877.990.966	0.0101010101010101	Rejected
Sgot	0.0227454636307124	-0.202243023622191	-109.070.456.241.277	204.642.579.379.086	0	Rejected
steroid	-0.0682117963877942	-0.305848633751856	-181.888.372.792.529	185.753.721.094.273	0	Rejected
`Liver Firm`	-0.321542976993779	-0.460301443419664	-175.240.839.637.963	180.243.888.266.836	0	Rejected
`Spleen Palpable`	-0.374618770761956	-0.0306401167429987	-252.396.262.936.227	0.444756681349813	0	Rejected
antivirals	-128.597.437.162.862	-162.963.810.871.679	-3.607.852.469.301	0.526988458694094	0	Rejected

Finalmente seleccionaremos las 6 variables que se encuentran confirmada que son (Protime, Albunim, Histology, Ascites, Anorexia y Varices), por que el modelo será el siguiente:

$$Class \sim Protime + Albunim + Histology + Ascites + Anorexia + Varices$$

5.5 Modelo predictivos

A continuación, definiremos unos conceptos para entender los modelos predictivos:

- "Recursive Partitioning and Regression Trees"(RPart) es el algoritmo que utilizaremos , funciona dividiendo de forma recursiva esto quiere decir que que proporcionara árboles que muestran la sucesión de reglas que deben seguirse para llegar a un valor.
  La regla de división elegida es aquella que sigue el criterio de reducir la entropía, la entropía es la cantidad de aleatoriedad que hay en cada uno de los nodos.
- Accuracy es el porcentaje del total de aciertos de nuestro modelo. Una confusión muy normal es no saber distinguir entre estos dos conceptos Accuracy y Precisión, ambos reflejan qué tan cerca está una medida en este caso la predicha de un valor real, para diferenciarlos Acurracy es la diferencia respecto al valor real en cambio la Precisión tiene que ver con la repetibilidad, aunque estas repeticiones estén lejos del valor real.

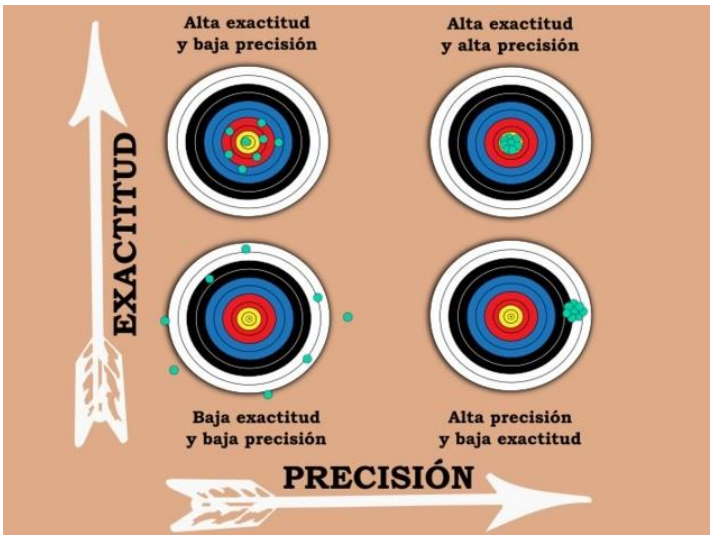

Fig. 45. Distribución precisión-Exactitud
Fuente: (Fernandes, s.f.)

- Matriz de confusión: índica como se distribuye el error, siempre intentaremos escoger el modelo donde el error se reparte de forma equitativa.
- Árbol de clasificación es un modelo de clasificación que permite predecir el valor de una variable categórica con una estructura tipo árbol.

A continuación, veremos la representación que obtenemos con el modelo seleccionando todas las variables de las que disponemos

5.5.1 Modelo utilizando todas las variables de las que disponemos.

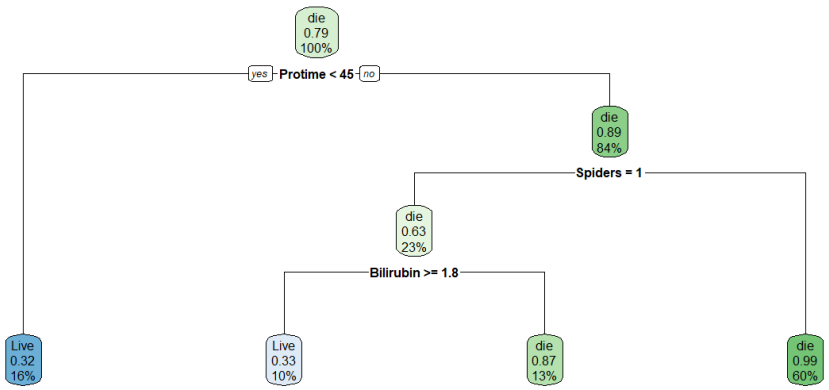

Fig. 46. Árbol de decisión con todas las variables
Fuente: Elaboración propia

Podemos observar (Fig.46.) los pacientes con la Protrombina menor que 45 y sobreviven representan el 16 % de los datos en cambio con la Protombina superior a 45 y que fallecen representan el 84% de este 84% solo el 10% de los pacientes con una bilirrubina mayor que 1.8 sobreviven, el resto fallecen 74%, como podemos observar intentando interpretar esta figura es algo complicado.

Obtenemos un 84.62% de Acurracy, una cifra muy buena además como podemos comprobar en la imagen siguiente el error no está repartido de forma equitativas, pero casi.

Tabla 6: Matriz de confusión con el modelo completo
Fuente: Elaboración propia

	DIE	LIVE
DIE	6	4
LIVE	2	27

5.5.2 Modelo utilizando las variables seleccionadas por el algoritmo Boruta

Class ~ Protime + Albunim + Histology + Ascites + Anorexia + Varices

Fig. 47. Modelo con las variables seleccionadas
Fuente: Elaboración propia

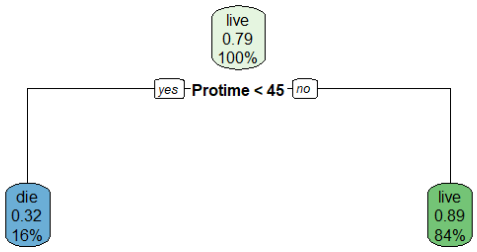

Fig. 48. Árbol de decisión con las variables seleccionadas (Fig. 46.)
Fuente: Elaboración propia

A diferencia de la figura 47 en esta figura vemos un árbol de decisión mucho más sencillo y que solo depende de una variable, que es la Protrombina si es menor 45 el 16% Sobrevive y si la Protombina es mayor que 45 el 84% de los pacientes fallece.

Obtenemos un 84.62% de Acurracy, la misma cifra que con el anterior modelo y solo seleccionando las 5 variables que nos ha proporcionado el algoritmo Boruta además conseguimos que el error este equilibrado en los dos tipos de errores.

Tabla 7: Matriz de confusión con el modelo simplificado (selección de atributos)
Fuente: Elaboración propia

	DIE	LIVE
DIE	5	3
LIVE	3	28

6. CONCLUSIONES Y PROPUESTAS

Podemos Concluir del apartado anterior, la importancia de una selección de atributos a la hora de conseguir un modelo más eficiente, además de conseguir un modelo que visualmente sea más fácil de sacar conclusiones para el estudio, y en este caso (no siempre) hemos conseguido la misma precisión que con el modelo completo además de conseguir que los errores se repartan de la misa forma, es decir, hemos mejorado modelo.

Una vez los hemos comparado podemos contestar a la pregunta ¿es importante la selección de atributos? Como he explicado anteriormente, deberemos responder con rotundo sí, puesto que hemos mejorada el modelo completo, hemos conseguido una visualización más simple, hemos conseguido un menor gasto computacional, es decir utilizamos menos recursos de nuestro ordenador.

Además, hemos demostrado la importancia de la discretización de las variables, en nuestro estudio ha sido la edad, pero puede ser el peso, la altura etc. Ya que estas variables por si sola no nos proporcionan información y con unos simples descriptivos podemos llegar a algunas conclusiones previas.

Podemos dar por satisfechos los objetivos que me había propuesto:

- ✓ Demostrar los beneficios de utilizar una selección de atributos para simplificar el modelo.
- ✓ Conseguir Predecir el posible desenlace de nuevos pacientes o que factores son críticos para cada desenlace.
- ✓ Aplicar los conocimientos y técnicas aprendidas en el Grado de Estadística Empresarial y conseguir una versión práctica de estas técnicas.
- ✓ Profundizar más en las técnicas que hemos mencionado anteriormente y ponerlas en práctica en un ejemplo de la vida real.

Por otro lado, a partir de estudio se abren muchas posibilidades:

- Aplicar otros modelos predictivos y comparar con la precisión obtenida en este estudio.
- Ampliar la selección de atributos en distintas enfermedades, no solo para obtener mejores modelos, si no, para descartar variables que quizás no deberíamos monitorizar (Ahorrar recursos).
- Una vez hemos detectado las variables importantes nos podríamos centrar en automatizar el proceso de recogida de datos de esas variables consiguiendo así una mayor muestra, con ello podríamos construir un sistema de alarmas con ciertos parámetros que nos indiquen la próxima de la muerte del paciente, para de alguna forma intentar remontar la enfermedad.

7. REFERENCIAS

Bay, S. D. (s.f.). *University of California, Irvine.* Obtenido de https://dl.acm.org/doi/pdf/10.1145/347090.347159

Dougherty, J., Kohavi, R., & Sahami, M. (1995). *Stanford Univesity.* Obtenido de http://robotics.stanford.edu/users/sahami/papers-dir/disc.pdf

Fernandes, A. Z. (s.f.). *Difenciador*. Obtenido de https://www.diferenciador.com/diferencia-entre-exactitud-y-precision/

Gonzalez, L. (04 de 01 de 2019). *aprendeIA*. Obtenido de https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/

Gonzalez, L. (30 de 04 de 2020). *aprendeIA*. Obtenido de https://aprendeia.com/reduccion-de-la-dimensionalidad-machine-learning/

Mehrotra, K., & Mohan, C. (11 de Febrero de 2009). *Surface Syracuse University.* Obtenido de https://core.ac.uk/download/pdf/215692923.pdf

NC State University. (s.f.). Obtenido de https://www.ise.ncsu.edu/fuzzy-neural/wp-content/uploads/sites/9/2022/01/k-means-clustering-Wikipedia-1.pdf

R. Calvo Hernández, J. C. (2006). *Scielo*. Obtenido de https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1130-01082006000700014

RDocumentation. (s.f.). Obtenido de https://www.rdocumentation.org/packages/arules/versions/1.7-3/topics/discretize

RDocumentation. (s.f.). Obtenido de https://www.rdocumentation.org/packages/Boruta/versions/7.0.0/topics/Boruta

Revista de investigación Industrial Data. (2021). Obtenido de https://www.redalyc.org/journal/816/81669876013/html/#:~:text=En%20este%20proyecto%2C%20se%20emplea,por%20el%20matem%C3%A1tico%20Stuart%20P.

Román, V. F. (2009). *Scielo*. Obtenido de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1130-01082009001000010#:~:text=En%20cuanto%20a%20los%20antivirales,tenofovir%20(an%C3%A1logos%20de%20nucle%C3%B3tidos).

Tan, K., Teoh, E., Yu, Q., & Goh, K. (2009). *journals.elsevier.* Obtenido de http://tarjomefa.com/wp-content/uploads/2017/09/7680-English-TarjomeFa.pdf

U.S department of veterans affairs. (s.f.). Obtenido de https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-alkaline-phosphatase.asp

Voltas, B., Ferrer, J. C., Sánchez, C., Marco, C., Sanz, P., & García, L. (2018). *Scielo*. Obtenido de https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112018000100245