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

Laboratorio 4 - Árboles de Decisión

Integrantes:	Ariel Aaron Argomedo Madrid
	Marcelo Israel Guzmán Gutiérrez
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco
Ayudante:	Gustavo Hurtado

11 de Noviembre de 2022

Tabla de contenidos

1. Introducción

En esta instancia de laboratorio se plantea la utilización de árboles de decisión, para la obtención de reglas que permitan asociar los diferentes parámetros del problema de la hepatitis, para de este modo, lograr comprender de mejor manera con una herramienta más avanzada las diferentes cualidades que diferencian a pacientes vivos de aquellos muertos. A grandes rasgos, un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que puede ser utilizado para tareas de clasificación como regresión IBM (2022). Dentro de las múltiples ventajas que ofrecen los árboles de decisión se pueden encontrar, por ejemplo: es válido tanto para variables numéricas como cuantitativas, son simples de entender e interpretar, la respuesta del algoritmo es fácilmente justificable a partir de la lógica booleana con la que trabaja Lastra (2020).

Considerando que en el laboratorio anterior, también se utilizó una herramienta de obtención de reglas de asociación, se espera que los resultados obtenidos en esta instancia sean parecidos o incluso mejores. Características como la Hypoalbuminemia, debería ser un factor presente en aquellos pacientes de la clase Muerto, de acuerdo a los resultados ya obtenidos previamente.

El siguiente estudio estará conformado en primer lugar por un marco teórico, el cual busca explicar brevemente los conceptos más importantes. A continuación, se presenta en detalle los resultados obtenidos para la obtención del árbol de decisión, y, por último, se realizará una comparación con las reglas de asociación previamente obtenidas para estudiar la consistencia de los resultados y analizar si se produjo alguna mejora.

1.1. Objetivos

1. Extraer conocimiento del problema de la Hepatitis, mediante el uso del software de la librería C50 para la obtención de un árbol de decisión adecuado.

2. Analizar y comparar los resultados obtenidos con la literatura encontrada y lo expuesto en la teoría.

3. Comparar los resultados obtenidos con la instancia de laboratorio anterior.

2. Marco teórico

2.1. Árbol de decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado, es decir, basa su aprendizaje en un conjunto de datos de entrenamiento previamente etiquetados. Este tipo de algoritmo se utiliza, tanto para tareas de clasificación como regresión. El árbol de decisión, tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja IBM (2022).

El proceso de aprendizaje de un árbol de decisión emplea la estrategia "divide and conquer" mediante una búsqueda exhaustiva para identificar los puntos de división óptimos dentro de un árbol IBM (2022).

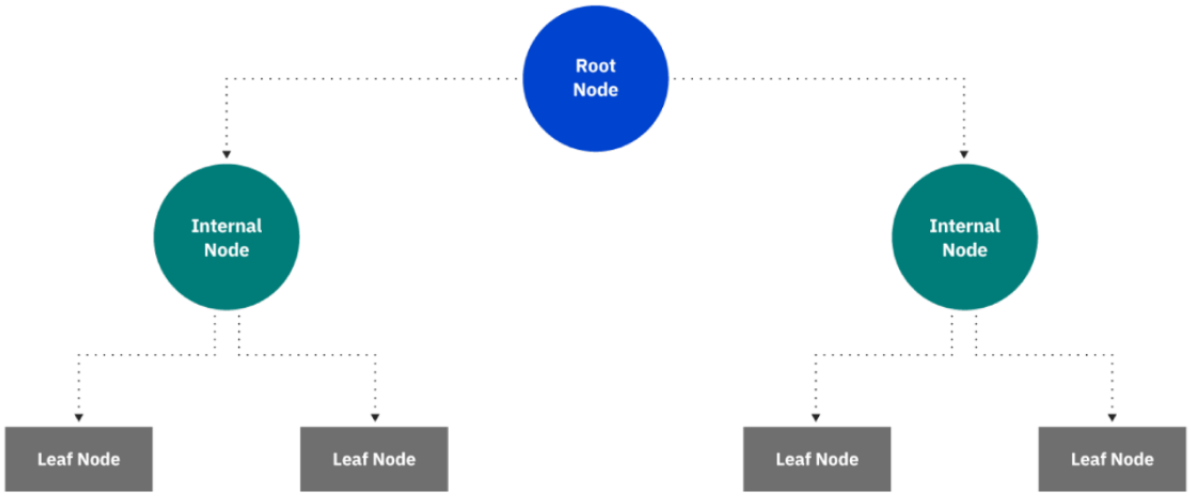

Figura 1: Estructura árbol de decisión IBM (2022).

2.2. Métricas de calidad

La mejor elección para medir el desempeño de un algoritmo de árbol de decisión, es una matriz de confusión. Cada fila de la matriz representa a un dato o sujeto real, mientras que la columna representa a un sujeto predicho. La matriz, está compuesta por verdaderos negativos (True negative), falsos positivos (False positive), falsos negativos (False negative) y verdaderos positivos (True positive) Johnson (2022).

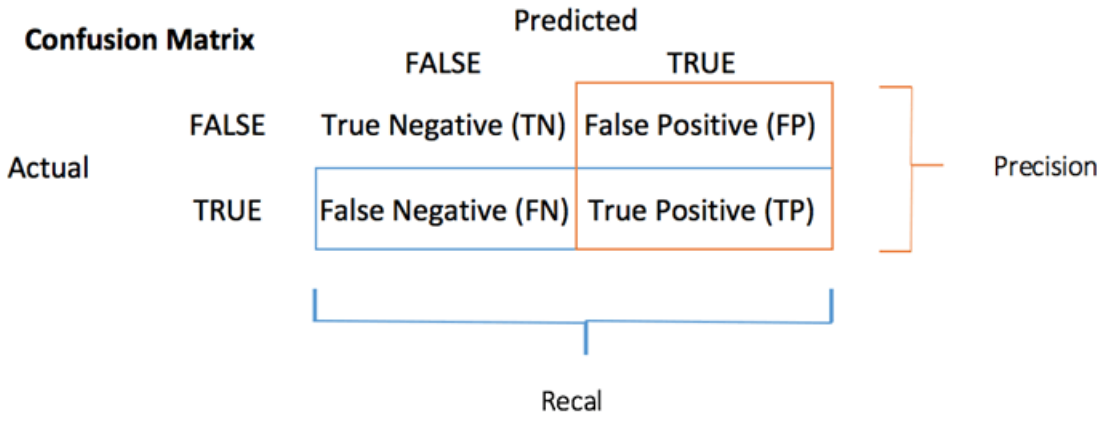

Figura 2: Matriz de confusión Johnson (2022).

Johnson (2022), sugiere realizar un test de precisión con los datos de la matriz de confusión:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

2.3. Reglas de asociación

Una regla de asociación es definida por Rodrigo (2018) como una implicación del tipo "si X entonces Y" ($X \rightarrow Y$), donde X e Y son *itemsets* o *items* individuales. A la izquierda de la regla se encuentra el antecedente o *left-hand-side* (LHS) y al lado derecho está el consecuente o *right-hand-side* (RHS).

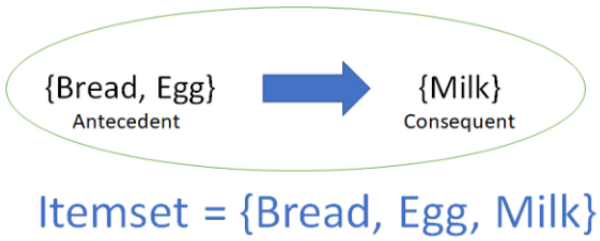

Figura 3: Ejemplo regla de asociación Garg (2018).

3. Obtención del árbol

3.1. Pre-procesamiento

Para la obtención del árbol de decisión, se realizó el mismo pre-procesamiento realizado previamente en las entregas anteriores, es decir, se eliminaron outliers, missing values y se adoptaron formatos correctos para cada tipo de variable.

3.2. Árbol de decisión

El árbol de decisión de la Figura 4, fue obtenido con ayuda de la función "C5.0" de la librería con el mismo nombre. Los datos fueron previamente trabajados (se realizó un proceso de limpieza), y a continuación se particionaron en dos grupos, el grupo de entrenamiento (train) conformado por un subset aleatorio que contiene el 70 % de los datos originales, y un grupo de prueba (test) con el 30 % restante de los datos. La elección de la partición de los datos fue escogida a través de un proceso de prueba y error, y además se consideró la investigación realizada por Nguyen et al. (2021), en la que se afirma que la capacidad predictiva de los modelos de machine learning es afectada en gran medida por los ratios train/test, donde el 70/30 presentó el mejor desempeño de los modelos. Además, se escogió como parámetro de semilla al valor 222, el cual fue el que presentó los mejores resultados del árbol de decisión, en cuanto a *accuracy*, *sensitivity* y *specificity*. Por último, a través del árbol de decisión obtenido, se logran reconocer 8 reglas, las cuales se presentan en la Tabla 1.

Antecedente	Consecuente	n/m	Confianza	Lift
Spleen_Palpable = Yes_Spleen_Palpable, Bilirubin >1.2, Albumin <= 3.7	class Muerto	9	0.909	4.6
Age >28, Ascites = Yes_Ascites, Sgot <= 80	class Muerto	7	0.889	4.5
Ascites = Yes_Ascites	class Muerto	15/5	0.647	3.3
Ascites = No_Ascites, Bilirubin <= 1.2	class Vivo	60/2	0.952	1.2
Age <= 28	class Vivo	16	0.944	1.2
Spleen_Palpable = No_Spleen_Palpable, Ascites = No_Ascites	class Vivo	71/4	0.932	1.2
Albumin >3.7	class Vivo	70/4	0.931	1.2
Spiders = No_Spiders	class Vivo	70/6	0.903	1.1

Cuadro 1: Reglas obtenidas mediante árbol de decisión.

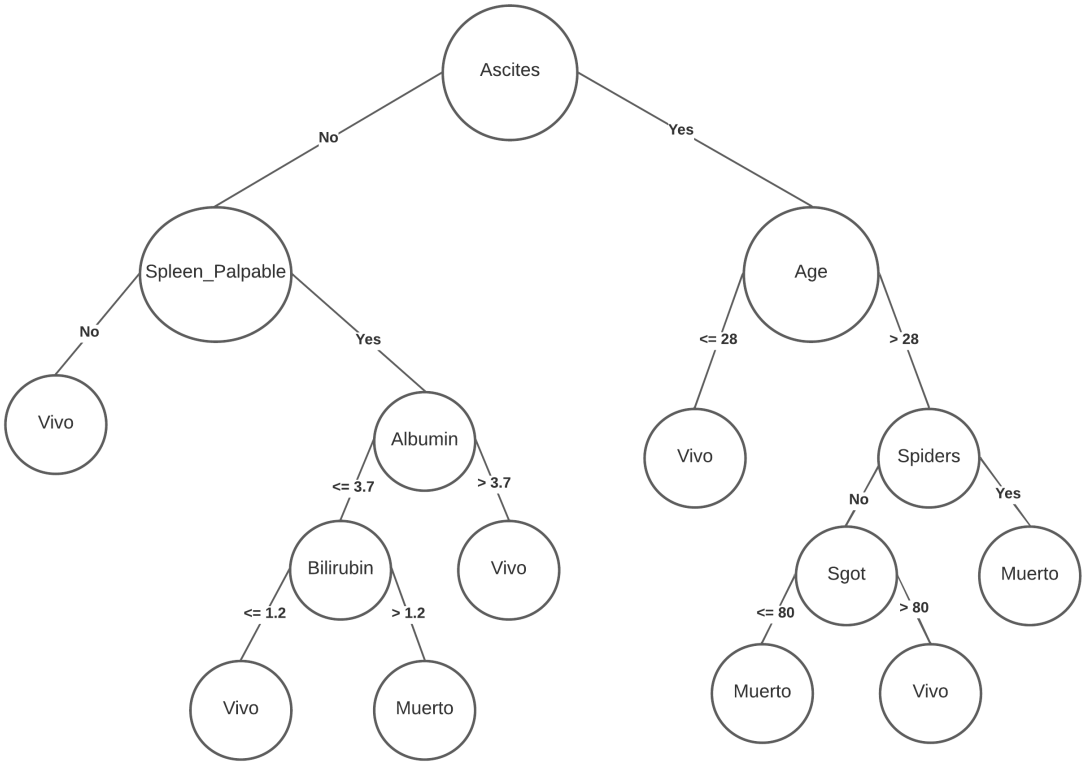

Figura 4: Árbol de decisión.

La Tabla 2 muestra la matriz de confusión obtenida, donde se logra apreciar que el modelo presentó un *accuracy* igual a 88.37 %, *sensitivity* igual a 80 % y una *specificity* igual a 89.47 %. La matriz de confusión como fue mencionado en la sección 2, permite observar aquellos casos clasificados correctamente a la clase a la que pertenecen, así como aquellos clasificados erróneamente.

	Muerto	Vivo
Muerto	4	4
Vivo	1	34

Cuadro 2: Matriz de confusión.

4. Análisis de resultados y comparación

Para este análisis, se discutirán las reglas basándonos en su clase, separando entre *vivo* y *muerto*, y luego se realiza una comparación con la literatura, con otro árbol de decisión. Posteriormente, se compararan las reglas obtenidas a través del algoritmo de árbol de decisión con aquellas generadas mediante el proceso de minería de reglas de asociación del laboratorio 3.

Por un lado, las reglas del árbol de decisión para la clase muerto son:

1. {Spleen Palpable= Yes, Biliburin $>$ 1.2, Albumin$\leq$3.7}
2. {Age $>$ 28, Ascites = Yes, Sgot $\leq$ 80}
3. {Ascites = Yes}

Para la regla número uno, de las que considera dentro de su consecuente a la clase *muerto*, se logra apreciar que el porcentaje de sujetos clasificados de esta manera, presentaron en su mayoría bazo palpable (*spleen palpable*), niveles de bilirrubina superiores a 1.2 (considerados como niveles elevados de acuerdo a los resultados presentados en la entrega anterior) y niveles de albúmina menores o igual a 3.7 (clasificación que cae dentro del rango considerado como niveles bajos de esta proteína).

La regla dos relaciona la ascitis y una edad de adulto joven hacia adelante como atributos que deciden la clasificación como no supervivencia, además de la presencia de niveles de la enzima *Sgot* inferiores a 80.

Por último, vemos la regla número tres, que es la más simple, al clasificar directamente con un atributo, en este caso, cuando el paciente presenta *Ascites*, se tiene influencia directa sobre la incidencia de la clase *muerto*, mostrando una medida del *lift* de 3.3.

En la misma línea de la clase *muerto*, contrastando con la experiencia de reglas de asociación, las reglas 1 y 3 de la experiencia anterior incluían en su antecedente a la falta de examen histológico o *histology*, lo que no se relaciona con los resultados obtenidos con el árbol de decisión.

Aquellas reglas que poseen en su estructura a la clase *vivo* como consecuente se enumeran a continuación:

1. {Ascites = No, Bilirubin $\leq$ 1.2}
2. {Age $\leq$ 28}
3. {Spleen Palpable = No, Ascites = No}
4. {Albumin $>$ 3.7}
5. {Spiders = No}

Cada una de las reglas que contemplan a la clase *vivo* en su consecuente son autoexplicativas, dado que, respaldado en las investigaciones llevadas a cabo en la experiencia anterior, la no presencia de ascitis, bazo palpable y spiders, son un signo claro de capacidad de supervivencia elevada por parte de los sujetos. Además, considerando que las reglas 4 y 7 de la Tabla 1 asocian a los pacientes con niveles de bilirrubina y albúmina normales, respectivamente, con la clase sobreviviente, entonces se espera que se encuentren en un buen estado de salud, por lo demás, sería extraño que estos pacientes fallecieran, a no ser que presenten otras cualidades que la base de datos no registra.

En una comparación entre las reglas de asociación y las reglas obtenidas a través del árbol de decisión, se observa una enorme diferencia entre la cantidad de reglas a obtener, de las que en la experiencia anterior solo se analizó un *top* con aquellas reglas con las mejores mediciones, mientras que ahora solo tenemos ocho reglas a utilizar. Cabe destacar que esta diferencia se debe a que anteriormente intentamos buscar *todas* las reglas interesantes del *dataset* (solo que se filtró por el consecuente con las clases que se deseaba analizar, es decir, *vivo* y *muerto*), de ahí el nombre de la minería de reglas de asociación, al que se tiene que entregar la clase objetivo y sus predictores.

En relación a lo encontrado en la literatura, se tiene una aplicación del algoritmo *random forest*, entregando por resultado el árbol de decisión mostrado en la figura 5. En el nodo inicial del árbol, se encuentra una coincidencia sobre el mismo atributo utilizado como punto de partida, es decir, la presencia de ascitis, mostrando, entonces, que realmente es fundamental evaluar este aspecto en forma médica, y no solo coincidencia de los árboles de decisión. También es posible identificar los atributos *spiders*, *albumin* y *spleen palpable* que coinciden con el árbol obtenido, sin embargo, ciertos nodos cambian y se consideran otros

atributos, tales como el género de la persona (*sex*), *SGOT* y aquellas columnas relacionadas con la textura del hígado (*liver firm* y *liver big*).

La presencia de ascitis como nodo raíz del árbol de decisión, no es casualidad, debido a que, tal como afirma Rudralingam et al. (2016), la acumulación excesiva de líquido intraperitoneal, conocida como ascitis, es una pista importante que apunta a una enfermedad subyacente significativa. Es decir, es un factor decisivo a la hora de clasificar pacientes de acuerdo a su capacidad de supervivencia.

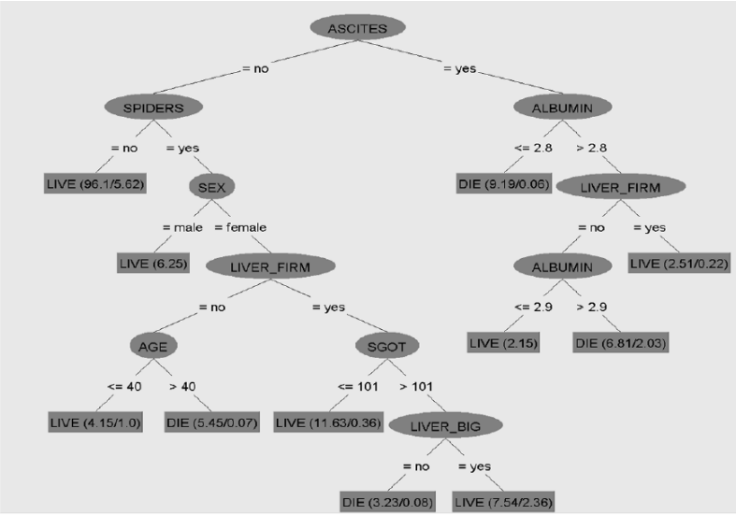

Figura 5: Árbol de decisión generado por diferente algoritmo para el mismo dataset (Karthikeyan and Thangaraju, 2013).

Aquellas reglas que, en el estudio llevado a cabo en el laboratorio sobre reglas de asociación, llevaban en su consecuente a la clase *muerto*, presentaron un *lift* igual a 4.1428. En esta ocasión, se obtuvo un *lift* superior a 3.3 para las 3 reglas que contemplan a la clase muerto en su consecuente. Es decir, antecedente y consecuente están correlacionados positivamente, es decir, existe una asociación de no independencia entre los factores.

En cuanto a los valores de *lift* para la clase *vivo* (presente como consecuente en las reglas 4, 5, 6, 7 y 8 de la Tabla 1), se logra apreciar que ligeramente superan el umbral que las define como reglas exentas del resultado de un suceso aleatorio (1.1 - 1.2). Este fenómeno se repitió de igual forma para las reglas obtenidas a través del proceso de minería de reglas de asociación llevado a cabo con anterioridad (entre 1.1419 - 1.1726).

Es importante señalar que la base de datos utilizada para la obtención del árbol

de decisión, presenta un desbalance desde su origen, donde la clase *vivo* abarca el 80.69 % del total de los sujetos, y la clase *muerto* el restante 19.31 %.

La Figura 6 permite apreciar de mejor manera el desbalance de clases presente en la base de datos. A pesar de esto, los resultados tanto de *sensitivity*, *specificity* y *accuracy* superan el 80 %, es decir, son valores confiables que permiten continuar con el estudio. Sin embargo, es probable que el modelo sea capaz de predecir mayoritariamente para la clase predominante en la base de datos (clase *vivo*).

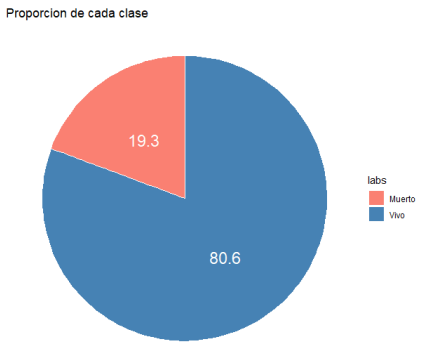

Figura 6: Representación gráfica del desbalance de clases.

5. Conclusiones

En el presente estudio, se llevó a cabo la construcción de un árbol de decisión, que fue de gran utilidad para la creación u obtención de reglas que permitieron asociar las diferentes cualidades que se ven implicadas en la determinación de la clase de los sujetos. La estructura del árbol de decisión, mostró, en primer lugar, como nodo raíz, a la Ascites o Ascitis, característica también presente en el estudio realizado con *random forest* por Karthikeyan and Thangaraju (2013). Sin embargo, los nodos de los niveles superiores denotan otras características, principalmente se cree que esto se debe a los parámetros de *train/test* y la semilla utilizada. Además, es importante considerar que *random forest* es un algoritmo que basa su funcionamiento en el uso de un gran número de árboles de decisión individuales que operan como un conjunto Yiu (2019). Es considerado una forma de perfeccionar los resultados obtenidos de un solo árbol de decisión, por lo que, es esperable que a través de él se obtengan mejores resultados.

Una de las principales ventajas que notamos en la generación del árbol con la librería "C50" es que a diferencia de las reglas de asociación, la función `C5.0` admite un *set* de datos con valores numéricos discretos y continuos, por lo que el uso de la discretización de la experiencia anterior no fue necesaria en esta ocasión. Por otro lado, aunque se tiene esta ventana de fácil uso, es importante tener en cuenta los resultados de Zhou (2014), donde se menciona que las reglas de asociación permitieron una mayor precisión que algunos enfoques tradicionales de clasificación como los árboles de decisión. Menciona que a diferencia de los demás métodos tradicionales, las reglas de asociación cuentan con dos características: se genera una gran variedad de reglas de asociación y la medida de confianza y soporte se utiliza para evaluar la importancia de las reglas de asociación.

Un factor a tener en cuenta es que las reglas obtenidas a través del algoritmo de árbol de decisión que contemplan en su consecuente a la clase vivo presentan un valor de *lift* ligeramente superior a 1, dado esto, es necesario ser precavidos y evitar el uso de generalización al interpretar los resultados, ya que, muy posiblemente las asociaciones encontradas para dicha clase son producto de la casualidad y no representan una realidad médica.

Bibliografía

Garg, A. (2018). Complete guide to association rules (1/2) - towards data science. https://towardsdatascience.com/association-rules-2-aa9a77241654.

IBM (2022). ¿Qué es un árbol de decisión? https://www.ibm.com/es-es/topics/decision-trees.

Johnson, D. (2022). Decision tree in R: Classification tree with example. https://www.guru99.com/r-decision-trees.html#7.

Karthikeyan, T. and Thangaraju, P. (2013). Analysis of Classification Algorithms Applied to Hepatitis Patients. *International Journal of Computer Applications*, 62(15).

Lastra, E. F. (2020). Qué es un árbol de decisión y su importancia en el Data Driven. https://artyco.com/que-es-un-arbol-de-decision-y-su-importancia-en-el-data-driven/.

Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., and Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021:1–15.

Rodrigo, J. A. (2018). Reglas de asociación y algoritmo apriori con r. https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion#:~:text=Una%20regla%20de%20asociacin%20se,hand-side%20(RHS).

Rudralingam, V., Footitt, C., and Layton, B. (2016). Ascites matters. *Ultrasound*, 25(2):69–79.

Yiu, T. (2019). Understanding random forest - towardsdatascience.com. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

Zhou, Z. (2014). A new classification approach based on multiple classification rules. *Mathematical Problems in Engineering*, 2014:1–7.