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

Laboratorio 4 - Árboles de decisión

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

18 de Julio de 2021

Tabla de contenidos

1. Introducción

A medida que pasa el tiempo se generan cada vez más datos, los cuales al ser procesados utilizando diversos métodos permiten obtener información, lo que resulta útil para la toma de decisiones. Una de las técnicas usadas para poder tomar decisiones corresponde a los árboles de decisión, el cual es un tipo de aprendizaje supervisado que busca lograr predecir el valor de una variable dependiente a partir de las características que posea el conjunto de datos, con el fin de asistir en la toma de decisiones.

En el siguiente informe se presenta el desarrollo del cuarto laboratorio del curso Inteligencia Computacional, el cual busca continuar con el análisis de la base de datos hepatitis y mostrar el desarrollo y análisis de los objetivos que propuestos en el enunciado del laboratorio. Entre los objetivos a cumplir se encuentran: implementar de manera correcta árboles de decisión sobre la base de datos mencionada mediante el lenguaje de programación R, comparar los resultados obtenidos con lo expuesto en la teoría y el laboratorio3 de reglas de asociación e identificar las principales diferencias entre reglas de asociación y árboles de decisión.

La motivación del trabajo es lograr obtener un mayor conocimiento a partir de los datos y ver si es que existen características que permitan dilucidar respecto a que variables son más relevantes para predecir si un pacientes sobrevivirá a la enfermedad(hepatitis). También se busca conocer si al variar las características utilizadas para la generación de los árboles, afecta el resultado final.

2. Marco Teórico

2.1. Arboles de decisión

Los árboles de decisión son uno de los algoritmos de aprendizaje supervisado y se pueden comportar como regresor o clasificador, dependiendo si las variables son continuas o discretas. Su principal objetivo es lograr generar una función que a partir de un grupo de variables independientes permita predecir el valor de la variable objetivo (sitio big data, 2019). La referencia anterior también indica que el árbol se construye mediante un enfoque top-down, siguiendo los siguientes pasos:

1. Se selecciona el mejor atributo utilizando alguna medida y se dividen el espacio de las variables independientes.

2. La variable seleccionada se convierte en nodo.

3. Si no quedan más atributos o instancias se finaliza el algoritmo, de lo contrario se vuelve al paso 1.

A continuación se definen conceptos importantes correspondientes a como está compuesto un árbol de decisión:

- **Nodo raız:** Es el nodo que divide a toda la población o muestra en dos conjuntos homogéneos. Representa a la característica con la mejor medida utilizada.

- **Nodo de decisión:** Corresponden a los nodos intermedios y al igual que el nodo raíz representan características.

- **Nodo hoja o final:** Son nodos que no tienen hijos y representan el resultado obtenido.

- **Ramas:** Es una subsección del árbol y representa una regla de decisión.

En base a lo anterior es que a partir de los árboles de decisión (clasificador) es posible obtener reglas de asociación, siguiendo los caminos generados desde el nodo raız hasta los nodos finales.

2.2. Métricas de calidad

Debido a la base de datos utilizada en el laboratorio es que el árbol de decisión a implementar tendrá un comportamiento de clasificador, por lo tanto, las medidas de calidad seleccionadas para establecer comparaciones corresponden al AUC y F1.

2.3. AUC

La curva de ROC permite representar de manera gráfica la relación de especificidad y sensitividad para un sistema de clasificación binario dado un punto de corte. Uno de los índices que permite medir esta curva corresponde al AUC, el que representa el área bajo la curva, es decir, la probabilidad de que dado dos casos la prueba los clasifique de manera correcta, por lo cual, mientras más cercano a uno sea su valor, mejor es el clasificador(Hanley and McNeil, 1982).

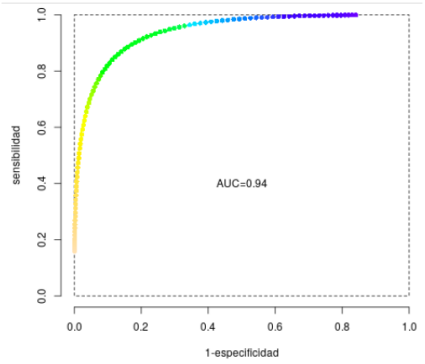

Figura 1: Figura 1: Curva de ROC

2.4. F1

F1 es una medida que combina la presición con el recall (sensitividad) y le asigna la misma importancias a ambas, sin embargo, la utilidad de este valor dependerá del caso de estudio (Heras, 2020). Mientras más cercano a 1 se encuentre F1, ambos parámetros tienen la misma importancia dentro del clasificador.

3. Obtención del Árbol

Los diferentes arboles de decisión generados a partir de la base datos Hepatitis se obtienen haciendo uso del algoritmo C5.0, el cual crea árboles de clasificación ya sea utilizando árboles de decisión simples, modelos basados en reglas o *ensembles* basados en boosting. Cabe señalar que para las divisiones del árbol, este utiliza la medida de pureza mediante la entropía y además los árboles generados pueden ser convertidos en modelos basados en reglas (**?**).

El algoritmo C5.0 de la librería C5 de R tiene entre sus parámetros el parámetro trials, el cual nos permite aplicar el concepto de boosting a la generación de nuestros arboles de decisión como se puede observar en la figura X. Además también se hace uso de Ramdom Forest como otra forma de obtener un árbol de decisión, utilizando el concepto de bagging. Luego podemos comparar los resultados de utilizar ambos conceptos para la generación de los árboles.

Para esta experiencia se han considerado conjuntos de entrenamiento con un 70 % hasta un 90 % de los datos. Los diferentes resultados obtenidos se pueden visualizar al ejecutar el código de la presente experiencia pues en el informe solo se hace énfasis en aquellos arboles que resultaron ser más concordantes con lo obtenido en la experiencia pasada (Laboratorio 3). En los capítulos siguientes se analizan y comparan los resultados obtenidos en ambas experiencias.

3.1. Arboles generados utilizando todos los atributos del dataset

Entre las pruebas realizadas con los diferentes grupos de entrenamiento(70 %, 75 %, 80 %, 85 %, 90 %) y utilizando todas las variables del dataset, se obtuvieron arboles con valores menores o iguales a 50 % de *Sensitivity*, es decir, solo son capaces de predecir el 50 % o menos de los casos. Sin embargo la prueba realizada con un 80 % de los datos para el entrenamiento nos entrego una *Sensitivity* de 54 % con un 83 % de casos acertados. El árbol generado se puede apreciar en la Figura 2, en el cual se observa el atributo *ascites* como nodo decisión. Recordando esta variable representa la presencia de acumulación de líquido seroso en la zona abdominal y como se menciona en (quimica.es, 2019), presentar líquido seroso en esa zona suele estar relacionado a cáncer hepático como en los pacientes con hepatitis.

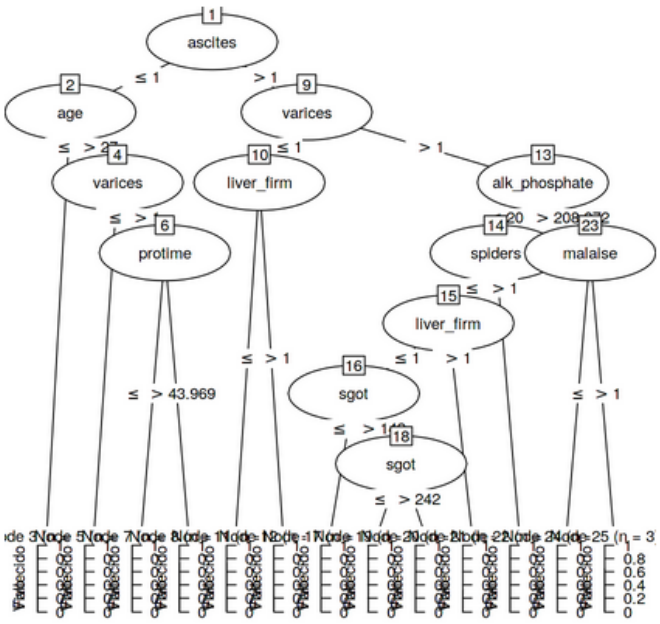

Figura 2: Árbol de decisión con 80 % de entrenamiento y usando toda la data

El árbol obtenido usando Random Forest y los datos anteriores nos entrego la Bilirrubina como variable de decisión asi como varias variables importantes que son consistentes con las variables de interés otenidas en la experiencia anterior. En la Figura 2 se puede observar la importancia de la bilirrubina y protime que son los atributos que se corresponden con la experiencia anterior. Cabe destacar que la 2 también tiene variables coincidentes con las obtenidas en la experiencia anterior, las cuales se encuentran en su mayoría al lado izquierdo del árbol.

Cabe señalar que para la prueba con un 50 % de los datos como conjunto de entrenamiento y con 5 trials, es decir, aplicando el concepto de boosting para la generación del árbol se obtuvo un 81 % de *Accuracy* y un 57 % de *Sensitivity*, luego se puede decir que al aplicar boosting para este tipo de estudio podría resultar favorecedor. El atributo decisión es*ascites* y siendo el lado derecho el cual destaca la bilirrubina y spleen-palpable ambos atributos también se han presentados como atributos de interés en las experiencias pasadas.

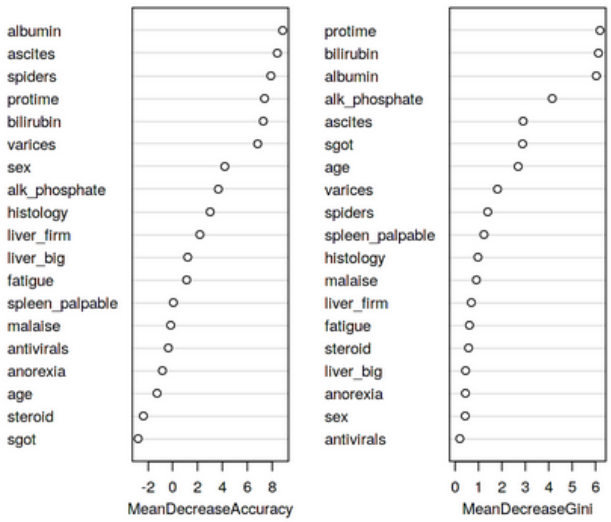

Figura 3: Árbol de decisión usando Random Forest con 80 % de entrenamiento y usando toda la data

3.2. Arboles generados utilizando atributo considerados relevantes

Para generar los siguientes arboles que se muestran a continuación, se trabajo la data, filtrando los atributos y dejando solo aquellos atributos considerados relevantes en las experiencias anteriores, luego los nueve atributos escogidos fueron, age, liver-big, siders, billirubin, alk-phosphate, sgot, albumin, protime y histology. Además de realizar un balanceo de los datos filtrados. Todos los arboles resultan ser bastantes interesantes y congruentes con las experiencias anteriores, además de tener porcentajes superiores al 80 % de *Accuracy* como de *Sensitivity*.

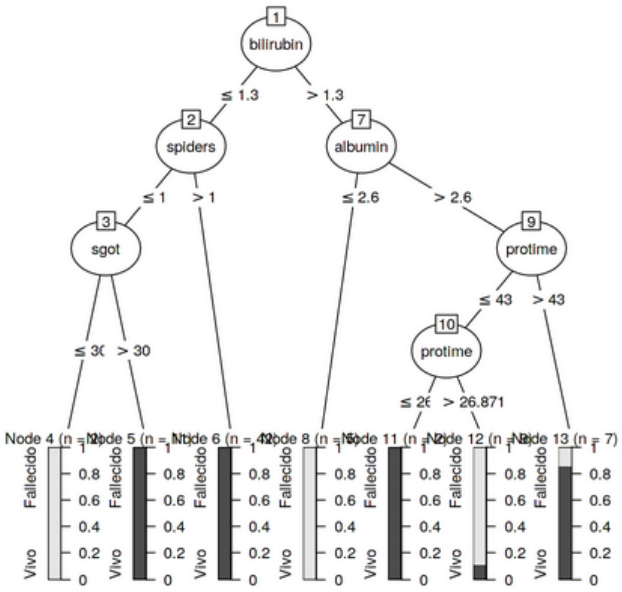

Figura 4: Árbol de decisión con 50 % de entrenamiento, utilizando el concepto de boosting y usando solo atributos relevantes

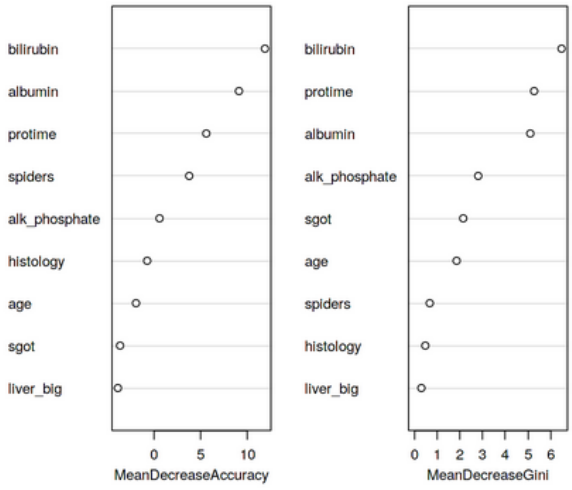

Figura 5: Árbol de decisión usando Random Forest con 50 % de entrenamiento, utilizando el concepto de boosting y usando solo atributos relevantes

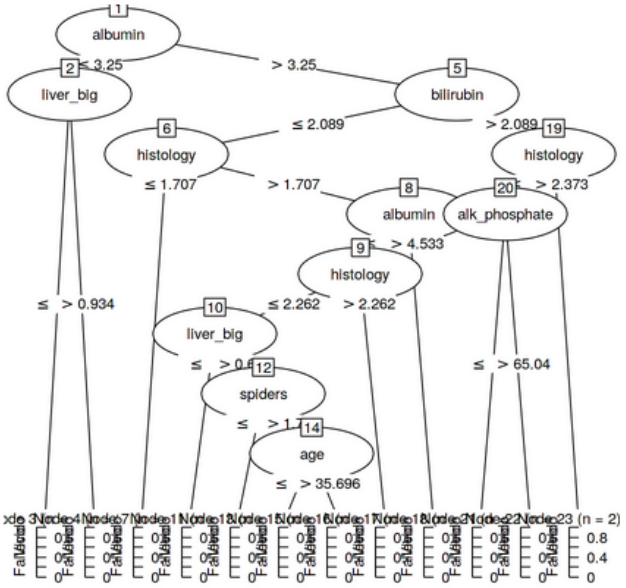

Figura 6: Árbol de decisión con 80 % de entrenamiento, utilizando el concepto de boosting, usando solo atributos relevantes y con datos balanceados

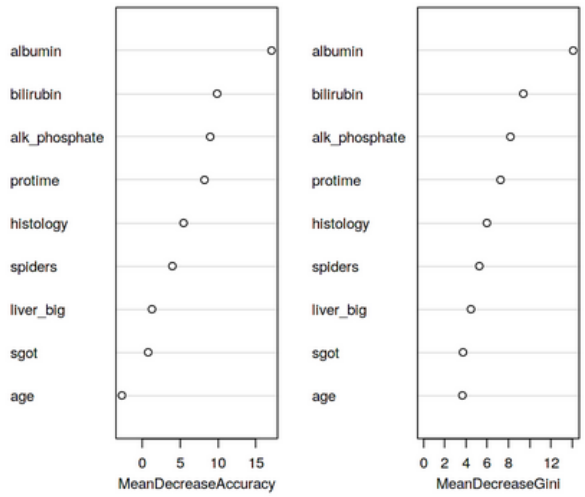

Figura 7: Árbol de decisión usando Random Forest con 80 % de entrenamiento, utilizando el concepto de boosting, usando solo atributos relevantes y con datos balanceados

4. Análisis de los resultados

De acuerdo a lo presentado en el apartado anterior, podemos notar que hay dos grupos principales de resultados, donde el primero corresponde a la generación de árboles con datos no filtrados, ni balanceados y otros que fueron aplicadas diferentes técnicas, por lo que desde estas situaciones, se generan diferentes formas de comprender las clasificaciones. Sin embargo, de la primera, a priori, se tiene que está algo alejada a las experiencias previas y a la literatura trabajada en los laboratorios anteriores.

4.1. Análisis de los árboles generados

De los primeros árboles generados, podemos notar que *ascites* es uno de los atributos más relevantes que se tiene como parte del origen o nodo principal, seguido por *varices*, *edad* y en algunos casos la *bilirubin*. Estas variaciones, se originan a partir de los diferentes conjuntos de entrenamiento, ya sea desde el 50 % hasta incluso el 90 %.

Dado lo anterior, se logra observar esta dependencia de *ascites* (si hay presencia de acumulación de líquido seroso o no) como el principal nodo, que de acuerdo a la literatura (Burakoff (2017)), es bastante recurrente utilizar este parámetro para identificar rápidamente algún tipo de afección entorno al hígado, pero no necesariamente algo ligado a la hepatitis.

Ahora bien, de acuerdo a lo evidenciado en los laboratorios anteriores, existe una cierta divergencia entorno a esta variable, dado que no se había presentado hasta ahora, como un componente de tal importancia. Es por ello que la exploración se expande en ajustar los datos a utilizar, así como también en la filtración de los principales atributos a considerar dentro del modelo de los árboles de decisión. Por otra parte, podemos recalcar que al aplicar Random Forest a este mismo grupo de entrenamiento, se puede evidenciar que existe fuerte presencia de algunos atributos que ya habían sido identificados en laboratorios anteriores, como por ejemplo la *bilirubin*, *albumin*, *protime*, $alk - phosphate$, *sgot*, *histology*, entre otros. Notar que si bien es cierto, no se muestra en los árboles decisión, estos si contribuyen en la precisión del modelo, acorde a la definición MD Acurracy y MD Gini, donde a mayor valor o índice, mayor es su importancia o contribución al modelo.

De este grupo de árboles generados con todo el dataset, en general se tienen precisiones

(accuracy dentro del los gráficos) menores al 80 % y sensibilidad entorno al 50 %, lo que se consideran resultados poco satisfactorios, considerando además lo anteriormente señalado. Si bien se tiene un porcentaje bastante menor respecto al error de la clasificación, sin embargo, la probabilidad de que se obtenga si un paciente puede sobrevivir o fallecer, entorno al *ascites*, deja bastante que desear.

Del apartado de los resultados con los atributos relevantes, acordes a lo evidenciado en los laboratorios anteriores, en general se obtienen resultados que mejoran con respecto a los datos 'sin procesar' o filtrar.

Al aplicar Synthetic Data Generation y considerar los atributos más relevantes, se tiene que la precisión y la sensibilidad están entorno al 80 % y al 90 % respectivamente, lo que se evidencia claramente que al tener este trabajo previo en el conjunto de datos, a priori, se obtiene una mejor clasificación en los árboles de decisión.

Notar además que al momento de disminuir el conjunto de entrenamiento, la precisión fue disminuyendo drásticamente, a diferencia de los resultados del primer grupo. De este conjunto de resultados, se logra observar que los atributos de *albumin*, *bilirubin*, *protime* y la *histology*, son los principales componentes que aportan al modelo predictivo, en donde además forman parte de los primeros nodos de los árboles decisión.

Notamos que al variar el conjunto de entrenamiento entre un 70 % y un 90 %, los nodos principales, se mueven entre la *albumin* y la *bilirubin*, componentes que ya han sido previamente identificados, como relevantes para la predicción del problema. Incluso para un conjunto de entrenamiento al 70 %, se logra observar que para la $bilirubin > 2,22$ tiene como consecuencia la muerte del paciente, con una precisión del 80 %, sensibilidad del 85 %. Además, dentro de esta misma clasificación, para valores de *albumin* inferiores a $\sim 3,25$ a pesar de tener niveles bajos de *bilirubin*, existen condiciones específicas que también conducen a la probabilidad de que el paciente pueda fallecer, como por ejemplo que los niveles de *protime* sea alta, es decir que la sangre toma mayor tiempo de lo esperado para coagularse y por tanto el hígado no contiene la suficiente cantidad de proteínas para realizar tal tarea, donde usualmente se explica por un daño importante en el órgano o por cirrosis (of Veterans Affairs (2006)).

En todos los grupos de resultados el parámetro trial, fue bastante significativo para mejorar lo presentado en este laboratorio, como método de boosting para los clasificadores.

4.2. Comparación con reglas

Las reglas obtenidas en este laboratorio, fue por medio de C5.0, una librería que nos permite también obtener las asociaciones que contribuyan a la explicación de la clase predictora del conjunto de datos.

Notamos que una configuración bastante interesante de observar, es la obtenida al aplicar SDG, filtrado de atributos relevantes, conjunto de entrenamiento del 80 % y un boosting de 5 iteraciones (trials = 5).

De acuerdo al clasificador de árboles de decisión, podemos ver que los atributos considerados, son en general significativos entorno a un uso del 84 % en promedio, por lo que esta configuración nos da indicios de estar en un entorno correcto, de hecho la precisión alcanza un 83 % y una sensibilidad del 91 %.

Por otra parte, las reglas que se obtuvieron de acuerdo la figura, se logra evidenciar que al tener una $bilirubin > 2,08$, $alk-phosphate > 65,04$ y presencia de la *histology*, tiene consecuencia el fallecimiento del paciente, obtiendose además un $lift = 1,9$. Esta regla encontrada, es bastante relavante desde el punto de vista de la literatura, por ejemplo la *bilirubin* en un adulto tiene niveles normales entorno a 1,2($\frac{mg}{dl}$) de sangre, para personas más jóvenes, estos níveles decrecen, por lo que, la regla que se obtiene es bastante consecuente con la realidad (Sabrina Felson (2021)), para casos en donde se presenta algún tipo de falla hepática, cirrosis o de hepatitis. Por otra parte, se tiene los niveles de $alk - phosphate$ para un adulto, se encuentra del orden de $20 - 140(U/L)$ (Ujjawal Sharma and Rajendra Prasad (2014)), por lo que la regla encontrada, está asociando a un número inferior, si se considera que para valores más elevados, el paciente podría eventualmente tener cirrosis, cáncer de hígado, hepatitis, sobre medicación de un hígado dañado, mal nutrición, entre otras condiciones.

Ahora bien, respecto a la *histology*, este atributo se hace presente nuevamente al igual que en los laboratorio previos, así como en particular en las reglas de asociación obtenidas con el algoritmo de Apriori.

Otra regla importante dentro de esta configuración, es cuando el paciente no presenta un híga-

do de tamaño grande ($liver - big = 1$) y que además los niveles de $albumin < 3,25(g/dL)$, que de acuerdo a (Medical Center of University of Rochester (2021)) los valores normales de un adulto oscilan entre $3,4$ a $5,4(g/dL)$, por lo tanto, esta asociación, tiene relevancia, dado que este tipo de pacientes, podrían presentar no solamente un fallo hepático, si no que también podrían involucrar patologías de los riñones (U.S Department of Veteran Affairs (2006)). Pero del punto de vista de la clasificación, no es concluyente, pero si importante de considerar.

Las otras reglas obtenidas en esta configuración, no son precisas respecto a los valores de corte, donde se acotan por valores muy generales y que no contribuyen a la consecuencia del problema.

Otra configuración que es importante recalcar, es similar a lo anterior, pero con un conjunto de entrenamiento del 70 %, donde algunas de las reglas que se logran evidenciar, donde la $bilirubin > 1,33$, $alk_phosphate > 49,36$, $sgot > 38,64$, $protime < 58,82$ y el hígado de tamaño normal, se tiene como consecuencia la muerte del paciente. Si consideramos que los niveles normales de *sgot* están entorno a 8 y $45(u/LS)$ (Ana Gotter (2018)) y para el caso de protime, no se logra encontrar evidencia respecto a la unidad de escala del dataset.

Respecto a las reglas obtenidas en el laboratorio 3, son similares entorno a los atributos relevantes que fueron evidenciados, pero los grupos difieren en información, y debido a una restricción del modelo Apriori donde los componentes debían ser discretizados (i.e bajo, medio y alto), los consecuentes son algo difusos.

5. Conclusión

Con respecto a los aspectos más significativos del laboratorio se encuentras en primera instancia el hecho de obtener representaciones similares a partir de algoritmos diferentes, es decir, que tanto las reglas de asociación como los árboles de decisión generen reglas, las cuales pueden ser comparadas en incluso llegar ser parecidas.

Entre las ventajas que se identificaron al utilizar arboles de decisión en comparación a los demás algoritmos utilizados en los laboratorios anteriores tales como clustering y reglas de asociación se encuentran la visualización que posee el árbol, pues esta resulta ser bastante intuitiva, luego no es necesario explicar exhaustivamente los resultados ni demorar en interpretarlos o de buscar alguna forma de representarlos.

Otra de las ventajas que presentan los arboles en comparación a las reglas de asociación es su tiempo de cómputo, el cual es menor al de las reglas, debido a que a medida que el algoritmo va iterando, acota el espacio de búsqueda de soluciones, en cambio las reglas generan todas aquellas que cumplan con una métrica establecida, por lo que puede incluso llegar a producir todas las combinaciones posibles. Por lo tanto, si se desean obtener reglas y se cuenta con una cantidad excesivamente grande de datos y poca capacidad de procesamiento, la elección deberıa ser el árbol. Sin embargo, se debe destacar que el árbol no genera exactamente las mismas reglas, si no que, tal como se vio en el desarrollo del presente laboratorio, un subconjunto de estas, las que además pueden ser redundantes.

En relación a los resultados obtenidos se puede destacar que las reglas de clasificación fueron mejores al utilizar C5.0(Arboles de decisión) versus las reglas generadas por el método de reglas de asociación, las cuales fueron obtenidas en la experiencia anterior. Con respecto a la comparación de los resultados los árboles de decisión generaron reglas bastante acordes a las obtenidas con las reglas de asociación, cabe destacar que con los arboles se obtuvieron parámetros que son acordes a lo revisado en la literatura para determinar la gravedad de la hepatitis. Destaca una de las reglas obtenidas del conjunto de datos con un 70 % de entrenamiento, la cual menciona 5 atributos relevantes como la bilirubina, alcalina fosfato, sgot, protime y big-liver dando valores que se consideran peligroso y consistentes con la literatura, luego un paciente pueda fallecer con niveles que superan casi los normales y en

algunos casos superándolos al igual como lo representa la regla obtenida.

Ante todo lo anterior, es posible determinar que los objetivos planteados en un inicio se cumplieron en su totalidad, los cuales fueron: Implementar de manera correcta árboles de decisión sobre la base de datos mencionada mediante el lenguaje de programación R y comparar los resultados obtenidos con lo expuesto en la teoría, la literatura y el laboratorio 3 de reglas de asociación e identificar las principales diferencias entre reglas y árboles.

Es importante destacar la ayuda de R, ya que facilita el manejo de los datos y cuentan con funciones y estructuras que permiten de manera sencilla obtener los resultados, ya sea de manera numérica o gráfica. Se espera que tanto el trabajo realizado en esta actividad como los laboratorios anteriores sean de ayuda para problemáticas que se pueden presentar a futuro. Además, se puede establecer que se logró obtener un conocimiento en mayor profundidad de los datos y del problema en general.

Bibliografía

Ana Gotter, h.-t. (2018). Sgot test.

Burakoff, C. D. (2017). How cirrhosis from chronic hepatitis can cause ascites.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Heras, J. M. (2020). Precision, recall, f1, accuracy en clasificación.

Medical Center of University of Rochester, h. (2021). Albumin (blood) article, health encyclopedia.

of Veterans Affairs, U. D. (2006). Prothrombin time, hepatatis c for patience. https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-prothrombin-time.asp.

quimica.es (2019). Ascitis.

Sabrina Felson, h.-t.-z.-g.-t. (2021). What is a bilirubin test?

sitio big data (2019). Árbol de decisión en machine learning (parte 1).

Ujjawal Sharma, D. P. and Rajendra Prasad, h. (2014). Alkaline phosphatase: An overview.

U.S Department of Veteran Affairs, h.-a. (2006). Viral hepatitis and liver disease.