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

Laboratorio 4 - Árboles de Decisión

Integrantes:	Christian Méndez Acosta
	Israel Arias Panez
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco
Ayudante:	Gustavo Hurtado A.

31 de Octubre de 2022

Tabla de contenidos

1. Introducción

En la presente experiencia de laboratorio se busca continuar con el estudio del hipotiroidismo y la base de datos *Allhypo*. Para esta experiencia se estudiará el problema mediante Árboles de decisión. Los árboles de decisión son modelos predictivos formados por reglas binarias con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta (Amat, 2017). Se hace interesante el estudiar el problema con árboles de decisión debido a que como sostiene (Amat, 2017), los árboles son fáciles de interpretar aun cuando las relaciones entre predictores son complejas, son muy útiles en la exploración de datos, ya que permiten identificar de forma rápida y eficiente las variables (predictores) más importantes y además pueden aplicarse a problemas de regresión y clasificación.

Estudiar este problema es relevante debido a que el hipotiroidismo es la situación clínica producida por un déficit de hormonas tiroideas (Barea et al., 2012). Las hormonas tiroideas afectan a la función de todos los órganos del cuerpo, regulan la actividad metabólica y son críticas en el desarrollo somático y cerebral en infantes (Brent, 2022). Además, a nivel mundial afecta hasta al 5 % de la población, y se calcula que otro 5 % no está diagnosticado (Chiovato et al., 2019).

Los objetivos para esta experiencia de laboratorio son:

- Extraer conocimiento respecto al hipotiroidismo, mediante el uso del software R, la librería C50 y la base de datos *Allhypo*, utilizando arboles de decisión.
- Analizar y comparar los resultados obtenidos con la literatura encontrada y lo expuesto en la teoría.
- Comparar los resultados obtenidos con las reglas de laboratorio número 3 de reglas de asociación, identificando las principales diferencias entre el análisis de ambos métodos.

En la siguiente sección se presentará un marco teórico útil para entender conceptos relevantes a la experiencia, para luego presentar la forma en la cual se obtuvo el árbol adecuado. Posteriormente a esto se interpretarán los resultados obtenidos, comparándolos a la literatura y se realizara una comparación del método de árboles de decisión con el método de reglas de asociación de la experiencia de laboratorio anterior en base a los resultados obtenidos.

2. Marco Teórico

2.1. Reglas de asociación

Una regla de asociación se define como una implicación del tipo "si X entonces Y" ($X \Rightarrow Y$), donde X e Y son items individuales. El lado izquierdo de la regla recibe el nombre de antecedente o left-hand-side (LHS) y el lado derecho el nombre de consecuente o right-hand-side (RHS). Por ejemplo, la regla $\{A, B\} \Rightarrow \{C\}$ significa que, cuando ocurren A y B, también ocurre C (**?**). La minería de reglas de asociación es un procedimiento cuyo objetivo es observar y encontrar patrones, correlaciones o asociaciones que se producen con frecuencia a partir de conjuntos de datos (**?**).

2.2. Árbol de decisión

Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Un árbol de decisión comienza con un nodo raíz, que no tiene ramas entrantes. Las ramas salientes del nodo raíz alimentan los nodos internos, también conocidos como nodos de decisión. En función de las características disponibles, ambos tipos de nodos realizan evaluaciones para formar subconjuntos homogéneos, que se indican mediante nodos hoja o nodos terminales. Los nodos hoja representan todos los resultados posibles dentro del conjunto de datos (IBM, sf). En la Figura 1, se puede apreciar un ejemplo de árbol de decisión.

2.3. Algoritmo C5.0

El algoritmo C5 es uno de los algoritmos más utilizados en el ámbito de los árboles de clasificación. Este algoritmo crea modelos de árbol de clasificación, permitiendo sólo variables de salida categórica. Las variables de entrada pueden ser de naturaleza continua o categórica (Parra, 2019).

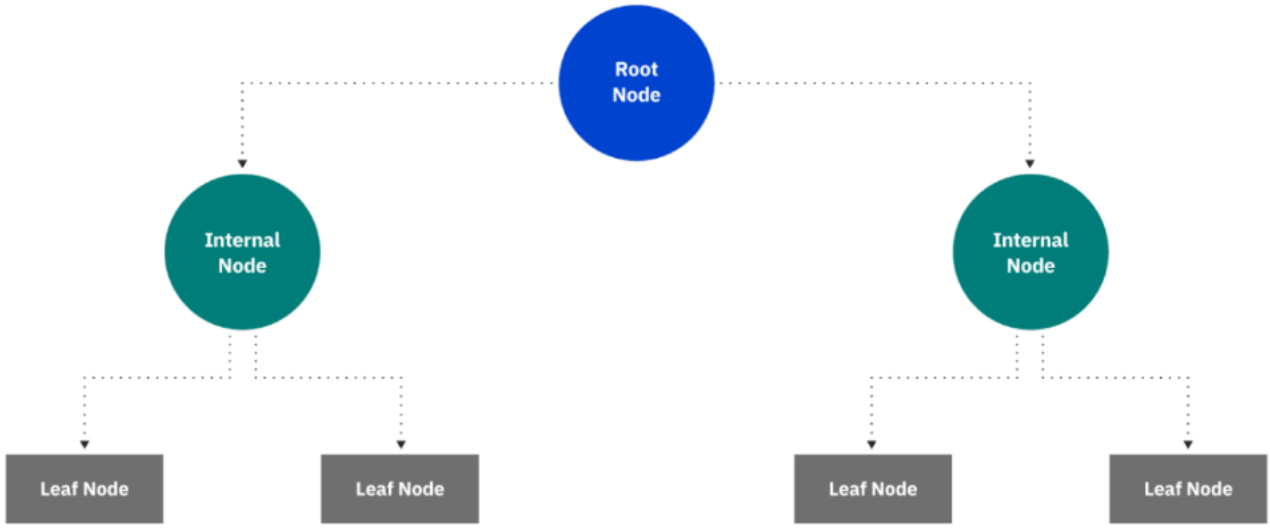

Figura 1: Ejemplo de árbol de decisión

2.4. Matriz de confusión

Es una tabla resumida utilizada para medir el rendimiento de un modelo de clasificación. La tabla indica la cantidad de predicciones correctas e incorrectas obtenidas a partir del modelo a evaluar. A continuación, en la Figura 2 se presenta un ejemplo de una matriz de 2x2, el resultado de una predicción tiene como posibles valores si y no. La cantidad de valores correctamente clasificados, es decir, el valor arrojado por un modelo coincide con el valor real, corresponde a las celdas verdaderos positivos y verdaderos negativos, por otra parte los valores clasificados de forma errónea, es decir, el valor arrojado por un modelo no coincide con el valor real, corresponden a los falsos positivos y los falsos negativos (Shin, 2020).

A partir de las cantidades obtenidas en la tabla de confusión, se establecen las siguientes razones para evaluar el modelo:

- Exactitud: indica la proporción de predicciones que el modelo clasifico correctamente, es el cociente entre la cantidad de predicciones correctas y la cantidad total de predicciones.

- Precisión: indica la proporción en que las predicciones positivas fueran correctas, es el cociente entre los verdaderos positivos y la suma de los verdaderos positivos con los falsos positivos.

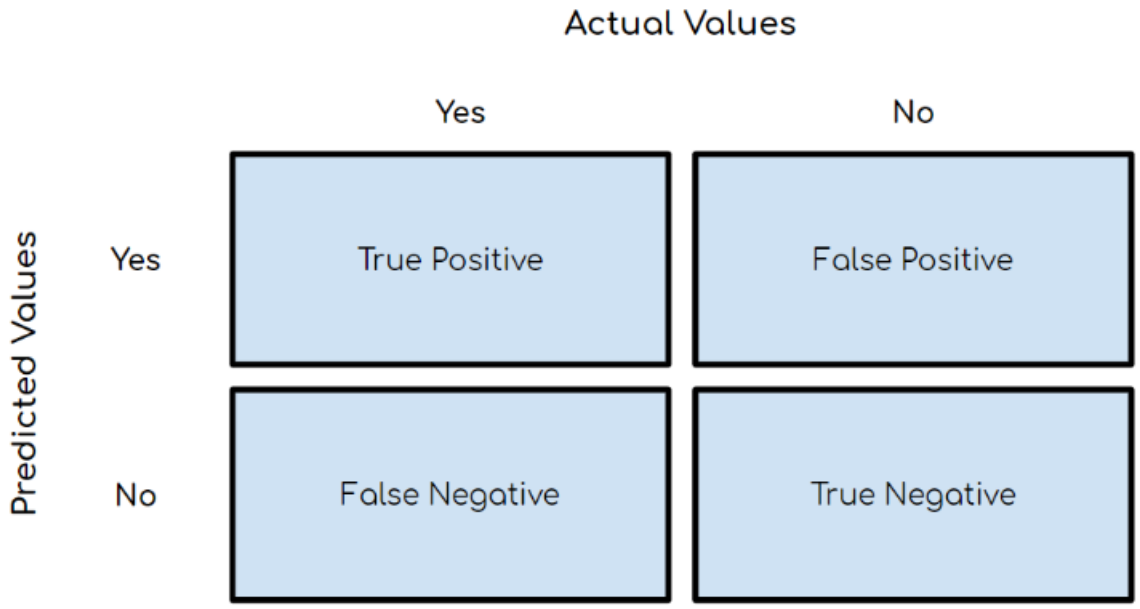

Figura 2: Ejemplo de matriz de confusión de 2x2

- Sensibilidad: indica la proporción de reales positivos que fueron clasificados correctamente, es el cociente entre los verdaderos positivos y reales positivos.
- Especificidad: es el opuesto de la sensibilidad, indica la proporción de reales negativos clasificados correctamente, es el cociente entre los verdaderos negativos y reales negativos.

Cabe aclarar que los **reales positivos** corresponden a la suma de verdaderos positivos y falsos negativos, por otra parte los **reales negativos** corresponden a la suma de verdaderos negativos y falsos positivos.

3. Obtención de árbol

La base de datos requiere de un procesamiento de datos antes de poder trabajar con ellos, para esta experiencia se repitió el mismo procedimiento realizado en la experiencia de laboratorio anterior (Arias and Méndez, 2022).

Como fue mencionado en la sección de introducción, el árbol fue obtenido mediante el algoritmo C5.0 disponible en la librería C50 en R. Como menciona (Gonzalez, 2018), el algoritmo de árbol de decisión corresponde a un algoritmo de aprendizaje supervisado, o sea se le indica que es lo que se está buscando predecir (la clasificación en este caso) y se requiere de la división de los datos en conjuntos de entrenamiento y prueba, con el objetivo de evaluar el modelo de conseguido. Mediante la experimentación de distintos valores, se decidió finalmente el usar valores de 70 % para el conjunto de entrenamiento y 30 % en el conjunto de prueba, ya que permite obtener árboles con más de 3 reglas, presentando árboles que tienen relación con los resultados obtenidos en la experiencia de laboratorio anterior, además fue fijada una semilla con valor 723 en la generación de los conjuntos de entrenamiento y prueba, de forma de obtener el mismo árbol. El árbol generado con las características anteriores se presenta en la Figura 3.

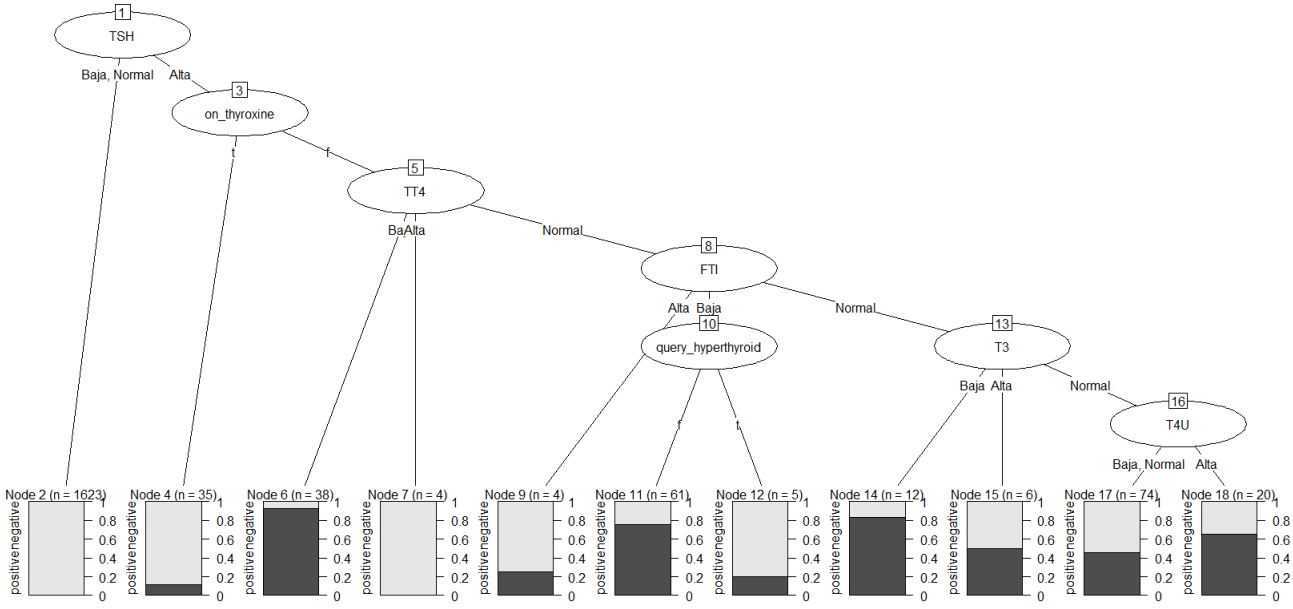

Figura 3: Árbol de decisión generado

En la Tabla 1, se presenta las reglas obtenidas en el árbol, cabe destacar que en

la columna n/m, la cantidad de datos bien clasificados son representados por n y la cantidad de datos mal clasificados por m.

N°	Antecedente	Consecuente	n/m	Confianza	Lift
1	TSH = {Bajo, Normal}	negative	1623/1	0.999	1.1
2	TT4 = Alta	negative	199	0.995	1.1
3	FTI = Alta	negative	337/1	0.994	1.1
4	T3 = Alta	negative	219/3	0.982	1.1
5	on_thyroxine = t	negative	230/4	0.978	1.1
6	T3 = Normal, T4U = {Baja, Normal}, FTI = Normal	negative	810/34	0.957	1.0
7	on_thyroxine = f, TSH = Alta, T3 = Baja	positive	46/4	0.896	11.4
8	on_thyroxine = f, query_hyperthyroid = f, TSH = Alta, FTI = Baja	positive	98/18	0.810	10.3
9	on_thyroxine = f, TSH = Alta, T3 = Normal, TT4 = Normal, T4U = Alta	positive	50/17	0.654	8.3

Tabla 1: Reglas obtenidas

En la Tabla 2, se presenta la matriz de confusión obtenida al evaluar el conjunto de prueba, la exactitud obtenida es de 95.78 %, la sensitividad es de 97.58 % y la especificidad es de 73.77 %. La exactitud indica que 95.78 % de las predicciones fueron correctas, la sensitividad indica que el 97.58 % del total de pacientes negativos fueron bien clasificados, la especificidad indica que el 73.77 % del total de pacientes positivos fueron bien clasificados.

	Negative	Positive
Negative	727	16
Positive	18	45

Tabla 2: Matriz de confusión

4. Comparación de métodos

A continuación, se presentará la comparación entre los resultados conseguidos en esta experiencia de laboratorio con el método de árboles de decisión y los conseguidos en la experiencia de laboratorio anterior con el método de reglas de asociación.

En la Figura 4 se presentan las reglas de asociación conseguidas en la experiencia anterior para la clase de pacientes con hipotiroidismo positivo.

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{thyroid_surgery=f, TSH=Alta, T3=Baja, TT4=Baja}	=>	{classification=positive}	0.01227679	0.9428571	0.01302083	12.01137	33
[2]	{thyroid_surgery=f, TSH=Alta, T3=Baja, TT4=Baja, FTI=Baja}	=>	{classification=positive}	0.01227679	0.9428571	0.01302083	12.01137	33
[3]	{thyroid_surgery=f, psych=f, TSH=Alta, T3=Baja, TT4=Baja}	=>	{classification=positive}	0.01227679	0.9428571	0.01302083	12.01137	33
[4]	{sick=f, thyroid_surgery=f, TSH=Alta, T3=Baja, TT4=Baja}	=>	{classification=positive}	0.01227679	0.9428571	0.01302083	12.01137	33
[5]	{thyroid_surgery=f, tumor=f, TSH=Alta, T3=Baja, TT4=Baja}	=>	{classification=positive}	0.01227679	0.9428571	0.01302083	12.01137	33

Figura 4: Reglas de asociación de la experiencia anterior - Clase positiva

En la Figura 5 se presentan las reglas de asociación conseguidas en la experiencia anterior para la clase de pacientes con hipotiroidismo positivo.

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{TSH=Normal, FTI=Normal}	=>	{classification=negative}	0.4665179	1	0.4665179	1.085184	1254
[2]	{TSH=Normal, TT4=Normal}	=>	{classification=negative}	0.5825893	1	0.5825893	1.085184	1566
[3]	{TSH=Normal, T4U=Normal, FTI=Normal}	=>	{classification=negative}	0.3422619	1	0.3422619	1.085184	920
[4]	{TSH=Normal, T3=Normal, FTI=Normal}	=>	{classification=negative}	0.3917411	1	0.3917411	1.085184	1053
[5]	{TSH=Normal, TT4=Normal, FTI=Normal}	=>	{classification=negative}	0.4471726	1	0.4471726	1.085184	1202

Figura 5: Reglas de asociación de la experiencia anterior - Clase negativa

4.1. Comparación reglas pacientes positivos

En la Figura 6 se presenta una comparación entre las reglas conseguidas con ambos métodos que caracterizan a los pacientes con hipotiroidismo (clase positiva), los antecedentes comunes son marcados con un recuadro del mismo color. Cabe destacar que los niveles de TSH altos caracterizan a los pacientes positivos en ambos métodos, esto se respalda mediante la literatura ya que el hipotiroidismo es detectado a partir de niveles de TSH altos y niveles de tiroxina (T4) bajos (Mayo Clinic, sf). Las reglas obtenidas mediante árbol de decisión no parecen caracterizar al paciente mediante los niveles de T3, ya que se obtuvieron reglas que caracterizan tanto por niveles bajos como normales (regla 7 y 9). Cabe destacar que las reglas 7 y 8 de árbol de decisión son más similares a las reglas conseguidas mediante reglas de asociación, en ambos métodos se tienen niveles altos para TSH y bajos para T3 y T4, con

respecto a los niveles hormonales que caracterizan a un paciente con hipotiroidismo (Mayo Clinic, sf) menciona que el hipotiroidismo es detectado a través de niveles altos de TSH y niveles bajos de tiroxina (T4). Por otra parte, las reglas de asociación caracterizan al paciente siempre por niveles bajos de T3 y T4. En cuanto a los indicadores, las reglas de asociación en la Figura 4 poseen una confianza de 94.28 %, por otro lado mediante árbol de decisión en la Tabla 1 se obtuvieron confianzas menores teniendo un 89.6 % para la regla 7, 81 % para la regla 8 y 65.4 % para la regla 9. El lift también fue superior para las reglas de asociación (Figura 4) teniendo un valor de 12.01, en arból de decisión (Tabla 1) el lift fue de 11.4 para la regla 7, 10.3 para la regla 8 y 8.3 para la regla 9. Si bien ambos métodos poseen valores altos de confianza y lift, las reglas conseguidas mediante el método de reglas de asociación supera en ambos aspectos a las reglas conseguidas mediante árbol de decisión.

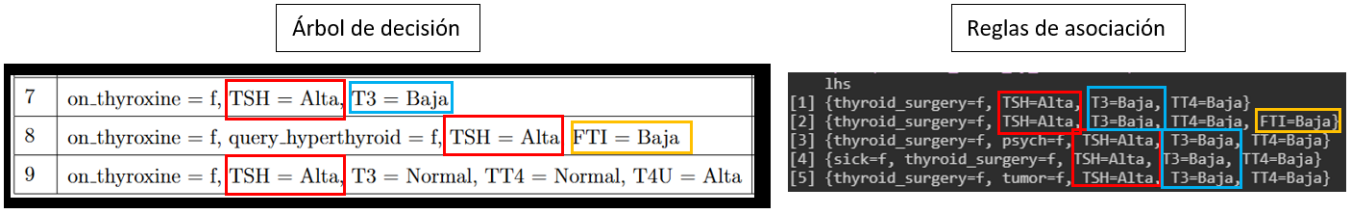

Figura 6: Comparación entre resultados para clase positiva

4.2. Comparación reglas pacientes negativos

En la Figura 7 se presenta una comparación entre las reglas conseguidas con ambos métodos que caracterizan a los pacientes sin hipotiroidismo (clase negativa), los antecedentes comunes son marcados con un recuadro del mismo color. Es posible observar que en este caso las reglas conseguidas con el método de árbol de decisión que más similares son a las reglas de asociación conseguidas son las reglas 1 y 6. La regla 1 menciona que con el antecedente TSH en niveles bajos o normales la persona no tiene hipotiroidismo, lo cual concuerda con la literatura, ya que el hipotiroidismo se caracteriza por niveles altos de TSH (Mayo Clinic, sf). Por otro lado, la regla 6 es la que más se parece a los resultados conseguidos por reglas de asociación, con antecedentes: T3 = Normal, T4U = Normal y FTI = Normal. Las reglas 2, 3, 4 y 5 no presentan semejanza alguna con las reglas de asociación.

Respecto a los indicadores para las reglas con clase negativa, observables en la Figura 5 (reglas de asociación clase negativa) y Tabla 1 (reglas árbol de decisión) es posible

observar que para todas las reglas conseguidas se tiene un lift mayor o igual a 1, lo cual es un buen indicador, a su vez todas tiene una confianza mayor al 95 %. A pesar de estos buenos indicadores hay reglas que poseen un solo antecedente, lo que provoca una confianza mayor. Las reglas de un solo antecedente no son muy interesantes de estudiar. Finalmente cabe destacar que las reglas de asociación de la experiencia anterior también presentan valores de lift mayor a 1, además de confianza del 100 % con reglas con más de un antecedente, entonces al igual que para el caso presentado con las reglas para pacientes positivos, el método de reglas de asociación supera a los resultados conseguidos con árboles de decisión.

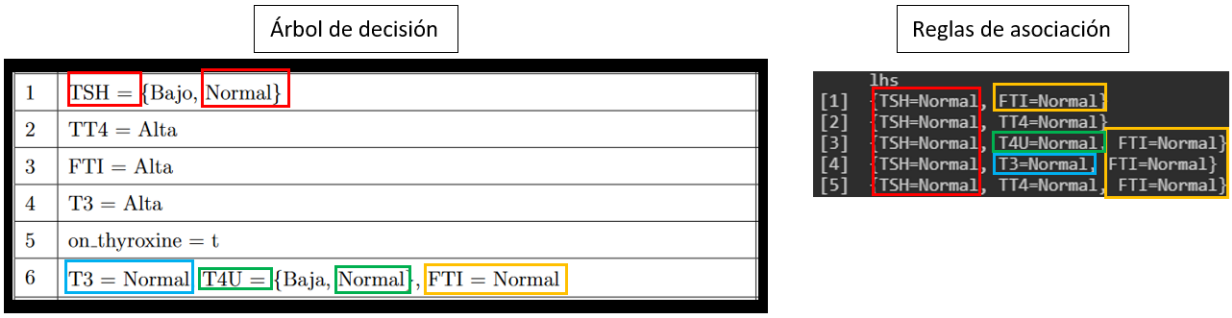

Figura 7: Comparación entre resultados para clase negativa

4.3. Cantidad de reglas

Mediante árbol de decisión en la Tabla 1 se obtuvieron 3 reglas para la clase positiva, por otra parte a través del método de reglas de asociación se obtuvieron 14.009 reglas. Para la clase positiva el método de árbol de decisión en la Tabla 1 obtuvo 6 reglas y el método de reglas de asociación obtuvo 1.379.518 reglas. Cabe destacar que la cantidad de reglas obtenidas mediante reglas de asociación resulta muy superior a las 9 reglas conseguidas mediante el árbol de decisión.

4.4. Forma de entregar los resultados

En las Figura 4 y 5 se puede apreciar cómo se entregan los resultados mediante reglas de asociación. En la Tabla 1 y la Figura 3 se puede apreciar cómo se entregan los resultados mediante árbol de decisión. Si bien ambos métodos entregan parámetros como la confianza y lift, resulta más complicado interpretar las reglas obtenidas mediante reglas

de asociación debido al gran número de reglas obtenidas, además de no contar con una interpretación gráfica de las reglas. Por otra lado, mediante árbol de decisión la cantidad de reglas fue de tan solo 9, además cuenta con la ventaja de una interpretación gráfica mediante un árbol jerárquico binario, donde los nodos internos corresponden a reglas y los nodos hoja a resultados, lo cual hace mucho más intuitiva la interpretación de los resultados obtenidos.

5. Conclusiones

En la presente experiencia de laboratorio se realizó el estudio de la enfermedad hipotiroidismo mediante el método de árboles de decisión, con el algoritmo C5.0 en el software R.

El estudio comenzó estableciendo un marco teórico para la comprensión del método y algoritmo utilizados, además de conceptos útiles. Posteriormente se explicó la forma en la cual se obtuvo el árbol de decisión, detallando los parámetros utilizados para su generación, además se presentó el árbol conseguido en la Figura 3, junto a las reglas en la Tabla 1 y su matriz de confusión en la Tabla 2. En la siguiente sección se efectuó la comparación y análisis de los resultados obtenidos con los resultados obtenidos en la experiencia anterior de laboratorio con el método de reglas de asociación, donde además se efectuó una comparación con la literatura.

Respecto a los resultados relevantes encontrados durante el estudio, se puede mencionar que el modelo obtenido con el método de árbol de decisión presenta los siguientes valores para la predicción de la clase de hipotiroidismo: exactitud obtenida de 95.78 %, sensitividad de 97.58 % y especificidad de 73.77 %. Ante estos resultados se puede concluir que el modelo tiene buen poder de predictivo, sin embargo, predice de manera más precisa los casos de pacientes sin la enfermedad (hipotiroidismo negativo). En cuanto a las reglas conseguidas se obtuvieron 9 reglas observadas en la Tabla 1, 3 para pacientes con hipotiroidismo (clase positiva) y 6 para pacientes no enfermos (clase negativa). Las reglas conseguidas para pacientes con hipotiroidismo se ven respaldadas por la literatura ya que (Mayo Clinic, sf) indica que los pacientes con hipotiroidismo son identificados a partir de niveles de TSH altos y niveles de T4 bajos. Las reglas conseguidas para pacientes negativos respaldadas por la literatura son la regla 1 que indican niveles de TSH bajos o normales y la regla 6 que indica niveles de T3 y T4 normales, como menciona (Mayo Clinic, sf) el hipotiroidismo es identificado mediante niveles alterados de TSH y T4.

Respecto a las reglas conseguidas, las conclusiones más relevantes son que se consiguieron algunas reglas que describían las mismas características de la enfermedad y además se veían respaldadas por la literatura, como lo son las reglas 7, 8 y 9 que presentan argumentos como TSH = Alta, FTI = Baja y T3 = Baja para la clase positiva. En el caso de la clase negativa se presentan las reglas 1 y 6 con resultados apegados a la literatura TSH

= Normal, T3 = Normal, T4U = Normal, FTI = Normal. Las reglas enumeradas también aparecen en las reglas de asociación de la experiencia anterior. Al efectuar una comparación con los resultados conseguidos con el método de reglas de asociación de la experiencia de laboratorio anterior, es posible observar que las reglas más importantes conseguidas: 1,6,7,8 y 9 también aparecen en las reglas de asociación, recordar que estas reglas tienen apoyo en la literatura según (Mayo Clinic, sf). Sin embargo, el método de árboles de decisión también generó reglas que no tienen sentido con el problema o lo estudiado en la literatura, cosa que no ocurrió con el método de reglas de asociación. Respecto a los índices, se pudo observar que las reglas de asociación siempre presentaron un mayor lift y confianza para todas las reglas en general en comparación a las reglas del árbol de decisión. Es por estos motivos que se concluye que para estudiar este problema el método de reglas de decisión lo realizó de mejor forma.

Respecto a los objetivos específicos detallados en la sección de introducción fueron cumplidos en su totalidad se extrajo conocimiento relevante del hipotiroidismo implementando algoritmos de árboles de decisión en el software R usando la librería C50. Estos resultados fueron comparados con la literatura encontrada, otorgándoles sentido. Finalmente se realizó una comparación con los resultados obtenidos de la experiencia de laboratorio anterior, por lo que se cumplieron todos los objetivos propuestos.

Para concluir, respecto a las posibles mejoras en el desarrollo de esta experiencia se propone el repetir la experiencia de laboratorio, pero utilizando un nuevo algoritmo de árboles de decisión más populares, como por ejemplo random forest, Random Forest se considera como la “panacea” en todos los problemas de ciencia de datos (Orellana, 2018). Quizá modificando el algoritmo de generación del árbol se hubiese logrado conseguir mejores resultados que el método de reglas de asociación.

Bibliografía

Amat, J. (2017). Árboles de decisión, random forest, gradient boosting y c5. 0. *RPubs by RStudio [en línea], disponible en: https://rpubs. com/Joaquin_AR/255596*.

Arias, I. and Méndez, C. (2022). Laboratorio 3 – reglas de asociación.

Barea, I. T., Blanco, M. C., Sánchez, C. C., and Aguilar-Diosdado, M. (2012). Hipotiroidismo. *Medicine-Programa de Formación Médica Continuada Acreditado*, 11(14):819–826. doi: 10.1016/S0304-5412(12)70390-6.

Brent, G. A. (2022). Thyroid hormone action. *UpToDate*. https://www.uptodate.com/contents/thyroid-hormone-action.

Chiovato, L., Magri, F., and Carlé, A. (2019). Hypothyroidism in context: where we've been and where we're going. *Advances in therapy*, 36(2):47–58.

Gonzalez, L. (2018). Aprendizaje supervisado: Decision tree classification. https://aprendeia.com/aprendizaje-supervisado-decision-tree-classification/.

IBM (s.f.). Árboles de decisión. https://www.ibm.com/es-es/topics/decision-trees.

Mayo Clinic (s.f.). Hipotiroidismo (tiroides hipoactiva). https://www.mayoclinic.org/es-es/diseases-conditions/hypothyroidism/diagnosis-treatment/drc-20350289.

Orellana, J. (2018). Ensambladores: Random forest. https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html.

Parra, F. (2019). Métodos de clasificación. https://bookdown.org/content/2274/metodos-de-clasificacion.html.

Shin, T. (2020). Comprensión de la matriz de confusión y cómo implementarla en python.