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

Laboratorio 1 - Análisis Estadístico

Integrantes: Matías Escudero
Joaquín Macías

Curso: Análisis de Datos

Sección A-1

Profesor: Max Chacón Pacheco

9 de Junio de 2023

Tabla de contenidos

1. Introducción

El hipertiroidismo (Cooper et al., 2006) es una afección en la cual la glándula tiroides produce demasiadas hormonas tiroideas. La tiroides es una pequeña glándula en forma de mariposa ubicada en la parte frontal del cuello y juega un papel importante en la regulación del metabolismo del cuerpo.

El tratamiento del hipertiroidismo depende de la causa subyacente y puede incluir medicamentos para reducir la producción de hormonas tiroideas, terapia con yodo radioactivo o, en casos graves, cirugía para extirpar la tiroides. Con el tratamiento adecuado, la mayoría de las personas con hipertiroidismo pueden controlar sus síntomas y llevar una vida normal y saludable.

En el presente informe se busca analizar y comprender un conjunto de datos relacionados a la enfermedad de hipertiroidismo con la finalidad de identificar patrones y factores importantes que puedan contribuir con el diagnóstico y tratamiento de esta enfermedad. Para esto, se utiliza un conjunto de datos proveniente del repositorio UCI Machine Learning (for Machine Learning and at the University of California, 2007), que almacena múltiples conjuntos de datos para aprendizaje automático (machine learning) y minería de datos (data mining) creada y mantenida por la Universidad de California en Irvine (UCI).

Además de proporcionar los datos, UCI Machine Learning también ofrece herramientas y recursos útiles para ayudar a los usuarios a comprender y trabajar con los datos, incluyendo descripciones detalladas de los conjuntos de datos, tutoriales y ejemplos de código en diferentes lenguajes de programación.

A continuación, se abordarán los objetivos de este primer laboratorio, descripción del problema desde el propio contenido de la base de datos, las variables y clases asociadas; y un análisis estadístico (descriptivo) e inferencial de nuestro conjunto de datos.

1.1. Objetivos

1.1.1. Objetivo General

- Estudiar e interpretar la información correspondiente a un conjunto de datos relacionado a la enfermedad de hipertiroidismo.

1.1.2. Objetivos Específicos

1. Aplicar técnicas de estadística descriptiva e inferencial al conjunto de datos a estudiar.
2. Utilizar técnicas de visualización de datos para representar gráficamente las relaciones entre atributos y clases.
3. Limpiar y preprocesar el conjunto de datos de manera adecuada para su uso en futuras experiencias de laboratorio.
4. Interpretar la correlación entre variables de nuestro conjunto de datos.
5. Construir una regresión logística entre nuestra variable dependiente y una o más variables independientes.

2. Descripción del Problema

2.1. Descripción de la Base de Datos

2.1.1. Origen y Contexto

La base de datos utilizada para el desarrollo del presente informe fue otorgada por la Universidad de California, sede de Irvine, por el investigador Ross Quinlan (Quinlan, 2007), durante su visita en 1987 para un taller de aprendizaje automático de ese mismo año. Si bien se otorgó una base de datos completa para el presente laboratorio, se utiliza y describe particularmente el conjunto de datos *allhyper*, asociado a la enfermedad de hipertiroidismo y otras condiciones asociadas a la producción excesiva de hormonas tiroideas. Este conjunto de datos se creó a partir de datos médicos y demográficos de pacientes que fueron evaluados por diversas condiciones relacionadas con la producción excesiva de hormonas tiroideas.

2.1.2. Objetivo del Conjunto de Datos

El objetivo del conjunto de datos *allhyper* es facilitar el desarrollo y la evaluación de modelos de predicción para la identificación de condiciones relacionadas con la producción excesiva de hormonas tiroideas, como el **hipertiroidismo**, la **tirotoxicosis por T3** y el **bocio multinodular tóxico**. Estos modelos se basan en técnicas de aprendizaje automático que pueden ayudar a mejorar la precisión y la eficiencia en el diagnóstico de estas condiciones.

2.1.3. Dimensión del Conjunto de Datos

El conjunto de datos *allhyper* consta de 2,800 registros de pacientes, más 972 registros para pruebas de rendimiento del modelo, cada uno de los cuales contiene información demográfica, clínica y de diagnóstico relevante para las condiciones relacionadas con la producción excesiva de hormonas tiroideas. Cada registro se compone de 30 variables numéricas y categóricas, las cuales hacen referencia variables como la edad, el sexo, los niveles de hormonas tiroideas y otras características clínicas relevantes para el diagnóstico de estas condiciones médicas.

2.2. Descripción de Variables y Clases

Se especifican las distintas variables y clases del conjunto de datos a utilizar. Para cada variable se indica una breve descripción, el tipo de variable a la que corresponde y los posibles valores que esta puede tomar.

2.2.1. Variables Demográficas

- **age**: Variable numérica discreta que indica la edad en años del paciente.
- **sex**: Variable categórica que indica el sexo del paciente ('M' para masculino y 'F' para femenino).
- **referral source**: Variable categórica que indica la fuente de derivación del paciente. Las distintas fuentes de derivación que existen para el set de datos estudiado son: 'SVHC', 'SVI', 'STMW', 'SVHD' y 'other'.

2.2.2. Variables Clínicas

- **on thyroxine**: Variable booleana que indica si el paciente está tomando tiroxina (medicamento para el tratamiento contra hipotiroidismo).
- **query on thyroxine**: Variable booleana que indica si el paciente ha consultado sobre el uso de tiroxina.
- **on antithyroid medication**: Variable booleana que indica si el paciente está tomando medicamentos antitiroideos.
- **sick**: Variable booleana que indica si el paciente está enfermo.
- **pregnant**: Variable booleana que indica si la paciente está embarazada.
- **thyroid surgery**: Variable booleana que indica si el paciente ha tenido cirugía de tiroides.
- **I131 treatment**: Variable booleana que indica si el paciente ha recibido tratamiento con yodo radiactivo.

- **query hypothyroid**: Variable booleana que indica si el paciente ha consultado sobre hipotiroidismo.
- **query hyperthyroid**: Variable booleana que indica si el paciente ha consultado sobre hipertiroidismo.
- **lithium**: Variable booleana que indica si el paciente está tomando litio.
- **goitre**: Variable booleana que indica si el paciente tiene bocio.
- **tumor**: Variable booleana que indica si el paciente tiene un tumor en la glándula tiroides).
- **hypopituitary**: Variable booleana que indica si el paciente tiene hipopituitarismo.
- **psych**: Variable booleana que indica si el paciente tiene antecedentes psiquiátricos.

2.2.3. Variables de Diagnóstico

- **TSH measured**: Variable booleana que indica si el paciente se midió la hormona estimulante de la tiroides (TSH).
- **TSH**: Variable numérica continua que indica el nivel de hormona estimulante de la tiroides (TSH).
- **T3 measured**: Variable booleana que indica si el paciente se midió la hormona tiroidea triyodotironina (T3).
- **T3**: Variable numérica continua que indica el nivel de hormona tiroidea triyodotironina (T3).
- **TT4 measured**: Variable booleana que indica si el paciente se midió la tiroxina total (TT4).
- **TT4**: Variable numérica continua que indica el nivel de tiroxina total (TT4).
- **T4U measured**: Variable booleana que indica si el paciente se midió la tiroxina no unida (T4U).

- **T4U**: Variable numérica continua que indica el nivel de tiroxina no unida (T4U).
- **FTI measured**: Variable booleana que indica si el paciente se midió el índice de tiroxina libre (FTI).
- **FTI**: Variable numérica continua que indica el índice de tiroxina libre (FTI).
- **TBG measured**: Variable booleana que indica si el paciente se midió la globulina transportadora de tiroxina (TBG).
- **TBG**: Variable numérica continua que indica el nivel de globulina transportadora de tiroxina (TBG).

2.2.4. Clase

- **class**: Variable categórica que hace referencia a la **condición tiroidea** del paciente. Al estudiar el comportamiento de la variable, se infiere que esta sigue la notación ***condición*|*identificador***, donde *condición* representa la naturaleza médica asociada al paciente e *identificador* es un número entero único que determina cada instancia de la condición. Quitando el identificador asociado a cada clase, se tienen cuatro posibles condiciones:

  1. **Hyperthiroid**: Clase que indica que el paciente tiene **hipertiroidismo**, condición en la cual la glándula tiroides expulsa una mayor cantidad de hormona tiroidea de la que el cuerpo necesita.
  2. **Goitre**: Clase que indica que el paciente tiene **bocio**, condición en la cual se presenta un agrandamiento de la glándula tiroides (James, 1972).
  3. **T3 Toxic**: Clase que indica que el paciente tiene **toxicidad de T3 o T3-toxicosis**, la cual es una clase de hipertiroidismo en donde los niveles de la hormona T3 son elevados, y los de hormonas TSH y T4 son normales (Sriphrapradang and Bhasipol, 2016).
  4. **Negative**: El paciente **no tiene ninguna de las condiciones asociadas a niveles elevados de hormonas tiroideas**, anteriormente mencionadas.

3. Análisis Estadístico e Inferencial

3.1. Análisis Descriptivo

Se realiza un análisis descriptivo de las distintas variables del conjunto de datos utilizado. Para esto, en primera instancia, se generan tablas de resumen estadístico separadas según el tipo de dato al que corresponden, entendiendo que en el conjunto de datos existen variables categóricas, numéricas y booleanas.

Variable	Máximo	Promedio	Mediana	Mínimo	NA	Q1	Q3	D.E
age	455.00	51.84	54.00	1.00	1.00	36.00	67.00	20.46
FTI	395.00	110.79	107.00	2.00	295.00	93.00	124.00	32.88
TBG					2800.00			
T3	10.60	2.02	2.00	0.05	585.00	1.60	2.40	0.82
T4U	2.12	1.00	0.98	0.31	297.00	0.88	1.08	0.19
TSH	478.00	4.67	1.40	0.00	284.00	0.44	2.60	21.45
TT4	430.00	109.07	104.00	2.00	184.00	88.00	125.00	35.39

Cuadro 1: Resumen estadístico de variables numéricas.

A partir de las estadísticas que se muestran en el Cuadro 1, se puede inferir:

- La variable de edad *Age* tiene un rango muy amplio, desde 1 hasta 455 años, lo cual sugiere un error en los datos, ya que es poco probable que se le realice una prueba de tiroides a un niño tan pequeño y es imposible que un humano llegue a tener esa edad. Se deben evaluar bajo los siguientes puntos:

  - Según (Claudia Godoy C., 2009), se realizó una revisión retrospectiva de cuadros clínicos de niños menores de 15 años, entre junio de 2004 y agosto de 2005, donde el diagnóstico de hipertiroidismo se efectuó con TSH suprimida y niveles elevados de hormonas tiroideas, concluyendo que el bocio fue el síntoma de HTA pediátrica más frecuente y la enfermedad de Graves la principal etiología.
  - En el caso de edad *Age* de 455 y similares **se descartarán para este análisis ya que corresponden a valores erróneos**.

- De acuerdo a las variables FTI, $T3$, $T4U$, TSH y $TT4$:
  - Tienen valores faltantes, lo que sugiere que no todos los pacientes tienen registros completos para estas pruebas.
  - Se encontraron rangos bastante amplios, lo cual sugiere una alta variabilidad en los datos. Además, en algunas de estas variables se tiene una alta desviación estándar, lo cual indica alta dispersión.
  - En los casos de FTI y $TT4$, las medias y medianas son bastante parecidas, lo cual sugiere una distribución similar a la normal.
- La variable TBG presenta únicamente valores nulos, por lo cual no sirve para el análisis y será descartada.

Variable	Categoría	Frecuencia	Porcentaje
sex	F	1830	65.36
	M	860	30.71
	N.A	110	3.93
referral source	other	1632	58.29
	STMW	91	3.25
	SVHC	275	9.82
	SVHD	31	1.11
	SVI	771	27.54
class	goitre	7	0.25
	hyperthyroid	62	2.21
	negative	2723	97.25
	T3 toxic	8	0.29

Cuadro 2: Resumen estadístico de variables categóricas.

A partir de el Cuadro 2, es posible inferir que:

- **La mayoría de los pacientes del conjunto de datos son mujeres (65.36 %)**, existiendo una **alta concentración (densidad) con una cantidad de 1830**. Es importante considerar esta proporción en futuros análisis, ya que podría afectar los resultados si no se ajusta de manera adecuada.

    - De acuerdo a la densidad encontrada en esta variable, (Romero, 2014) menciona que según un estudio se hallaron 29.947 pacientes tratados para hipotiroidismo en 82 ciudades colombianas. **La mayoría (79,1 %) eran mujeres, con una edad media de 63,2 $\pm$ 16,1 años**.

- La gran mayoría de los paciente provienen de la fuente *other* **(58,29 %) (1632)**, seguida por SVI **(27,54 %) (771)**. El resto de las categorías son significativamente menores, por lo cual se debe tener en cuenta en caso de considerar esta variable dentro del análisis.

- Para la variable de clase se tiene que **la gran mayoría da negativo para hipertiroidismo con un 97.25 % de los pacientes (2723)**, seguido de *hyperthyroid* con un 2.21 % (62). Es importante tener en cuenta esta amplia diferencia entre clases ya que los análisis posteriores podrían verse afectados por este desequilibrio.

- A partir del análisis anterior **se decide eliminar todos los registros que pertenezcan a las clases** *goitre* **y** $T3\ Toxic$, de esta forma se genera la variable booelana de clase *hyperthiroid* que **indicará si un paciente tiene diagnóstico por hipertiroidismo o no**.

Variable	N° Falso	N° Verdadero	N° NA	% Verdadero	% Falso
FTI_measured	295	2505		89.46	10.54
goitre	2775	25		0.89	99.11
hypopituitary	2799	1		0.04	99.96
I131_treatment	2752	48		1.71	98.29
lithium	2786	14		0.50	99.50
on_antithyroid_medication	2766	34		1.21	98.79
on_thyroxine	2470	330		11.79	88.21
pregnant	2759	41		1.46	98.54
psych	2665	135		4.82	95.18
query_hyperthyroid	2627	173		6.18	93.82
query_hypothyroid	2637	163		5.82	94.18
query_on_thyroxine	2760	40		1.43	98.57
sick	2690	110		3.93	96.07
T3_measured	585	2215		79.11	20.89
T4U_measured	297	2503		89.39	10.61
TBG_measured	2800	0		0.0	100.0
thyroid_surgery	2761	39		1.39	98.61
TSH_measured	284	2516		89.86	10.14
TT4_measured	184	2616		93.43	6.57
tumor	2729	71		2.54	97.46

Cuadro 3: Resumen estadístico variables booleanas.

A partir del Cuadro 3 sobre las variables booleanas, se puede inferir:

- **La gran mayoría de las variables booleanas presentan una distribución muy desequilibrada**, con un alto porcentaje de valores $False$ en la mayoría de los casos. Esto quiere decir que la mayoría de los clientes no padecen o presentan las características o condiciones que estas variables desean medir.

- Las variables relacionadas a mediciones presentan valores $True$ relativamente altos, lo

cual sugiere que la mayoría de los pacientes tiene resultados para estos índices y pueden utilizarse para estudiar su relación con la variable de clase.

- Por otra parte, las variables *on thyroxine*, *query hyperthyroid*, y *query hypothyroid* presentan valores $True$ más altos.

- La variable TBG *measured* presenta únicamente valores $False$, lo cual explica por qué la variable TBG presenta únicamente valores nulos. También será descartada de futuros análisis.

3.2. Análisis Gráfico

Se analiza mediante un gráfico de violín unificado la distribución, densidad de probabilidad y posibles outliers que presentan las variables numéricas.

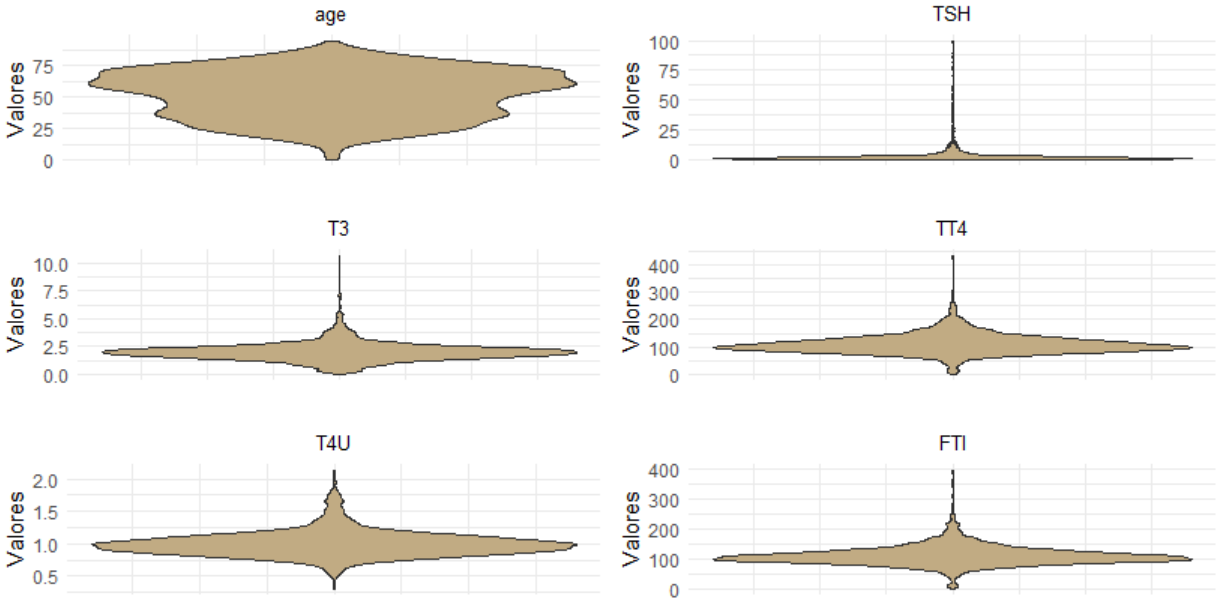

Figura 1: Gráficos de violín para las variables numéricas.

De acuerdo a la Figura 1 es posible inferir que (con datos de referencia Cuadro 1):

- Para la variable *Age*, se presenta una **alta concentración (densidad), entre los valores entre 50 a 60 años**, con una mediana en 54,00, y su rango intercuartílico entre 36 y 67. Es posible notar una leve inclinación hacía la parte superior del violín.

- Para la variable TSH existen valores mayores a 100 (0.6 % de los datos) que se escapan totalmente de las tendencias que presenta esta variable y no permitían siquiera visualizar correctamente el gráfico, por lo cual se eliminan ya que corresponden a outliers. Posterior a esto, se presenta una **alta concentración (densidad), entre los valores entre 0 a 2**.

- Para la variable $T3$ se muestra una **alta concentración (densidad), entre los valores entre 1,5 a 2,5**, con una mediana en 2,00, y su rango intercuartílico entre 1,60 y 2,40. Acá es posible visualizar outliers para valores cercanos a 10, por lo cual se eliminarán estos registros.

- Para la variable $TT4$ se presenta una **alta concentración (densidad), entre los valores entre 100 a 110**. Acá es posible observar outliers para valores cercanos a los 400, los cuales serán removidos del conjunto de datos.

- Para la variable $T4U$ se presenta una **alta concentración (densidad), entre los valores entre 0,90 a 1,10**. Para este caso particular no se observan outliers de acuerdo al gráfico.

- Para la variable FTI se muestra una **alta concentración (densidad), entre los valores entre 100 a 110**. Acá se observan outliers cercanos a los 400, los cuales serán removidos.

3.3. Estudio de Correlaciones

3.3.1. Variables numéricas

Una matriz de correlación es una tabla que indica los coeficientes de correlación de Pearson entre dos variables numéricas. Cada celda de la tabla muestra la conexión entre los dos factores.

Entendiendo que se estudia la correlación existente entre variables independientes de un eventual modelo, es importante notar qué variables de estas están fuertemente correlacionadas para evitar multicolinealidad en el modelo (Universidad de Granada, 1900). A continuación

se muestra la matriz de correlación, para las variables numéricas de nuestro conjunto de datos. Según la Figura 2, es posible inferir:

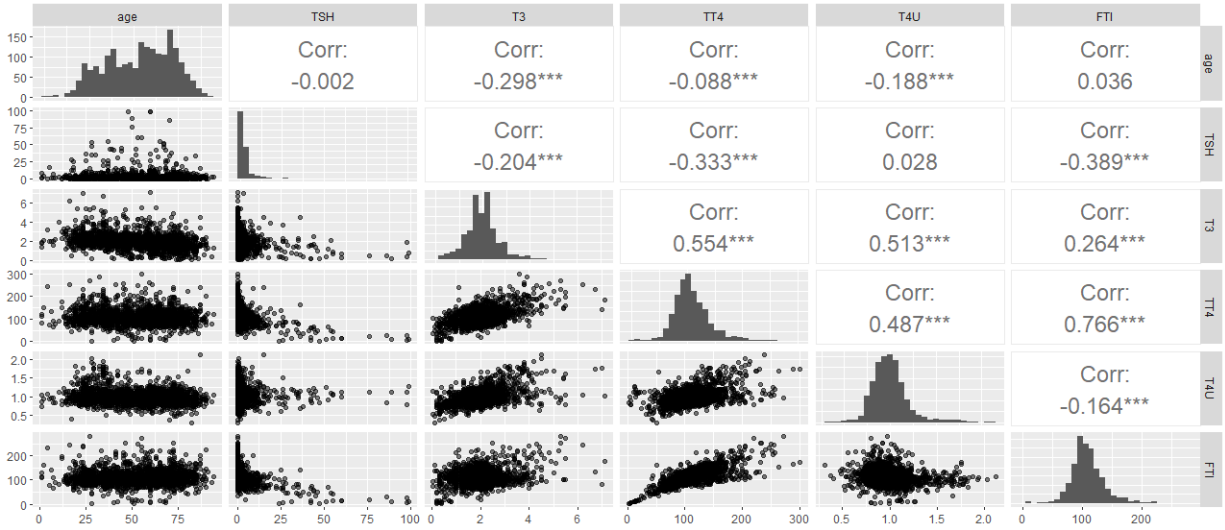

Figura 2: Matriz de correlación para las variables numéricas.

- Existe una correlación **fuertemente positiva entre** FTI **y** $TT4$ **con un valor 0,766**, justificada por que en el hipertiroidismo, lleva a niveles elevados de T4 libre y T4 total en la sangre. En consecuencia, **tanto el** FTI **como el** $TT4$ **estarán elevados en la mayoría de los casos de hipertiroidismo**. En cambio, en el hipotiroidismo, la glándula tiroides no producen suficientes hormonas tiroideas, lo que lleva a niveles bajos de T4 libre y T4 total en la sangre. **Tanto el** FTI **como el** $TT4$ **estarán disminuidos en la mayoría de los casos de hipotiroidismo** (CIGNA, 2022).

- Existe una correlación **medianamente positiva entre** $TT4$ **y** $T3$ **con un valor 0,554**, justificada por que **en el hipertiroidismo, los niveles de T4 y T3 estarán elevados en la sangre**. Los niveles de T4 pueden ser más elevados que los de T3 debido a la mayor producción de T4 por parte de la tiroides, aunque los niveles de T3 también pueden estar elevados debido a la conversión de T4 en T3. En el hipotiroidismo, la glándula tiroides no produce suficientes hormonas tiroideas, incluyendo tanto T4 como T3. En general, **en el hipotiroidismo, los niveles de T4 y T3 estarán disminuidos en la sangre** (MedlinePlus, 2000).

- Existe una correlación **medianamente positiva entre** $T4U$ **y** $T3$ **con un valor 0,513**, justificada por que **en el hipertiroidismo, los niveles de T4U pueden estar disminuidos debido a una mayor unión de T4 y T3 a las proteínas transportadoras en la sangre**. Además, los niveles de T3 también pueden estar elevados debido a la conversión periférica de T4 en T3. En el hipotiroidismo, **los niveles de T4U pueden estar elevados debido a una menor unión de T4 y T3 a las proteínas transportadoras en la sangre**, ya que hay menos hormonas tiroideas disponibles para unirse a ellas.

- Existe una correlación **medianamente negativa entre** FTI **y** TSH **con un valor -0.389**, justificada por que **en el hipertiroidismo, se puede observar una disminución en los niveles de TSH y un aumento en los niveles de FTI**. Sin embargo, es importante tener en cuenta que la relación entre los niveles de FTI y TSH puede variar según la causa subyacente del hipertiroidismo y la gravedad de la enfermedad. En el hipotiroidismo, **se puede observar un aumento en los niveles de TSH y una disminución en los niveles de FTI**. Sin embargo, es importante tener en cuenta que la relación entre los niveles de FTI y TSH puede variar según la causa subyacente del hipotiroidismo y la gravedad de la enfermedad.

3.3.2. Variables numéricas v/s Variable de clase

Se realizan gráficos de caja para comparar el comportamiento de las variables numéricas en pacientes con hipertirodismo v/s pacientes sin la enfermedad, y a partir de esto, se determina qué variables parecen tener más correlación con la variable dependiente, que según lo encontrado, podríamos incluirlas en una primera instancia del modelo de regresión que se desea construir.

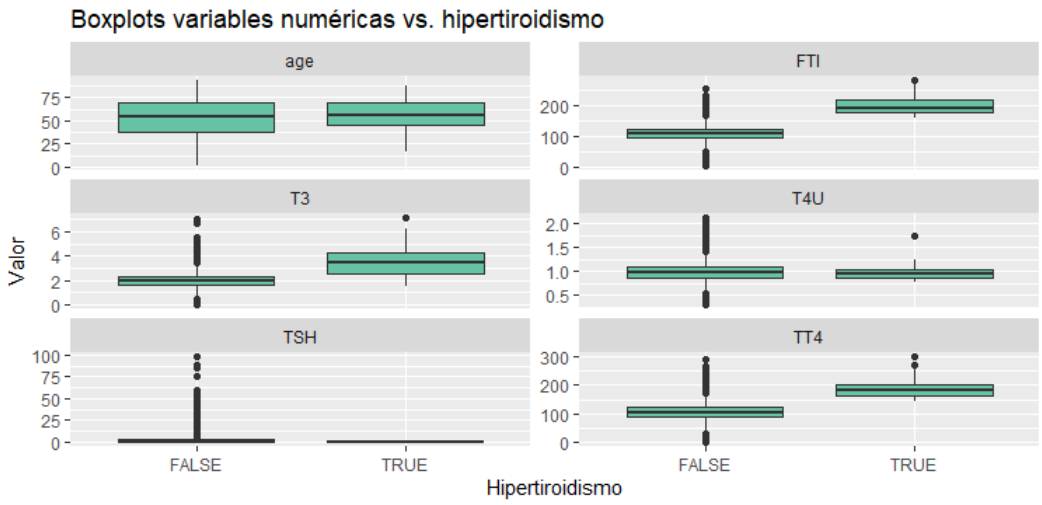

Figura 3: Gráficos de caja para las variables numéricas de pacientes con y sin hipertiroidismo.

A partir de la Figura 3, es posible inferir que para las seis variables numéricas estudiadas, se observan diferencias notables entre cada una de las generadas. Si bien hay casos en los que esta diferencia es más notoria, en todos los casos existe una diferencia entre las medianas de ambas cajas, o en sus rangos intercuartílicos, o en la presencia por outliers, por lo cual, en primera instancia **serán todas incluidas cómo variables independientes en el modelo a construir utilizando regresión logística** (S., 2014).

3.3.3. Variables categóricas v/s Variable de clase

En vista de que no es posible estudiar la correlación entre la variable dependiente, las variables booleanas y categóricas mediante la matriz de correlación utilizada anteriormente, **se genera una tabla de contingencia mediante el cálculo de independencia de chi cuadrado** (IBM, 2021). Esto indicará si existe una asociación significativa entre la variable de clase y las variables independientes que no son numéricas. De esta forma, se hace este test para todas las variables y se obtiene un p-value para cada una. A continuación se muestran únicamente las variables que presentan un p-value menor a 0.05.

Para la construcción de nuestro primer modelo se utilizan únicamente las variables booleanas y categóricas, según: $TSH\ Measured$, $T3\ measured$, $TT4\ measured$, $FTI\ measured$, $query\ hyperthyroid$ y sex.

Variable	$p - value$
$TSH\ measured$	0.0000
$T3\ measured$	0.0000
$TT4\ measured$	0.0000
$FTI\ measured$	0.0000
$query\ hyperthyroid$	0.0000
sex	0.0006

Cuadro 4: Valores de p-value para pruebas de independencia chi cuadrado.

3.3.4. Test de Welch

La prueba t de Welch (Pablo Livavic-Rojas, 2006), es la t de dos muestras se utiliza para comparar las medias de dos conjuntos de datos independientes diferentes. Sin embargo, es posible aplicar una prueba T de dos muestras en aquellos grupos de datos que comparten la misma varianza. Ahora, para comparar dos grupos de datos que tienen diferentes varianzas, usamos la prueba t de Welch. Se considera el equivalente paramétrico de la prueba T de dos muestras.

A continuación se especifican un caso de estudio particular, seleccionando variables relevantes a partir del análisis de correlación, con el fin de determinar la influencia de ellas entorno a la clase:

Caso 1: El tratamiento de litio y su afectación con la T3.
Al momento de calcular las varianzas de ambas variables, se obtiene **0.00545 para la variable** $lithium$ y **0.611 para la variable** $T3$. Como la diferencia de varianza es superior a 4:1, es posible ejecutar la prueba t de Welch, de acuerdo a las siguientes hipótesis:

H_0: El tratamiento de litio disminuyendo la disponibilidad de T3 en pacientes con este tipo de tratamiento (carbonato de litio).

H_1: El tratamiento de litio no disminuye la disponibilidad de T3 en pacientes con este tipo de tratamiento (carbonato de litio).

Ahora, el cálculo de t de Welch para ambas variables es dado por el **estadístico 0.20136** y el valor **p-value 0.8443**. Dado lo anterior, no es posible rechazar la hipótesis nula y concluir que **no hay evidencia suficiente para afirmar que existe una diferencia significativa en la disponibilidad de T3 entre pacientes que reciben tratamiento con litio y aquellos que no lo reciben**.

3.4. Regresión Logística

En vista de que se busca explicar el comportamiento de una variable booleana en función de otras variables independientes, **se utiliza una regresión logística**. Para la selección de variables, se determinan, en primera instancia, aquellas que creemos relevantes de acuerdo al análisis de correlaciones, para posteriormente, aplicar un método de selección de variables por paso. En primera instancia, el modelo generado es el siguiente:

$$\begin{aligned}\text{logit}(p) = \beta_0 &+ \beta_1 \cdot \text{age} + \beta_2 \cdot \text{FTI} + \beta_3 \cdot \text{T3} + \beta_4 \cdot \text{T4U} \\ &+ \beta_5 \cdot \text{TSH} + \beta_6 \cdot \text{TT4} + \beta_7 \cdot \text{TSH\_measured} \\ &+ \beta_8 \cdot \text{T3\_measured} + \beta_9 \cdot \text{TT4\_measured} \\ &+ \beta_{10} \cdot \text{FTI\_measured} + \beta_{11} \cdot \text{query\_hyperthyroid} + \beta_{12} \cdot \text{sex}_M \end{aligned} \tag{1}$$

Tras la aplicación de la selección de variables mediante el método *Stepwise* (IBM, 2022), el modelo entrega los siguientes resultados con las variables seleccionadas:

Variable	**Coeficiente**	**Error estándar**	**p-value**
Intercepto	-2.24	1.30	0.0851
T3	0.65	0.23	0.00471
T4U	-9.07	1.76	2.53e-07
TSH	-4.20	2.06	0.0412
TT4	0.05	0.0083	1.91e-09
sexM	-1.52	0.74	0.0406

Cuadro 5: Coeficientes del modelo de regresión logística.

Quedando la ecuación de la siguiente manera:

$$\text{logit}(p) = -2{,}24 + 0{,}65 \cdot T3 - 9{,}07 \cdot T4U - 4{,}20 \cdot TSH + 0{,}05 \cdot TT4 - 1{,}52 \cdot \text{sexM} \qquad (2)$$

A partir del Cuadro 5 y la Ecuación 2, es posible inferir:

- Un aumento en la variable $T3$ está asociada con un aumento en la probabilidad de tener hipertiroidismo. Particularmente, **por cada unidad de medida adicional en $T3$, la probabilidad de tener hipertiroidismo aumenta aproximadamente 1.91 veces**. Además, de acuerdo a su $p - value$ es estadísticamente significativa.

- Un aumento en la variable $T4U$ está asociada con una disminución en la probabilidad de tener hipertiroidismo. De esta forma, por **cada unidad de medida adicional en $T4U$, la probabilidad de tener hipertiroidismo disminuye aproximadamente 8615 veces**. Esta variable es **altamente significativa**, con un $p-value$ de $2{,}53 \times 10^{-7}$.

- Un aumento en la variable TSH se asocia con una disminución en la probabilidad de tener hipertiroidismo. **Por cada unidad adicional en TSH, la probabilidad de tener hipertiroidismo disminuye aproximadamente 67 veces**. También es estadísticamente significativa.

- Un aumento en la variable $TT4$ se asocia con un aumento en la probabilidad de tener hipertiroidismo. **Por cada unidad de medida adicional en $TT4$, la probabilidad de tener hipertiroidismo aumenta aproximadamente 1.05 veces**. Además esta variable es altamente significativa con un $p - value$ de $1{,}91 \times 10^{-9}$.

- Ser hombre ($sexM$) está asociado con una disminución en la probabilidad de tener hipertiroidismo. Comparado con las mujeres, **los hombres tienen aproximadamente 4.57 veces menos probabilidad de tener hipertiroidismo**. Es estadísticamente significativa de acuerdo al $p-value$ y esperable de acuerdo a la bibliografía mencionada anteriormente.

En términos generales, el modelo parece ser adecuado para predecir la presencia de hipertiroidismo utilizando las variables seleccionadas por el método *Stepwise*. La desviación residual **132.50** del modelo es menor que su desviación nula **419.03**, lo cual indica que el modelo con las variables incluidas explica una mayor variabilidad en los datos que el modelo sin ellas.

4. Conclusiones

El Thyroid Disease Data Set es una base de datos muy valiosa para la investigación médica y científica, ya que contiene información detallada sobre pacientes con enfermedades de la tiroides y características de sus condiciones de salud.

Un buen análisis descriptivo e inferencial de los datos puede ayudar a los profesionales médicos y científicos a identificar patrones y tendencias en los datos, lo que puede llevar a una mejor comprensión de las enfermedades de la tiroides y a la mejora de los tratamientos y diagnósticos. Además, un buen análisis de los datos puede ayudar a detectar posibles errores o inconsistencias en la base de datos, lo que puede ser crítico para la precisión y confiabilidad de los resultados.

En el presente estudio se analizó con éxito un conjunto de datos de pacientes para predecir la presencia de hipertiroidismo utilizando un modelo de regresión logística. Antes de la construcción del modelo, se describió la base de datos y se comprendieron a cabalidad las variables del set de datos *allhyper*. Posterior a esto, se realizó un meticuloso análisis exploratorio de los datos el cual incluyó la limpieza de los datos, el tratamiento de *outliers* y la identificación de correlaciones entre las distintas variables.

A partir del análisis y el uso de pruebas estadísticas, se logró identificar aquellas variables que parecían guardar una mayor importancia sobre la variable de clase. A continuación, se construyó un modelo con todas estas variables y se aplicó el método *Splitwise* de selección de variables para refinar aún más el modelo.

A partir del modelo final se demostró que las variables T3, T4U, TSH, TT4 y sexo de un paciente son significativas sobre la predicción del hipertiroidismo.

En resumen, el presente estudio nos ayuda a tener una mejor compresión de las variables que influyen en la presencia de hipertiroidismo en los pacientes y puede ser utilizado para predecir la probabilidad de hipertiroidismo en otro conjunto de datos. Además, es información valiosa para el desarrollo de las próximas experiencias de laboratorio que se asocian con el mismo conjunto de datos.

Bibliografía

CIGNA (2022). Pruebas de hormona tiroidea.

Claudia Godoy C., Marcela Acevedo M., A. B. N. (2009). Hipertiroidismo en niños y adolescentes.

Cooper, D., McDermott, M., and Wartofsky, L. (2006). Hipertiroidismo.

for Machine Learning, C. and at the University of California, I. S. (2007). Machine learning repository.

IBM (2021). Prueba de chi-cuadrado.

IBM (2022). Métodos de selección de variables en el análisis de regresión lineal.

James, P. (1972). Amyloid goitre. *Journal of Clinical Pathology*, 25(8):683–688.

MedlinePlus (2000). Hipertiroidismo provocado.

Pablo Livavic-Rojas, Guillermo Vallejo, P. F. (2006). Procedimientos estadÍsticos alternativos para evaluar la robustez mediante diseÑos de medidas repetidas.

Quinlan, R. (2007). Rulequest research pty ltd.

S., J. D. (2014). Regresión logística.

Sriphrapradang, C. and Bhasipol, A. (2016). Differentiating graves' disease from subacute thyroiditis using ratio of serum free triiodothyronine to free thyroxine. *Annals of medicine and surgery*, 10:69–72.

Universidad de Granada, E. (1900). Multicolinealidad.