

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



Laboratorio 1 - Análisis Estadístico

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

25 de Septiembre de 2022

Tabla de contenidos

1. Introducción	1
2. Descripción del problema	2
2.1. Descripción de la base de datos	5
2.2. Descripción de clases y variables	6
3. Análisis estadístico e inferencial	10
4. Conclusión	22
Bibliografía	24

1. Introducción

La glándula tiroides “tiene forma de mariposa y normalmente se localiza en la parte de adelante del cuello, su trabajo es formar las hormonas tiroideas, volcarlas al torrente sanguíneo y entregarlas a todos los tejidos del cuerpo”. (American Thyroid Association, sf) Es una parte importante del cuerpo humano ya que como sostiene Brent:

Las hormonas tiroideas son críticas en el desarrollo somático y cerebral en infantes y regulan la actividad metabólica en adultos, también afectan a la función de todos los órganos del cuerpo. Es por esto por lo que las hormonas tiroideas deben de estar constantemente disponibles para poder funcionar correctamente. (2022)

Debido a la importancia de la glándula tiroides y de sus hormonas para el normal funcionamiento del cuerpo se hace importante el entender y estudiar sobre las afecciones que puede presentar, siendo una de estas el hipotiroidismo, el cual es el problema que se trabajará a lo largo de este semestre. “El hipotiroidismo es la situación clínica producida por un déficit de hormonas tiroideas” (Barea et al., 2012, p. 819).

Para poder estudiar el problema se dispone de la base de datos *Allhypo*, la cual en general contiene una serie de exámenes tomados a distintas personas, en donde se detalla: edad, sexo, las medidas de la hormonas tiroideas como la hormona estimulante de la tiroides (TSH), triyodotironina (T3), tiroxina (T4) y la clasificación del individuo para hipotiroidismo. Tanto la base de datos como el problema serán abordados con mayor detalle en secciones posteriores del presente informe.

Los objetivos de esta experiencia de laboratorio son el adquirir un conocimiento profundo respecto al problema del hipotiroidismo y de la base de datos *Allhypo*. Respecto al problema del hipotiroidismo se busca entender sus causas, síntomas, problemas que genera, que indicadores influyen, etc. Respecto a la base de datos, se busca entender su estructura, variables y clases, para finalmente utilizar los datos disponibles en esta y realizar un estudio basado en todo lo aprendido respecto al problema, relacionando la teoría a los datos.

2. Descripción del problema

El objetivo en esta experiencia es estudiar el hipotiroidismo, para lograrlo se analizarán los datos almacenados en la base de datos *Allhypo* la cual contiene información recolectada en un estudio por el Instituto Garavan. Se desea conocer qué variables disponibles son relevantes para diagnosticar la enfermedad, para lograr esto se hace necesario el estudiar sobre el hipotiroidismo.

En primer lugar, se hace relevante el entender el funcionamiento de la tiroides, la organización Radiology Info aporta la siguiente información al respecto:

La glándula tiroides utiliza yodo (que se absorbe principalmente de la comida) para producir hormonas que controlan la forma en que su cuerpo utiliza la energía. La glándula pituitaria y el hipotálamo, un área que se encuentra en la base del cerebro, controlan la velocidad con la cual la tiroides produce y secreta estas hormonas. La función principal de la glándula tiroides es la secreción de una hormona denominada tiroxina o T4, que es convertida a una hormona denominada T3. Ambas hormonas circulan en el torrente sanguíneo y ayudan a regular su metabolismo. La cantidad de T4 producida por la glándula tiroides está determinada por una hormona producida por la glándula pituitaria denominada TSH, u hormona estimulante de la tiroides. Dependiendo de los niveles de T4 en el cuerpo, la glándula pituitaria produce más TSH o menos TSH para inducir a la tiroides a producir la cantidad adecuada de T4. (2020)

Respecto a los síntomas del hipotiroidismo Lozano sostiene que:

El espectro clínico es muy amplio y está condicionado por tres factores fundamentales: las dos grandes acciones generales de las hormonas tiroideas (consumo de oxígeno y efectos termogénicos), la intensidad del déficit hormonal y la edad del individuo en el momento de instaurarse el déficit. Así, todos los pacientes presentarán un grado variable de astenia y de letargia (disminución del consumo de oxígeno), de intolerancia al frío (reducción de la termogenia) y retraso en el crecimiento y desarrollo psicomotor en la infancia y adolescencia. Pero, además, todos los órganos, aparatos y sistemas orgánicos están afectados. (2006, p. 63)

Como se mencionó en la introducción el hipotiroidismo se puede describir como la condición provocada por déficit de hormonas tiroideas, sin embargo, es importante también conocer las causas por las cuales se puede adquirir la enfermedad, respecto a esto la organización Medline Plus menciona:

La causa más común de hipotiroidismo es la tiroiditis de Hashimoto. La hinchazón y la inflamación dañan las células de la glándula tiroides.

Otras causas de este problema incluyen:

- El sistema inmunitario ataca la glándula tiroides.
- Embarazo (a menudo llamado tiroiditis posparto).
- Determinados medicamentos, como litio o amiodarona y algunos tipos de quimioterapia.
- Anomalías congénitas (de nacimiento).
- Terapias de radiación al cuello o al cerebro para tratar diversos cánceres.
- Yodo radiactivo usado para tratar una tiroides hiperactiva.
- Extirpación quirúrgica de parte o de toda la glándula tiroidea.
- Tumor hipofisario o cirugía de la glándula pituitaria.

(2020)

Además de las causas por las cuales se produce el hipotiroidismo, es relevante destacar la existencia de tres distintas clasificaciones de la enfermedad: primario, secundario y subclínico. “El hipotiroidismo primario, la forma más frecuente de hipotiroidismo, no es sino una enfermedad autoinmune que aparece en general como secuela tras una tiroiditis de Hashimoto. Produce un tiroides hipotrófico y fibroso, con función escasa o nula.” (Lozano, 2006, p. 62), por otro lado “El hipotiroidismo secundario se desarrolla cuando el hipotálamo produce una cantidad insuficiente de hormona liberadora de tirotropina (TRH) o cuando la hipófisis produce una cantidad insuficiente de TSH.” (Manual MSD, 2020). Por último “El hipotiroidismo subclínico es el aumento de las concentraciones séricas de TSH en pacientes sin síntomas o con mínimos síntomas de hipotiroidismo y con concentraciones séricas normales de tiroxina (T4) libre.” (Manual MSD, 2020). El hipotiroidismo subclínico se hace relevante de observar,

debido a que “Diversos estudios han demostrado que el 30 % de los pacientes con hipotiroidismo subclínico desarrolla hipotiroidismo a 10 años y solo un 4 % de los pacientes con hipotiroidismo subclínico normalizan sus valores de TSH.” (Léniz, sf)

Una vez conocidas las distintas clasificaciones del hipotiroidismo queda la interrogante de conocer el cómo se diagnostica cada tipo de hipotiroidismo, según el libro de medicina interna de McMaster detalla a través de la Figura 1 que con niveles normales de TSH y Tiroxina T4 libre normales se tiene una tiroides con funcionamiento normal, o sea sin hipotiroidismo. Sin embargo, con niveles de TSH algo elevados y T4 libre normal se tiene hipotiroidismo subclínico, en el caso de niveles de TSH altos y T4 libre bajos se tiene hipotiroidismo primario, finalmente si los niveles de TSH son normales o bajos y además la t4 libre es baja se tiene hipotiroidismo secundario o también conocido como central.

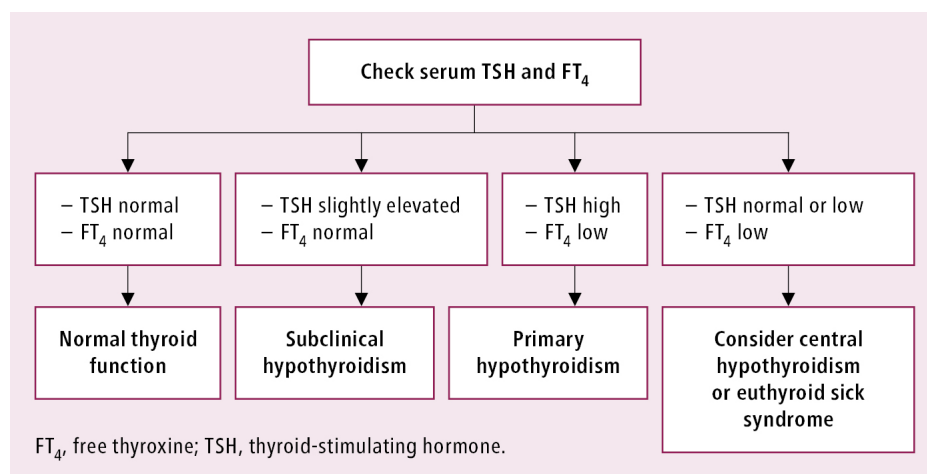


Figura 1: Algoritmo de diagnostico de hipotiroidismo basado en los niveles de TSH y T4 libre. *Adaptado de: McMaster Textbook of Internal Medicine (2021).* <https://empendium.com/mcmtextbook/chapter/B31.II.9.1>

El algoritmo de diagnostico del hipotiroidismo se basa en los niveles de las hormonas TSH y T4, es debido a esto que surge la interrogante, ¿qué es un nivel hormonal normal, alto o bajo? La clínica Cleveland Clinic detalla que:

Para el caso de la hormona estimulante de la tiroides (TSH) El rango normal para un adulto es: 0.40 - 4.50 mIU/mL (mili unidades internacionales por mililitro de sangre), para la tiroxina (T4) “Los rangos normales para un adulto es: 5.0 – 11.0 ug/dL (microgramos por decilitro de sangre) y en el caso de la triyodotironina

(T3) “El rango normal para un adulto es: 2.3 - 4.1 pg/mL (Picogramos por mililitro de sangre). (2019)

Con esta descripción inicial del problema es posible continuar a describir de mejor forma la base de datos.

2.1. Descripción de la base de datos

La base de datos entregada para la experiencia se encuentra contenida en un repositorio de aprendizaje de Machine learning de la Universidad de California en el interior del repositorio se encuentran diversos archivos con datos sobre distintas patologías, sin embargo los fundamentales para la realización de la experiencia son: *HELLO*, *allhypo.names* y *allhypo.data*.

En el archivo *HELLO*, se detalla una descripción general de las distintas bases de datos con registros sobre enfermedades a la tiroides aportados instituto de Garavan y por la Universidad de California durante la visita de Ross Quinlan en el taller de Machine Learning en 1987, además se indica la siguiente caracterización en las bases de datos:

- Muchos atributos (alrededor de 29).
- Atributos con valores numéricos o booleanos.
- Dominios de la enfermedad tiroidea.
- Varios valores de atributos perdidos (simbolizados por “?”).
- Poco número de clases (menos de 10).
- 2800 instancias en cada set de datos.
- 972 instancias en cada set de prueba.

(Dua and Graff, 2017)

En el archivo *allhypo.names*, se detalla el nombre y posibles valores para las 29 variables utilizadas, cabe destacar que el orden en el que se presentan las variables coincide con el archivo *allhypo.data*, además se indican 4 posibles clases (tipos de hipotiroidismo). Finalmente en el archivo *allhypo.data*, contiene 2800 instancias donde cada línea representa

los datos asociados a un paciente, en cada instancia los valores de las distintas variables se encuentran separados por comas, además de su clase y un identificador.

2.2. Descripción de clases y variables

Como se mencionó en la sección anterior, la base de datos contiene 4 clases distintas, las cuales corresponden a los distintos tipos de hipotiroidismo, estas clases son:

- **negative:** Significa que la persona estudiada no tiene hipotiroidismo.
- **primary hypothyroid:** Significa que la persona estudiada tiene hipotiroidismo primario, como se mencionó en la sección de descripción del problema, el hipotiroidismo primario es el tipo de hipotiroidismo más común y se caracteriza por una tiroides con una función escasa o directamente nula. Es diagnosticada cuando los niveles de la hormona TSH son altos, pero los de la hormona T4 son bajos.
- **secondary hypothyroid:** Significa que la persona estudiada tiene hipotiroidismo secundario, como se mencionó en la sección de descripción del problema, en el hipotiroidismo secundario, el fallo viene por parte del hipotálamo o la hipófisis, produciendo una baja cantidad de las hormonas TRH o TSH las cuales estimulan a la tiroides a secretar hormonas tiroideas. Es diagnosticada cuando los niveles de la hormona TSH son normales o bajos y los niveles de la hormona T4 son bajos.
- **compensated hypothyroid:** Significa que la persona tiene hipotiroidismo subclínico “El hipotiroidismo subclínico (también conocido como hipotiroidismo compensado o hipotiroidismo leve).” (Redford, 2017). Como sostiene la asociación Children’s Oncology Group:

Ocurre cuando la glándula pituitaria tiene que sobre trabajar la glándula tiroidea para mantener el nivel normal de hormonas tiroideas en la sangre. Esto puede ser un problema temporal luego de estar expuesto a radiación, o puede ser un signo que la glándula tiroidea está empezando a fallar. (sf)

Es por esto y como se mencionó en la sección de descripción del problema que este tipo de hipotiroidismo se diagnostica cuando los niveles de hormona TSH son elevados, pero los niveles de T4 son normales.

Una vez descritas las clases que presenta la base de datos, a continuación se realizará una breve descripción de cada una de las variables contenidas en la base de datos.

1. **Age:** Indica la edad del paciente.
2. **Sex:** Indica el sexo del paciente, toma los valores M, F correspondientes a male (hombre) y female (mujer).
3. **On thyroxine:** Indica si el paciente está en tratamiento de hipotiroidismo con tiroxina.
4. **Query on thyroxine:** Indica si el paciente ha estado previamente en tratamiento de hipotiroidismo con tiroxina.
5. **On antithyroid medication:** Indica si el paciente está en tratamiento para hipertiroidismo.
6. **Sick:** Indica si el paciente se encuentra enfermo.
7. **Pregnant:** Indica si la paciente se encuentra embarazada.
8. **I131 treatment:** Indica si el paciente ha sido sometido a tratamiento con yodo radioactivo. Respecto a este tratamiento Mumtaz et al. sostienen que:

El yodo radiactivo (I-131) o RAI, como se denominará ahora, se ha utilizado comúnmente para el tratamiento de afecciones tiroideas tanto benignas como malignas desde la década de 1940. El objetivo de la terapia es tratar el hipertiroidismo destruyendo suficiente tejido tiroideo para que el paciente sea eutiroides o hipotiroideo. (2009)

9. **Query hypothyroid:** Pregunta al paciente si cree tener hipotiroidismo.
10. **Query hyperthyroid:** Pregunta al paciente si cree tener hipertiroidismo.
11. **Lithium:** Indica si el paciente se encuentra en tratamiento con litio, respecto a esto “El litio es ampliamente utilizado como tratamiento y resulta eficaz para los trastornos bipolares. Su uso está asociado con anormalidades tiroideas” (Zantour and Chebbi, 2014, p. 1). “Las acciones antitiroideas del litio se investigaron por primera vez en detalle cuando se observó que los pacientes con enfermedades psiquiátricas tratados con carbonato de litio desarrollaban hipotiroidismo y bocio.” (Surks, 2021)

12. **Goitre:** Indica si el paciente tiene bocio. “El término “bocio” simplemente se refiere a un agrandamiento anormal de la glándula tiroides.” (American Thyroid Association, 2017)
13. **Tumor:** Indica si el paciente tiene un tumor.
14. **Hypopituitary:** Indica si el paciente posee hipopituitarismo “El hipopituitarismo es una afección en la que la hipófisis (una pequeña glándula situada en la base del cerebro) no produce una o más de sus hormonas, o no lo hace en cantidad suficiente.” (Ratini, 2020). Cabe recordar que en la sección de descripción del problema se mencionó que es esta glándula la encargada de secretar TSH.
15. **Psych:** Indica si el paciente posee algún trastorno psiquiátrico.
16. **TSH measured:** Indica si se ha medido el nivel de la hormona TSH al paciente.
17. **TSH:** Indica los valores de hormona TSH del paciente, como se mencionó en la sección de descripción del problema, los valores normales de esta hormona para un adulto son: 0.40 - 4.50 mIU/mL (mili unidades internacionales por mililitro de sangre).
18. **T3 measured:** Indica si se ha medido el nivel de la hormona T3 al paciente.
19. **T3:** Indica los valores de la hormona T3 del paciente, como se mencionó en la sección de descripción del problema, los valores normales de esta hormona para un adulto son: 2.3 - 4.1 pg/mL (Picogramos por mililitro de sangre).
20. **TT4 measured:** Indica si se ha medido el nivel de la hormona T4 Total al paciente.
21. **TT4:** Indica el valor de la hormona T4 total del paciente, como se mencionó en la sección de descripción del problema, los valores normales de esta hormona para un adulto es: 5.0 – 11.0 ug/dL (microgramos por decilitro de sangre).
22. **T4U measured:** Indica si se ha medido el radio de utilización de tiroxina (T4U)
23. **T4U:** Indica el valor del radio de utilización de la hormona T4.
24. **FTI measured:** Indica si se ha medido el nivel de hormona T4 libre en el paciente. También es conocida como FT4

25. **FTI:** Indica los valores de la hormona T4 libre del paciente, según Haldeman-Englert:

Los límites normales para la T4 libre en adultos es de 0.8 a 1.8 nanogramos por decilitro (ng/dl). Un nivel de T4 libre más alto de lo normal podría significar que tiene una tiroides hiperactiva. Las afecciones asociadas con el hipertiroidismo incluyen la enfermedad de Graves, un trastorno autoinmunitario. (2020)

26. **TBG measured:** Indica si fue medida la proteína globulina fijadora de tiroxina (TBG) en el paciente, Krawiec et al. describen y mencionan la función de esta proteína:

La globulina transportadora (TBG), componente menor de las α -globulinas, es una glicoproteína de 54 kDa, monomérica, que transporta el 70 % de T4 y T3 con una diferencia de afinidad de 10 a 20 veces a favor de la primera. Posee un solo sitio para unir ambas hormonas que es ocupado hasta el 25 %. (sf, p. 6)

27. **TBG:** Indica el valor de la proteína TBG medido en el paciente, “El rango normal es 13 a 39 microgramos por decilitro (mcg/dL) o 150 a 360 nanomoles por litro (nmol/L).” (Medline Plus, 2022)

28. **Referral source:** Corresponde a la fuente de referencia, puede tomar los valores de: SVI, SVHD, SVHC, STMW y other.

Tomar en consideración que la gran mayoría de variables son booleanas, es decir, toman uno de dos valores dicotómicos. De las variables previamente descritas Age, TT4, T4U, FTI y TBG son numéricas, mientras el resto de las variables son booleanas.

Cabe destacar que la descripción de ciertas variables como Tumor y sick no están directamente denotadas en la base de datos o en la literatura, por lo cual sus descripciones se realizaron en base al contexto del hipotiroidismo y su nombre.

3. Análisis estadístico e inferencial

Previo a realizar los análisis estadísticos, fue necesario realizar una limpieza a los datos del archivo *allhypo.data*, ya que como se detalla en el archivo *HELLO* disponible en la base de datos “Los archivos de documentación (con la extensión “names”) tienen un formato para ser leído por el programa de árbol de decisión de Quinlan” (Dua and Graff, 2017) y no solo el archivo con extensión names tiene ese formato, el archivo *allhypo.data* también lo posee. Luego los datos fueron filtrados, solo dejando las observaciones de los pacientes que tienen todos sus datos completos, eliminando cualquier observación con alguna variable incompleta, representada por el símbolo ? en la base de datos. Además las variable *TBG measured* y *TBG* no son consideradas debido a que no fueron medidas para ningún paciente, por lo que no aportan información. Al aplicar estos filtros las observaciones se redujeron de 2800 a 1946, dividiéndose en 1789 pacientes negativos, 50 pacientes con hipotiroidismo primario, 1 paciente con hipotiroidismo secundario y 106 pacientes con hipotiroidismo subclínico. Debido a que se cuenta con un único paciente con hipotiroidismo secundario es que se descarta esta clase para el estudio estadístico al no aportar información.

En la figura 2 se puede observar una gráfica de puntos de la edad de cada paciente dependiendo de su clasificación. A simple vista pareciera no haber ninguna tendencia o diferencia entre las edades de los pacientes de los tres grupos.

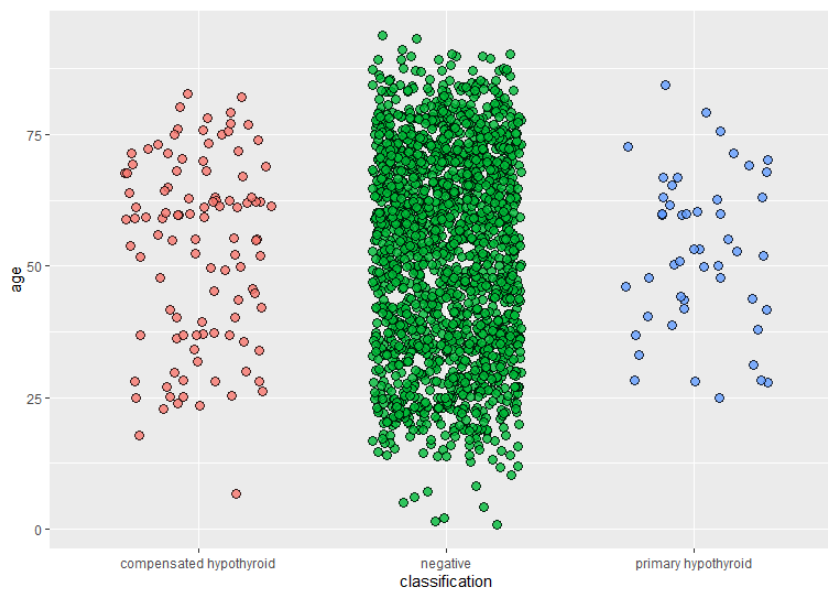


Figura 2: Gráfico de puntos: edad por clasificación de hipotiroidismo

En la figura 3 se puede observar una gráfica de puntos del radio de utilización de tiroxina (T4U) de cada paciente según su clasificación, al igual que el caso anterior pareciese no haber una diferencia entre los valores de los grupos.

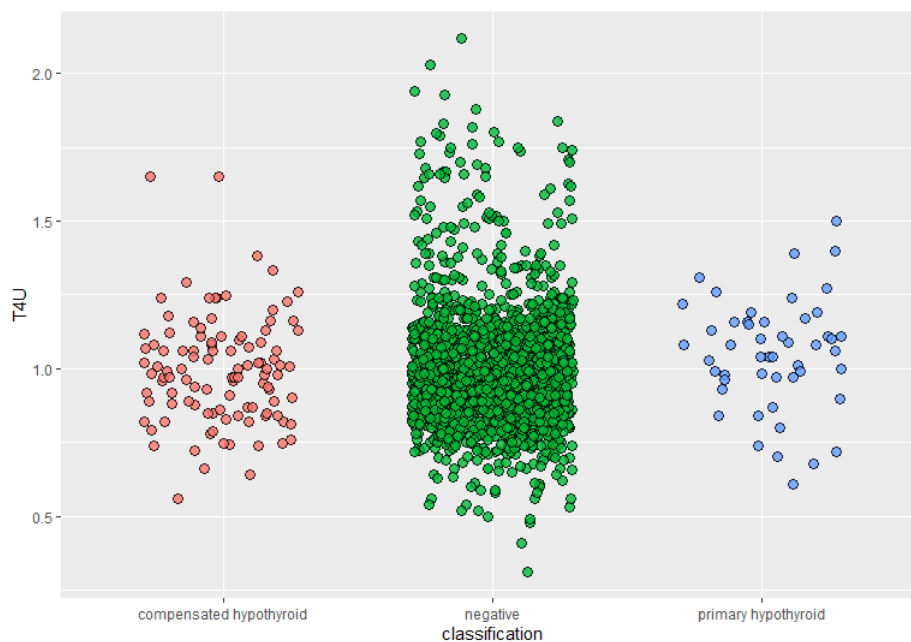


Figura 3: Gráfico de puntos: radio de utilización de tiroxina por clasificación de hipotiroidismo

En la figura 4 es posible observar una gráfica de puntos de la medición de hormona T3 medida en pg/mL de cada paciente según su clasificación, a simple vista pareciese no haber diferencia entre los valores de cada grupo.

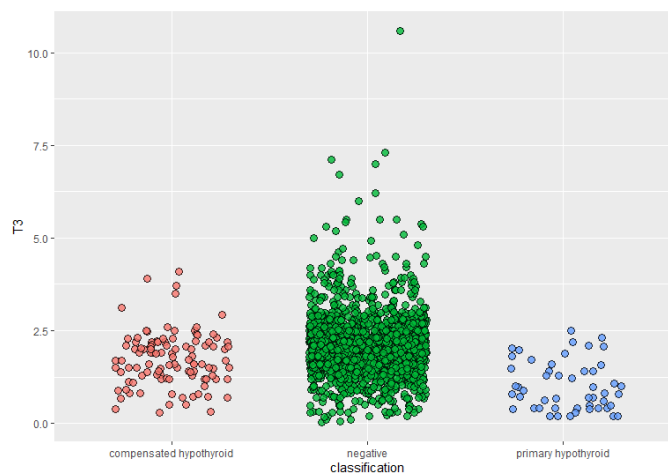


Figura 4: Gráfico de puntos: hormona T3 por clasificación de hipotiroidismo

En la figura 5 es posible observar una gráfica de puntos de la medición de hormona T4 total medida en ug/dL de cada paciente según su clasificación, en este caso es posible observar a simple vista como los valores para el grupo de pacientes con hipotiroidismo primario pareciesen ser menores en promedio que el grupo de los pacientes negativos y con hipotiroidismo subclínico.

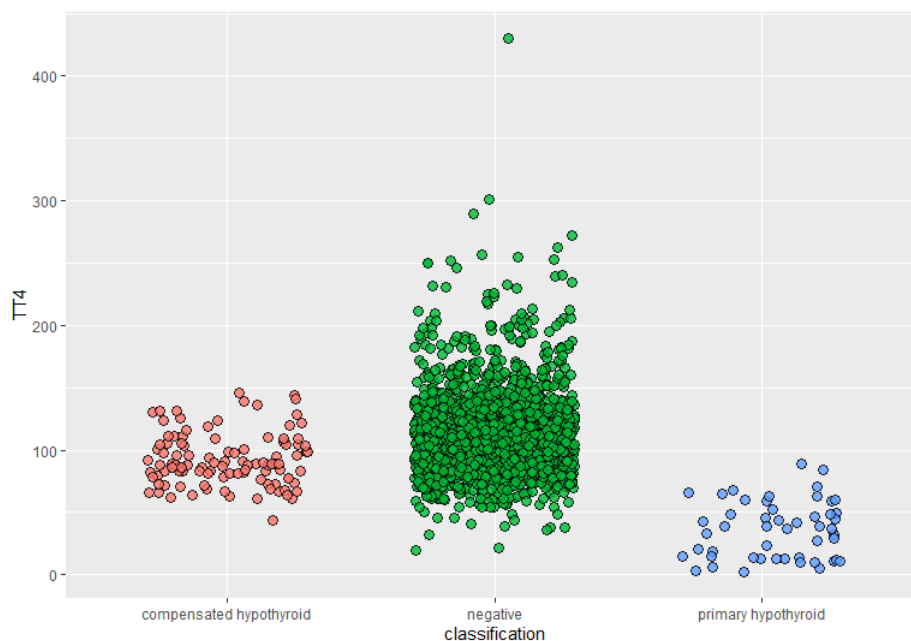


Figura 5: Gráfico de puntos: hormona T4 total por clasificación de hipotiroidismo

En la figura 6 se observa una gráfica de puntos de la medición de la hormona TSH en mIU/mL de cada paciente según su clasificación, en este caso también es posible observar a simple vista como los niveles promedio de esta hormona pareciesen ser mayores en el caso del grupo de hipotiroidismo primario en comparación al grupo de hipotiroidismo subclínico y negativos.

En la figura 7 se observa una gráfica de puntos de la medición de la hormona T4 libre en ng/dl de cada paciente según su clasificación, es posible observar a simple vista como los niveles promedio de esta hormona pareciesen ser menores en el caso del grupo de hipotiroidismo primario en comparación al grupo de hipotiroidismo subclínico y negativos.

Tomando en consideración las variables expuestas en los gráficos de puntos y con el fin de comprobar las tendencias aparentes en las distintas variables observadas a simple vista es que se realizaron distintos test de hipótesis. Al tener 3 clases a comparar y al comprobar la

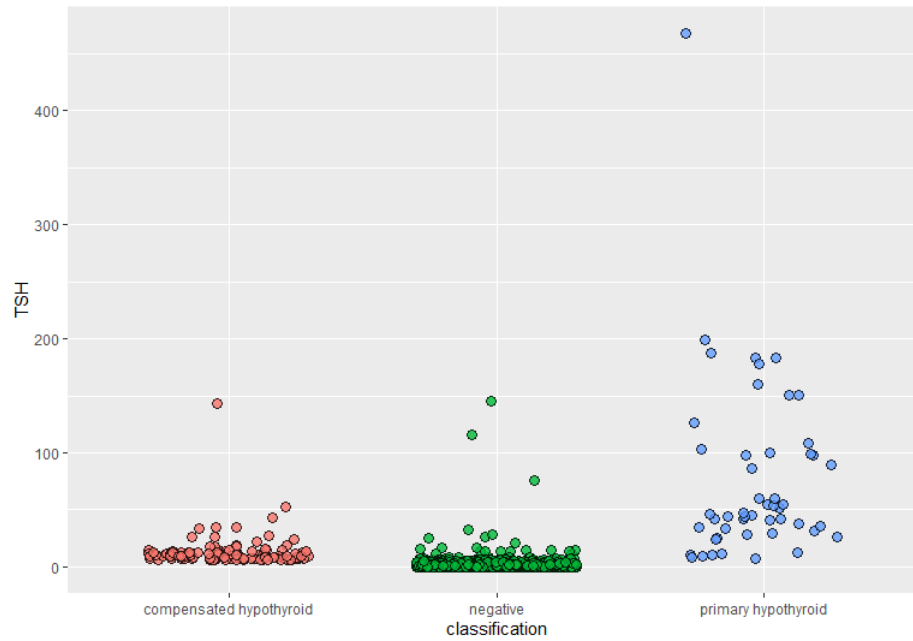


Figura 6: Gráfico de puntos: hormona TSH por clasificación de hipotiroidismo

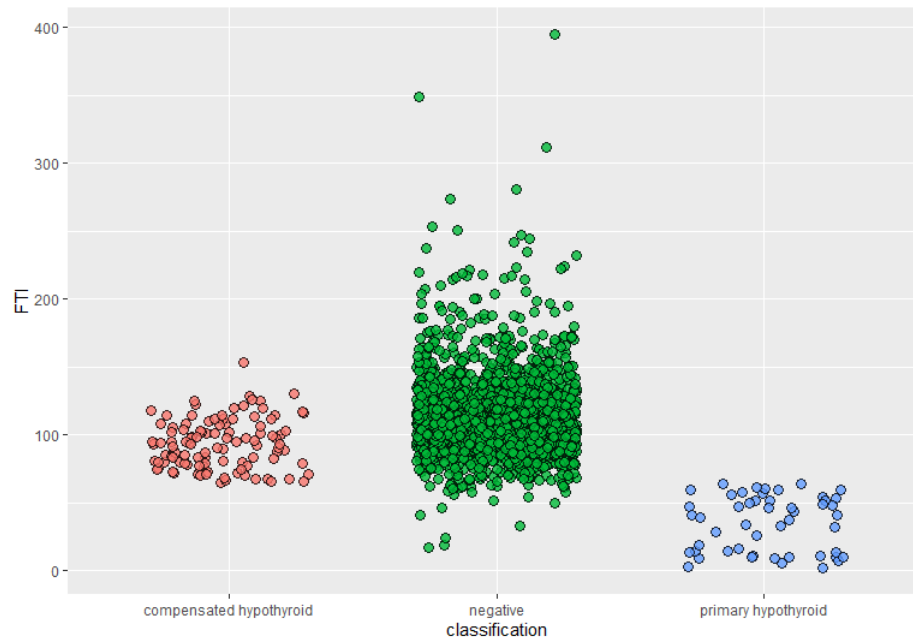


Figura 7: Gráfico de puntos: niveles de hormona T4 libre por clasificación de hipotiroidismo

normalidad de los datos, los métodos robustos de ANOVA para medias truncadas se presentan como una buena opción para poner a prueba las hipótesis al cumplir las condiciones para su aplicación.

	p.value
Compensated vs. Negative	0
Compensated vs. Primary	0
Negative vs. Primary	0

Tabla 1: Procedimiento Post-hoc para TSH.

Para la variable edad (**age**) no se aprecia una diferencia aparente en la media entre las distintas clases (figura 2), por lo cual la **prueba de hipótesis** realizada consiste en comprobar si **existe diferencia en la edad promedio de los pacientes con hipotiroidismo primario, subclínico y negativo**. El resultado de esta prueba de hipótesis arroja un valor ($t1way \mid p = 0.9127$), de modo que se puede concluir con 95 % de confianza que en efecto y como se observó a simple vista **no existe diferencia en la edad promedio de las distintas clases**.

Para la variable **TSH** se aprecia una diferencia aparente en la media de los pacientes con **hipotiroidismo primario** (figura 6) siendo esta **mayor que la de los pacientes con hipotiroidismo subclínico y negativos**, por lo cual la **prueba de hipótesis** realizada consiste en verificar si **existe diferencia en los niveles TSH promedio en los pacientes con hipotiroidismo primario, subclínico y negativo**. El resultado de esta prueba de hipótesis arroja un valor ($t1way \mid p = 0$), por lo cual es necesario realizar una **prueba post-hoc** para identificar entre que grupos existe diferencia.

Mediante los resultados de la prueba post-hoc observada en la tabla 1 se puede concluir que **existe diferencia en las medias de TSH entre los pacientes con hipotiroidismo primario, subclínico y negativo**.

Dado que se verificó que existen diferencias entre las medias de los distintos grupos de pacientes resulta interesante verificar aquellos grupos con mayores niveles de TSH, es por esto que a continuación se realizan pruebas de hipótesis entre únicamente dos grupos.

Como se mencionó anteriormente en el gráfico 6 los pacientes con hipotiroidismo primario pareciesen tener una media mayor que los grupos de hipotiroidismo subclínico y negativos, es por esto que en primer lugar se realizara una **prueba de hipótesis** para verificar si **en promedio los niveles de TSH son mayores en pacientes con hipotiroidismo primario con respecto a pacientes con hipotiroidismo subclínico**. El resultado de

esta prueba de hipótesis arroja un valor (Yuen T test | $p = 4.837e-07$), por ende se puede concluir que **en promedio los niveles de TSH de los pacientes con hipotiroidismo primario son mayores a los pacientes con hipotiroidismo subclínico.**

De manera análoga, la siguiente prueba de hipótesis a realizar consiste en verificar si **los niveles TSH, en promedio son mayores en pacientes con hipotiroidismo primario que en pacientes negativos.** El resultado de esta prueba de hipótesis arroja un valor (Yuen T test | $p = 2.397e-08$), por ende se puede concluir que **en promedio los niveles de TSH de los pacientes con hipotiroidismo primario son mayores a los pacientes negativos.**

Quedaría únicamente comparar las medias de TSH entre los pacientes con hipotiroidismo subclínico y negativos, por ende la **prueba de hipótesis** restante a contrastar consiste en verificar si **los niveles promedio de TSH en pacientes de hipotiroidismo subclínico son mayores a los pacientes negativos.** El resultado de esta prueba arroja un valor (Yuen T test | $p < 2.2e-16$), por lo cual se puede concluir que **los niveles de TSH promedio son mayores para los pacientes con hipotiroidismo subclínico con respecto a los pacientes negativos.**

Los resultados de estas últimas pruebas son totalmente concordantes a lo expuesto en la sección de descripción del problema, como se explicó y se puede ver resumido en la figura 1 las personas con hipotiroidismo subclínico se caracterizan por tener los niveles de hormona TSH ligeramente elevados en comparación a quienes poseen una tiroides con normal funcionamiento y a su vez las personas que sufren de hipotiroidismo primario se caracterizan por tener un alto nivel de hormona TSH, más alto que quienes sufren de hipotiroidismo subclínico. Lo cual son los mismos resultados conseguidos a nivel estadístico mediante las pruebas de hipótesis aplicadas a los datos.

Con respecto a la hormona T4 libre (**FTI**), en el gráfico 7 se aprecia un nivel menor en los pacientes con hipotiroidismo primario, por lo cual en primera instancia la **prueba de hipótesis** a realizar consiste en comprobar si **existe diferencia en los niveles de FTI promedio entre los pacientes con hipotiroidismo primario, subclínico y negativo.** El resultado de esta prueba arroja un valor (t1way | $p = 0$), por ende es necesario aplicar una prueba post-hoc para identificar en que grupos existe diferencia. En base a los resultados de la prueba post-hoc en la tabla 2, se puede concluir que **existe diferencia en las medias**

	p.value
Compensated vs. Negative	0
Compensated vs. Primary	0
Negative vs. Primary	0

Tabla 2: Procedimiento Post-hoc para FTI.

de FTI entre los pacientes con hipotiroidismo primario, subclínico y negativo.

Dado que se verificó que los niveles de FTI son distintos en los pacientes, restaría comprobar si los niveles de FTI promedio son menores en pacientes con hipotiroidismo primario, por tanto a continuación se realizan pruebas de hipótesis entre dos grupos.

En primer lugar se comparan las medias entre los pacientes con hipotiroidismo primario y subclínico, la **prueba de hipótesis** a realizar consiste en verificar **si en promedio los niveles de FTI son menores en pacientes con hipotiroidismo primario con respecto a los pacientes con hipotiroidismo subclínico**. El **resultado** de esta prueba arroja un valor (Yuen T test | $p = 7.513e-16$), por lo que se puede concluir que **en promedio los niveles de FTI son menores en pacientes con hipotiroidismo primario con respecto a los pacientes con hipotiroidismo subclínico**.

Análogamente, se realiza la misma prueba de hipótesis entre los pacientes con hipotiroidismo primario y negativos. Así mismo, el **resultado** de esta prueba arroja un valor de (Yuen T test | $p < 2.2e-16$), por ende se puede concluir que **en promedio los niveles de FTI son menores en pacientes con hipotiroidismo primario con respecto a los pacientes negativos**.

Finalmente queda comprobar que sucede con las medias entre los pacientes con hipotiroidismo subclínico y negativos. Dado que en el gráfico 7 los niveles de FTI de los pacientes con hipotiroidismo subclínico parecen ser menores, la **prueba de hipótesis** consiste en comprobar si **los niveles de FTI promedio en pacientes con hipotiroidismo subclínico es menor que en pacientes negativos**.

El resultado de esta prueba arroja un valor de (Yuen T test | $p = 1.298e-10$), por ende **se puede concluir que los pacientes con hipotiroidismo subclínico tienen en un nivel de FTI en promedio menor a los pacientes negativos**.

Una vez finalizadas las pruebas de hipótesis y con el fin de poder descubrir más

información sobre el problema, se construyó una matriz de correlación usando VdeCramer para todas las variables que aportan información, es posible visualizarla en la figura 8. De esta matriz es posible observar que existe una alta correlación de la variable embarazo con los niveles de T4U, lo que puede indicar que los niveles de radio de utilización de la hormona T4 (T4U) aumentan durante el embarazo, para confirmar esto se ha hecho una prueba de hipótesis, la cual consiste en comparar la media entre las mujeres embarazadas y las mujeres no embarazadas, como hipótesis alternativa se comprueba si la media del radio de utilización de la hormona T4 es mayor en las mujeres embarazadas con respecto a las mujeres no embarazadas, los resultados de la prueba arrojan (Yuen T test | $p = 1.272e-11$), por ende se puede concluir que efectivamente el promedio del radio de utilización de la hormona T4 es mayor en las mujeres embarazadas. Respecto al embarazo, Medici et al. sostiene:

Durante el embarazo, se producen profundos cambios en la fisiología tiroidea, lo que da lugar a intervalos de referencia de la hormona estimulante del tiroides (TSH) y de la tiroxina libre (FT 4) diferentes en comparación con el estado de no embarazo. (2015, p. 704)

Sin embargo, lo más interesante es la alta correlación que tiene la variable de clasificación a las variables de las hormonas TSH, TT4 y FTI. Lo que podría apuntar a que estas tres hormonas tienen mucha relevancia al momento de determinar la clasificación de hipotiroidismo para un paciente.

Para confirmar e indagar aún más a nivel estadístico cuales son las variables más relevantes para determinar la clasificación de hipotiroidismo para un paciente se decidió el entrenar tres modelos de regresión logística. El primer modelo busca predecir entre las clases negativo y hipotiroidismo primario, en la figura 9 es posible observar la importancia calculada para cada variable en la predicción de la clasificación del paciente. Como se puede apreciar las variables más importantes son los niveles de las hormonas TSH, FTI y TT4, siendo concordante a los resultados conseguidos en la matriz de correlación. El modelo fue entrenado y evaluado usando validación cruzada repetida de 10 pliegues con 3 repeticiones, usando la curva ROC como parámetro a optimizar. Los predictores escogidos automáticamente por este modelo fueron dos: TSH y FTI, lo que concuerda con las dos variables con más nivel de correlación de la matriz, además Jamal et al. mencionan que:

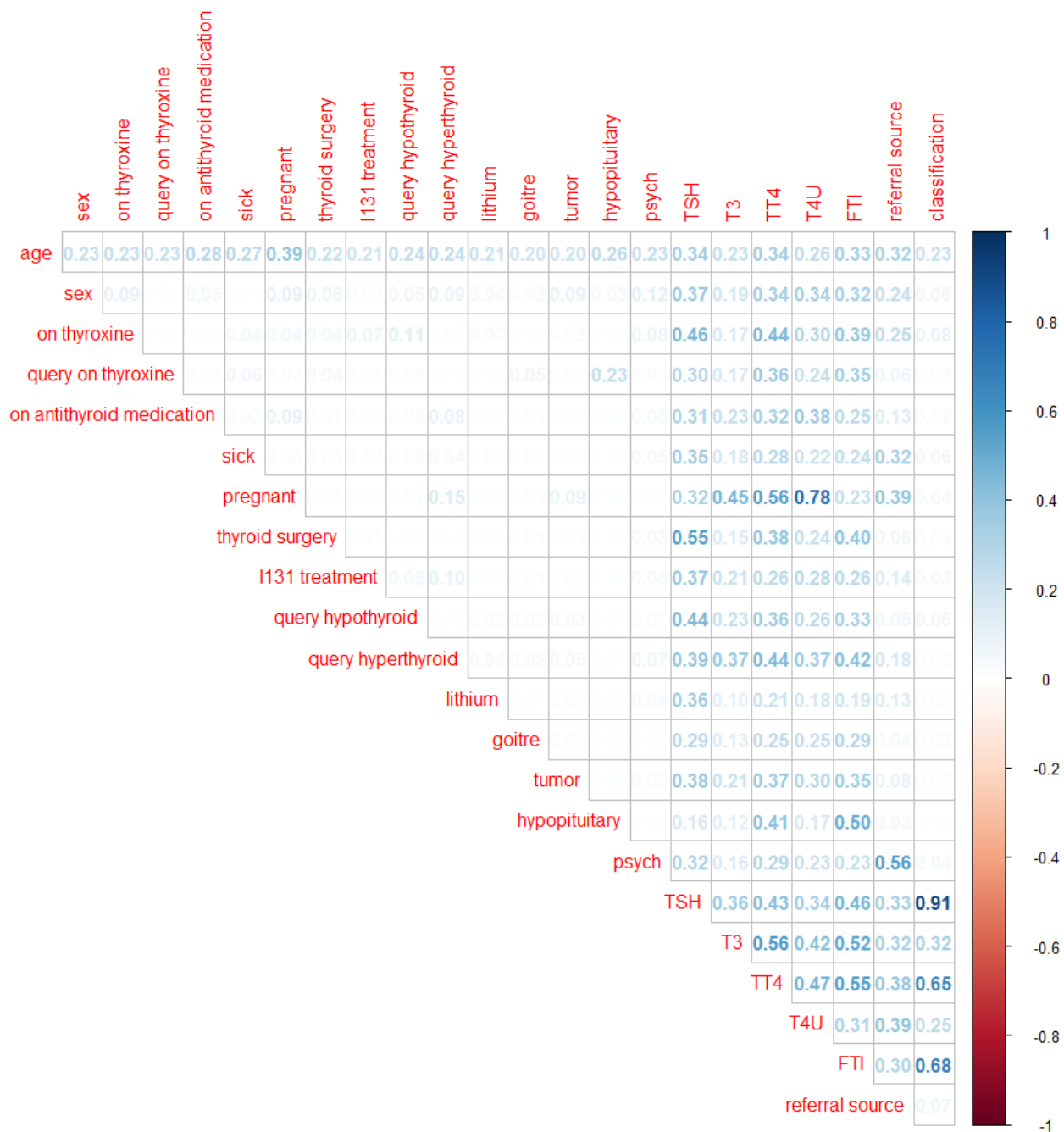


Figura 8: Matriz de correlación con VdeCramer para todas las variables

El nivel de T4 libre es actualmente la mejor medida del estado de la tiroides en conjunto a los niveles de TSH en la sangre. El nivel de T4 libre se mantiene bastante estable en los individuos a lo largo de los años en un “punto de ajuste” óptimo para la evaluación de la función tiroidea. (2021)

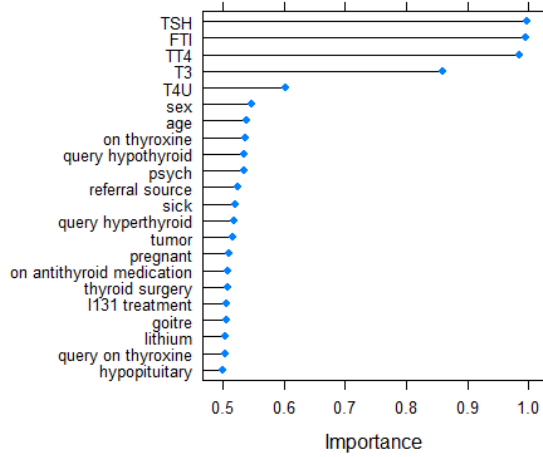


Figura 9: Importancia de cada variable: modelo negativo vs hipertiroidismo primario

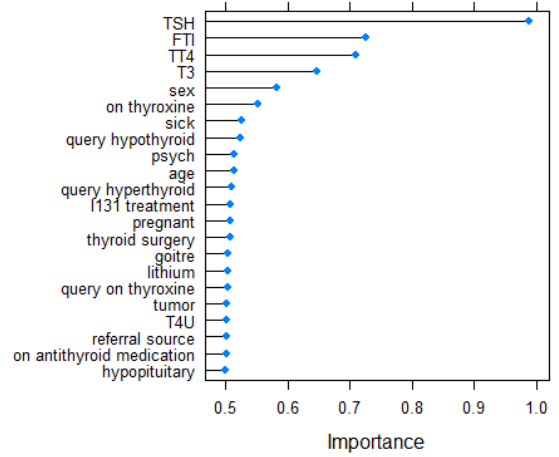


Figura 10: Importancia de cada variable: modelo negativo vs hipertiroidismo subclínico

Esto da a entender que el modelo está escogiendo los dos predictores que son reconocidos en la literatura como los más idóneos para evaluar el estado de la tiroides y si recordamos el algoritmo de diagnóstico de hipotiroidismo de la figura 1, los predictores escogidos por el modelo son exactamente los mismos usados en la medicina para el diagnóstico de los distintos tipos de hipotiroidismo.

El modelo consiguió una precisión de 0.99 en el conjunto de datos de prueba, tiene altos valores de sensibilidad, especificidad y del coeficiente Kappa, lo que da a entender que es un buen modelo, el resumen de esto se puede apreciar en la figura 11.

El segundo modelo de regresión logística fue entrenado para predecir sobre las clases hipotiroidismo subclínico y negativo, ocupando los mismos parámetros de entrenamiento que el modelo anterior, en la figura 10 se puede observar la importancia calculada por el modelo para cada variable en la predicción de la clasificación del paciente, es interesante observar que al igual que para el modelo anterior, las variables con más importancia fueron TSH, FTI y TT4. Además los predictores escogidos por el modelo fueron también TSH y FTI, concordando con el modelo anterior. El resumen de las estadísticas de este modelo en los datos de prueba se puede observar en la figura 12.

El ultimo modelo de regresión implementado es un modelo de regresión logística multinomial, el cual busca predecir al mismo tiempo sobre las tres clases de hipotiroidismo.

Confusion Matrix and Statistics		
Prediction	Reference	
	negative	primary hypothyroid
negative	533	1
primary hypothyroid	3	14
Accuracy : 0.9927		
95% CI : (0.9815, 0.998)		
No Information Rate : 0.9728		
P-Value [Acc > NIR] : 0.0007635		
Kappa : 0.8713		
McNemar's Test P-Value : 0.6170751		
Sensitivity : 0.9944		
Specificity : 0.9333		
Pos Pred Value : 0.9981		
Neg Pred Value : 0.8235		
Prevalence : 0.9728		
Detection Rate : 0.9673		
Detection Prevalence : 0.9691		
Balanced Accuracy : 0.9639		
'Positive' Class : negative		

Figura 11: Matriz de confusión y estadísticas
modelo: negativo vs hipotiroidismo primario

Confusion Matrix and Statistics			
Prediction	Reference		
	compensated hypothyroid	negative	primary hypothyroid
compensated hypothyroid	12	4	
negative	19	532	
Accuracy : 0.9594			
95% CI : (0.9398, 0.9741)			
No Information Rate : 0.9453			
P-Value [Acc > NIR] : 0.078482			
Kappa : 0.4917			
McNemar's Test P-Value : 0.003509			
Sensitivity : 0.38710			
Specificity : 0.99254			
Pos Pred Value : 0.75000			
Neg Pred Value : 0.96552			
Prevalence : 0.05467			
Detection Rate : 0.02116			
Detection Prevalence : 0.02822			
Balanced Accuracy : 0.68982			
'Positive' Class : compensated hypothyroid			

Figura 12: Matriz de confusión y estadísticas
modelo: negativo vs hipotiroidismo subclínico

La importancia de las distintas variables evaluadas por el modelo se presentan en la figura 13. Es posible observar como al contrario de los otros modelos, la variable más relevante para el modelo es la variable booleana *on thyroxine*, seguida de TSH, FTI y thyroid surgery. Esto resulta interesante, debido a que deja entrever que no solo los niveles hormonales son importantes para poder clasificar el tipo de hipotiroidismo que tiene el paciente.

Resulta interesante también el observar que este modelo, eligió múltiples predictores, al contrario de los otros que escogieron solo dos, los predictores seleccionados por el modelo se pueden observar en la figura 15, lo que refuerza la idea anterior de que no solo los niveles hormonales son relevantes para la clasificación.

Finalmente los resultados del modelo se pueden observar en la figura 14, se puede observar una precisión de 0.97, buenos niveles de especificidad y sensibilidad, lo que indica que es un buen modelo.

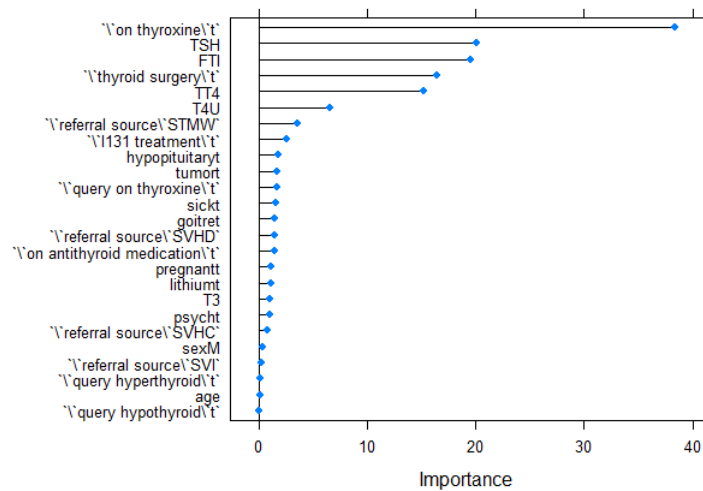


Figura 13: Importancia de cada variable: modelo regresión logistica multinomial

Confusion Matrix and Statistics				
Prediction	Reference			
	compensated hypothyroid	negative	primary hypothyroid	
compensated hypothyroid	23	2	2	
negative	8	532	2	
primary hypothyroid	0	2	11	
Overall Statistics				
Accuracy : 0.9725				
95% CI : (0.9557, 0.9842)				
No Information Rate : 0.921				
P-Value [Acc > NIR] : 1.322e-07				
Kappa : 0.8826				
McNemar's Test P-Value : 0.1328				
Statistics by Class:				
	Class: compensated hypothyroid	Class: negative	Class: primary hypothyroid	
Sensitivity	0.74194	0.9925	0.73333	
Specificity	0.99274	0.7826	0.99647	
Pos Pred Value	0.85185	0.9815	0.84615	
Neg Pred Value	0.98559	0.9800	0.99297	
Prevalence	0.05326	0.9210	0.02577	
Detection Rate	0.03952	0.9141	0.01890	
Detection Prevalence	0.04639	0.9313	0.02234	
Balanced Accuracy	0.86734	0.8876	0.86490	

Figura 14: Resultados modelo de regresión logística multinomial

"age"	"sexM"	"on thyroxine"
"query on thyroxine"	"on antithyroid medication"	"sickt"
"pregnant"	"thyroid surgery"	"I131 treatment"
"query hypothyroid"	"query hyperthyroid"	"lithium"
"goitret"	"tumort"	"hypopituitary"
"psycht"	"TSH"	"T3"
"TT4"	"T4U"	"FTI"
"referral source STMW"	"referral source SVHC"	"referral source SVHD"
"referral source SVI"		

Figura 15: Predictores seleccionados por el modelo de regresión logística multinomial

4. Conclusión

En la presente experiencia de laboratorio se realizó el estudio de la enfermedad hipotiroidismo la cual afecta a la glándula tiroides, además se realizó un análisis estadístico ocupando una base de datos que contenía distinta información de variables relacionadas a esta enfermedad. Todo esto con el objetivo interiorizarse y entender a fondo el problema y la base de datos que se ocuparan a lo largo de este semestre en la asignatura, al no ser expertos en el tema fue necesario el investigar a fondo respecto a la enfermedad y las distintas variables que contiene la base de datos, esto se logró mediante la lectura de diversos estudios, revistas médicas, libros de medicina, artículos relevantes, entre otros.

Respecto a resultados relevantes que se encontraron durante el estudio, se puede mencionar que en las gráficas de puntos revelaban a simple vista ciertas tendencias, como por ejemplo que la hormona TSH era mayor en los pacientes con hipotiroidismo primario en comparación a los otros grupos, además en pacientes con hipotiroidismo subclínico también se aprecian niveles mayores que en pacientes negativos. Esto fue confirmado a través de una prueba de hipótesis y mediante la literatura, la cual indicaba este mismo comportamiento.

Además de la hormona TSH, los comportamientos de las hormonas TT4 y FTI también tenían tendencias marcadas a simple vista, las cuales fueron confirmadas mediante pruebas de hipótesis y también como el caso anterior, concordaban los resultados con lo que indicaba la literatura.

Para poder descubrir posibles fenómenos en los datos se construyó una matriz de correlación, de la cual se descubrió que las mujeres embarazadas tienen un mayor radio de utilización de tiroxina (T4U) que las mujeres que no están embarazadas, esto se confirmó mediante un test de hipótesis y además la literatura también reforzó este fenómeno.

De la matriz de correlación también llamó la atención los altos índices de correlación de la variable de clasificación de la enfermedad a las variables de las hormonas TSH y FTI, para indagar más sobre esta idea se construyeron tres modelos de regresión logística. Estos modelos evidenciaron fuertemente la importancia de estas dos hormonas para la clasificación de la enfermedad, esto era esperable, debido a que en la literatura estudiada se mencionaba que el estado de una tiroides se mide principalmente bajo los niveles de esas dos hormonas y los diagnósticos de los distintos tipos de hipotiroidismo también se realizan en base a ellas.

Uno de los modelos entrenados fue un modelo de regresión logística multinomial, el cual reveló que las variables on thyroxine y thyroid surgery tienen una alta importancia cuando se trata de predecir sobre todas las clases al mismo tiempo, estas variables tratan respecto a si la persona esta en tratamiento con tiroxina y si ha recibido una cirugía a la tiroides. Estos resultados fueron inesperados al no haber logrado encontrar otro estudio respecto a clasificación con multiples variables, aunque se hace la conjetura que las importancias de estas variables tienen sentido tomando en consideración que las personas en tratamiento con tiroxina o que han recibido cirugía a la tiroides es porque poseen algún tipo de afección a la tiroides, por lo que se encuentran si o si dentro de algún grupo distinto al negativo.

Respecto a posibles mejores en el desarrollo de esta experiencia, queda pendiente la utilización de distintos métodos en modelos de predicción con el fin de extraer aún más información de las variables involucradas e incluso confirmar aquellas variables que poseen una gran importancia para predecir el estado de un paciente.

Bibliografía

- American Thyroid Association (2017). Bocio. http://www.thyroid.org/wp-content/uploads/patients/brochures/espanol/ata_bocio_w.pdf.
- American Thyroid Association (s.f.). Hipotiroidismo. <https://www.thyroid.org/hipotiroidismo/>.
- 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>.
- Children’s Oncology Group (s.f.). Thyroid problems. <https://www.childrensoncologygroup.org/index.php/thyroidproblems>.
- Cleveland Clinic (2019). Thyroid blood tests. <https://my.clevelandclinic.org/health/diagnostics/17556-thyroid-blood-tests>.
- Dua, D. and Graff, C. (2017). UCI machine learning repository.
- Haldeman-Englert, C. (2020). T4 libre. https://myhealth.ucsd.edu/Spanish/RelatedItems/167,free_t4_hydroxycortisol.
- Jamal, M. T., Li, Q.-L., Li, Q.-Y., Liang, W.-Y., Wang, L.-H., Wei, J.-H., Liang, Q., Hu, N.-Q., and Li, L.-H. (2021). Association of thyroid hormones with blood pressure and arterial stiffness in the general population: The dali study. *The Journal of Clinical Hypertension*, 23(2):363–372.
- Krawiec, L., Juvenal, G., and Pisarev, M. (s.f.). Bases fisiológicas.
- Lozano, J. (2006). Hipotiroidismo. *Offarm*, 25(1):61–65.
- Léniz, J. (s.f.). Hipotiroidismo subclínico: ¿tratar o no tratar? <https://medicina.uc.cl/publicacion/hipotiroidismo-subclinico-tratar-o-no-tratar/>.
- Manual MSD (2020). Hipotiroidismo (mixedema). <https://www.msdmanuals.com/es-cl/professional/trastornos-endocrinol%C3%B3gicos-y-metab%C3%B3licos/trastornos-tiroideos/hipotiroidismo>.

- McMaster Textbook of Internal Medicine (2021). Hypothyroidism.
<https://empendium.com/mcmtextbook/chapter/B31.II.9.1>.
- Medici, M., Korevaar, T. I., Visser, W. E., Visser, T. J., and Peeters, R. P. (2015). Thyroid function in pregnancy: what is normal? *Clinical chemistry*, 61(5):704–713.
- Medline Plus (2020). Hipotiroidismo. <https://medlineplus.gov/spanish/ency/article/000353.htm>.
- Medline Plus (2022). Examen de tbg en la sangre.
<https://medlineplus.gov/spanish/ency/article/003374.htm>.
- Mumtaz, M., Lin, L. S., Hui, K. C., and Khir, A. S. M. (2009). Radioiodine i-131 for the therapy of graves’ disease. *The Malaysian journal of medical sciences: MJMS*, 16(1):25.
- Radiology Info (2020). Enfermedad de la glándula tiroides.
<https://www.radiologyinfo.org/es/info/thyroid-disease>.
- Ratini, M. (2020). Hypopituitarism. <https://www.webmd.com/brain/hypopituitary>.
- Redford, Christopher y Vaidya, B. (2017). Subclinical hypothyroidism: Should we treat? *Post reproductive health*, 23(2):55–62.
- Surks, M. I. (2021). Lithium and the thyroid. *UpToDate*.
<https://www.uptodate.com/contents/lithium-and-the-thyroid>.
- Zantour, B. and Chebbi, W. (2014). Lithium treatment and thyroid disorders. *Thyroid Disorders Ther*, 3(143):2.