

Introducción a la Ciencia de Datos

Trabajo teórico-práctico

Antonio Manjavacas Lucas

10/12/2020

Índice

1. Análisis exploratorio de los datos	1
1.1. Dataset <i>wankara</i>	1
1.1.1. Estructura y variables	2
1.1.2. Distribución de los datos	3
1.1.3. Normalidad	5
1.1.4. Correlación	5
1.1.5. Hipótesis	9
1.2. Dataset <i>newthyroid</i>	13
1.2.1. Estructura y variables	14
1.2.2. Balanceo de clases	16
1.2.3. Distribución de los datos	16
1.2.4. Normalidad	19
1.2.5. Correlación	21
1.2.6. Conclusiones	21
2. Regresión	23
2.1. Regresión lineal	23
2.2. Regresión lineal múltiple	23
2.3. Regresión mediante <i>k-NN</i>	26
2.4. Comparativa	26
3. Clasificación	27
3.1. Clasificación mediante <i>k-NN</i>	27
3.2. Clasificación mediante <i>LDA</i>	27
3.3. Clasificación mediante <i>QDA</i>	29
3.4. Comparativa	30

1. Análisis exploratorio de los datos

En esta primera sección del trabajo realizaremos el análisis exploratorio de los datos a emplear en los problemas de regresión y clasificación propuestos.

1.1. Dataset *wankara*

Comencemos estudiando el dataset sobre el que aplicaremos regresión: *wankara*. Este conjunto de datos contiene la información meteorológica de Ankara (Turquía) desde el 01/01/1994 hasta el 28/05/1998. A partir

Tabla 1: Primeras filas del dataset wankara

Max_temperature	Min_temperature	Dewpoint	Precipitation	Sea_level_pressure	Standard_pressure	Visibility	Wind_speed	Max_wind_speed	Mean_temperature
40.3	16.9	24.6	0.05	30.12	26.83	8.5	4.37	13.80	30.7
73.0	43.2	34.9	0.00	30.19	27.03	8.1	7.48	12.70	55.6
30.6	11.8	14.1	0.00	30.47	27.03	2.6	3.11	9.21	19.3
35.6	28.0	23.2	0.00	30.15	26.81	8.2	1.27	8.06	30.5
55.6	46.4	43.5	0.10	29.73	26.57	8.2	6.56	10.20	50.1
82.4	53.6	46.3	0.00	29.78	26.66	8.9	9.55	20.80	69.0
31.3	0.3	4.6	0.00	30.30	26.82	6.9	0.00	13.31	14.1
55.4	24.4	28.1	0.00	30.04	26.76	7.2	5.75	14.90	39.8
66.2	50.0	44.1	0.00	29.85	26.67	9.2	12.20	57.40	58.7
84.2	48.2	42.1	0.00	29.79	26.72	8.6	4.60	11.30	67.9

de una serie de atributos meteorológicos, el objetivo será predecir la temperatura media alcanzada en un día.

1.1.1. Estructura y variables

Una vez cargado el conjunto de datos, se lleva a cabo una exploración preliminar de los mismos:

- Contamos con **1609 filas** (observaciones) y **10 columnas** (9 atributos y 1 variable a predecir). La Tabla 1 muestra un pequeño subconjunto de los datos con los que trabajaremos.
- Además, si analizamos los datos, observamos que todas las columnas albergan datos de tipo numérico y que no existe ningún *missing value* (*NA*) ni registro duplicado.
- Finalmente, no existen variables compuestas ni redundantes y, en general, los datos se interpretan correctamente.

A continuación se procede a estudiar las variables del dataset, de las cuales se supone la siguiente interpretación¹:

- *Max_temperature*: temperatura máxima alcanzada en el día (se asume en °F).
- *Min_temperature*: temperatura mínima alcanzada en el día (se asume en °F).
- *Dewpoint*: punto de rocío. Se trata de la temperatura a la cual una masa de aire debe enfriarse para provocar condensación (se asume en °F).
- *Precipitation*: cantidad de precipitaciones. Dado el rango de valores, parece estar medida en pulgadas (in).
- *Sea_level_pressure*: presión atmosférica a nivel del mar. Del rango de valores se asume que está medida en pulgadas de mercurio (inHg), siendo 29,92 la presión normal a nivel del mar. En general, suele encontrarse entre 29 y 31, como ocurre en nuestro conjunto de datos.
- *Standard_pressure*: presión atmosférica en la superficie. Al igual que en el caso anterior, se mide en inHg.
- *Visibility*: nivel de visibilidad. Normalmente se mide en metros (m) e indica la distancia a la cual un objeto o luz puede ser claramente identificado.
- *Wind_speed*: velocidad de viento. Se asume en millas por hora (MPH).
- *Max_wind_speed*: velocidad máxima de viento (MPH).
- *Mean_temperature*: temperatura media alcanzada en el día (se asume en °F). Es la variable a predecir.

A partir de estas variables, son varias las **hipótesis** preliminares que podemos plantear. Por ejemplo:

- ¿Los días en los que se registran mayores temperaturas la visibilidad es mayor? Esto podría deberse a una mayor incidencia de la luz solar en la superficie.
- ¿En los días con más precipitaciones cuál tiende a ser la temperatura media? ¿Las tormentas son más comunes con altas o bajas temperaturas?
- ¿Cómo afecta la presión a la temperatura media?
- ¿Y el viento?

¹Para esta interpretación de los datos se ha tomado como principal referencia la fuente: https://en.wikipedia.org/wiki/Surface_weather_observation.

Planteadas estas cuestiones trataremos de darles respuesta más adelante.

1.1.2. Distribución de los datos

Veamos ahora cómo se distribuyen los datos. Primero, observemos los boxplots e histogramas de las Figuras 1 y 2, respectivamente. De los boxplots detectamos una mayor cantidad de valores anómalos en las variables *Precipitation*, *Visibility*, así como *Wind_speed* y *Max_wind_speed*. Es comprensible que se dé esta situación, ya que se trata de variables que pueden tomar valores extremos en situaciones muy concretas. Podemos argumentar que, aunque no todos los días llueve en grandes cantidades, o no todos los días la velocidad del viento es muy alta, al ser algo que puede ocurrir si se dan las condiciones necesarias, estos no son *outliers* que realmente deseemos eliminar (no son errores de medición, sino casos atípicos).

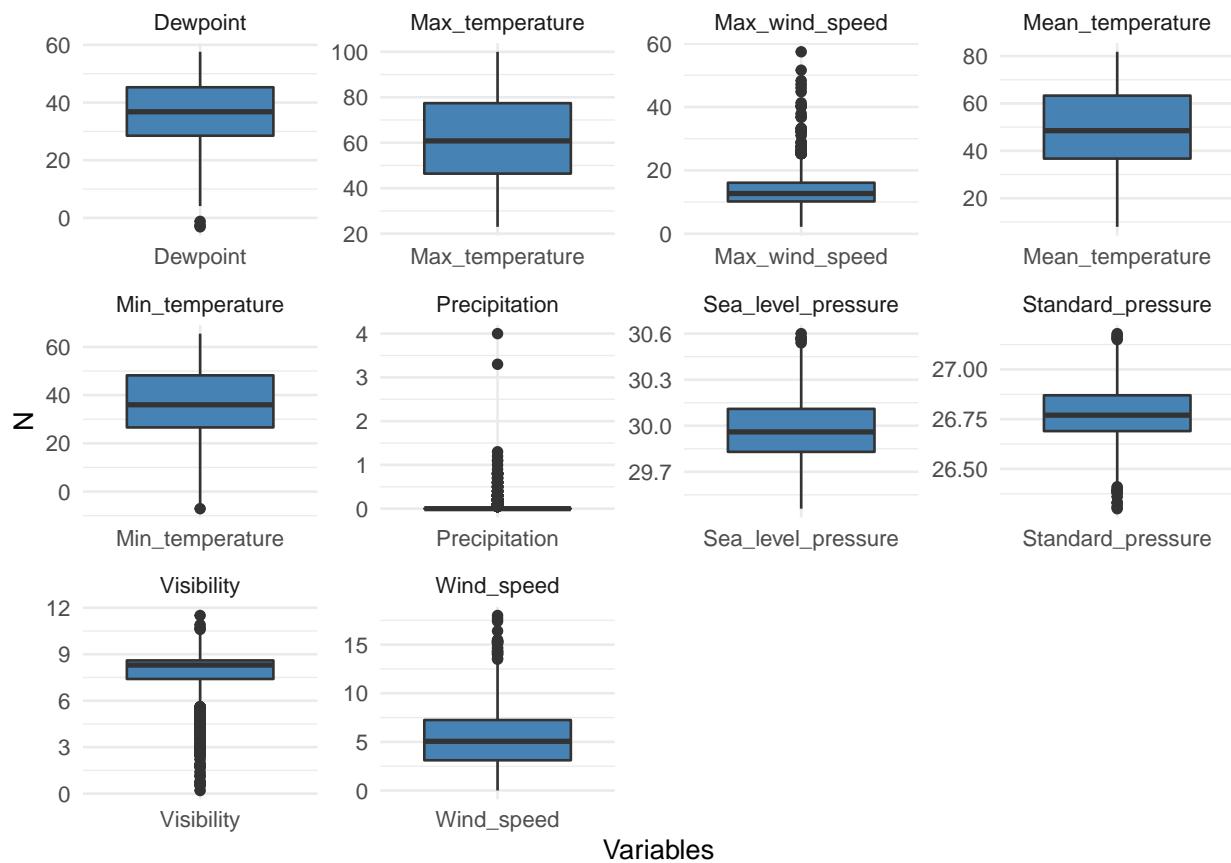


Figura 1: Distribución de los datos: boxplots

Tras observar gráficamente la distribución de los datos, calculamos su asimetría (*skewness*), obteniendo los siguientes resultados (ver Tabla 2):

- *Skewness* negativa (distribución desplazada a la derecha): *Dewpoint*, *Visibility*.
- *Skewness* positiva (distribución desplazada a la izquierda): *Precipitation*, *Sea_level_pressure*, *Wind_speed*, *Max_wind_speed*.
- *Skewness* cercana a cero (distribución centrada): *Max_temperature*, *Min_temperature*, *Mean_temperature*, *Standard_pressure*.

También podemos conocer la curtosis (*kurtosis*) de las distribuciones (ver Tabla 3). Mayores valores de curtosis implican curvas más pronunciadas, mientras que valores bajos implican distribuciones más achatadas.

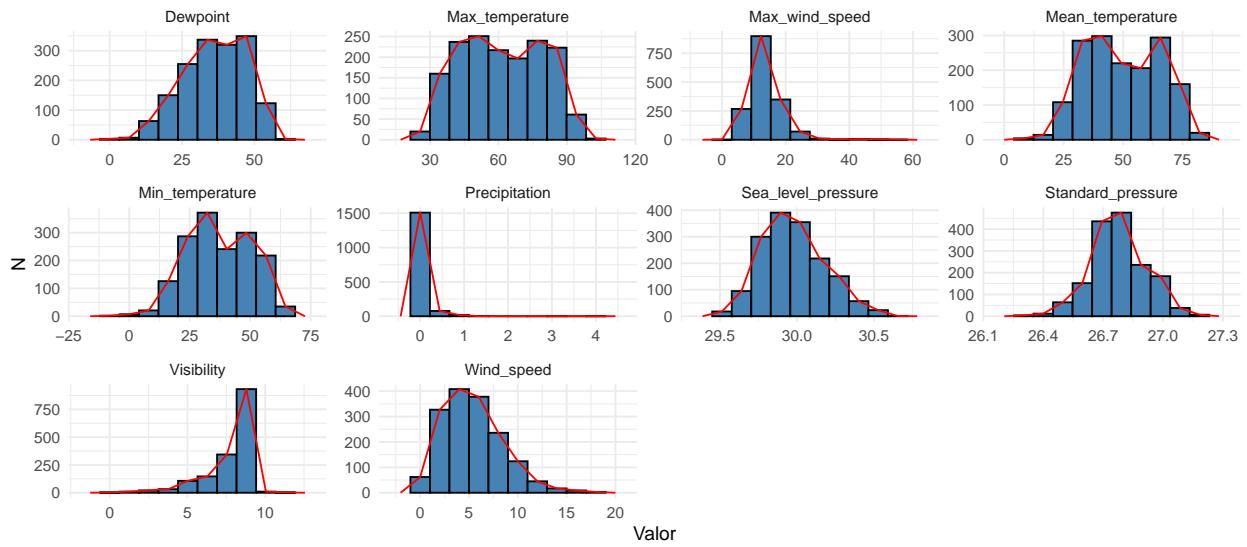


Figura 2: Distribución de los datos: histogramas

Tabla 2: Asimetría de las variables

	skewness
Max_temperature	0.0198032
Min_temperature	-0.0554807
Dewpoint	-0.3744065
Precipitation	11.0895643
Sea_level_pressure	0.4089653
Standard_pressure	-0.0083340
Visibility	-1.9496689
Wind_speed	0.7328759
Max_wind_speed	1.8568241
Mean_temperature	0.0342041

Tabla 3: Curtosis de las variables

	kurtosis
Max_temperature	1.864299
Min_temperature	2.293897
Dewpoint	2.520686
Precipitation	197.364517
Sea_level_pressure	2.818564
Standard_pressure	3.153448
Visibility	7.216549
Wind_speed	3.517378
Max_wind_speed	11.638890
Mean_temperature	1.942302

Tabla 4: Normalidad de las variables

vars	statistic	p_value	sample
Max_temperature	0.9658556	4.820105e-19	1609
Min_temperature	0.9804859	5.531039e-14	1609
Dewpoint	0.9772369	2.723394e-15	1609
Precipitation	0.2960197	0.000000e+00	1609
Sea_level_pressure	0.9853185	9.759597e-12	1609
Standard_pressure	0.9951595	4.616980e-05	1609
Visibility	0.7748121	0.000000e+00	1609
Wind_speed	0.9650800	2.897973e-19	1609
Max_wind_speed	0.8878523	0.000000e+00	1609
Mean_temperature	0.9682132	2.386456e-18	1609

En general, las variables con curtosis y asimetría más cercanas a 0 presentarán una mayor normalidad. De forma preliminar, los resultados obtenidos en esta sección nos revelan que es bastante improbable que alguna de nuestras variables siga una distribución normal. Aun así, trataremos de corroborar esta afirmación aplicando el test estadístico correspondiente.

1.1.3. Normalidad

Evaluaremos la normalidad de las variables aplicando el test de *Shapiro-Wilk*, obteniendo los resultados que se muestran en la Tabla 4. Como podemos observar, para ningún p-value se cumple: $p\text{-value} > 0,05$, lo cual implica que **no podemos asumir la normalidad de ninguna de las variables**. Realmente este no es un problema a la hora de emplear regresión, ya que no requiere que los datos sigan una distribución normal para ser aplicada. Igualmente, es una información útil a tener en cuenta por si posteriormente fuese necesario realizar algún tipo de test estadístico sobre los datos.

- Reincidiendo en la idea de la no normalidad de nuestros datos, y viendo los resultados del test estadístico aplicado, la variable que tal vez presenta una mayor cercanía a una distribución normal es *Standard_pressure*. Aun así, como ya hemos adelantado, no podemos asumir su normalidad, tal y como se muestra en el grafico Q-Q de la Figura 3.

Algo que sí resulta de especial interés a la hora de aplicar regresión (especialmente si vamos a emplear modelos con interacción entre variables o *k-NN*) es la **estandarización** de nuestros datos (*Z-score*). Tras aplicarla, obtenemos las distribuciones que se muestran en la Figura 4, que son similares a las ya vistas pero con $\mu = 0$ y $\sigma = 1$.

1.1.4. Correlación

Un paso previo esencial antes de aplicar regresión es estudiar la correlación entre las variables de las que disponemos. Aquellas variables más correladas con la variable objetivo serán mejores candidatas a formar parte de nuestro modelo. Estas correlaciones se muestran en detalle en las Figuras 5 y 6.

Las variables que presentan una mayor correlación con la temperatura media son: las temperaturas máxima y mínima, el punto de rocío, la presión atmosférica a nivel del mar y la visibilidad. Por otro lado, el viento, las precipitaciones y la presión estándar parecen estar menos relacionadas con nuestra variable objetivo.

- Si atendemos a estas relaciones, tiene sentido que las variables *Max_temperature* y *Min_temperature* estén positivamente relacionadas con *Mean_temperature*, ya que la temperatura media siempre se encuentra entre ambas. Debemos tener en cuenta que variables como el viento o las precipitaciones pueden tener cierta influencia sobre la temperatura media, haciéndola variar ligeramente, por lo que la

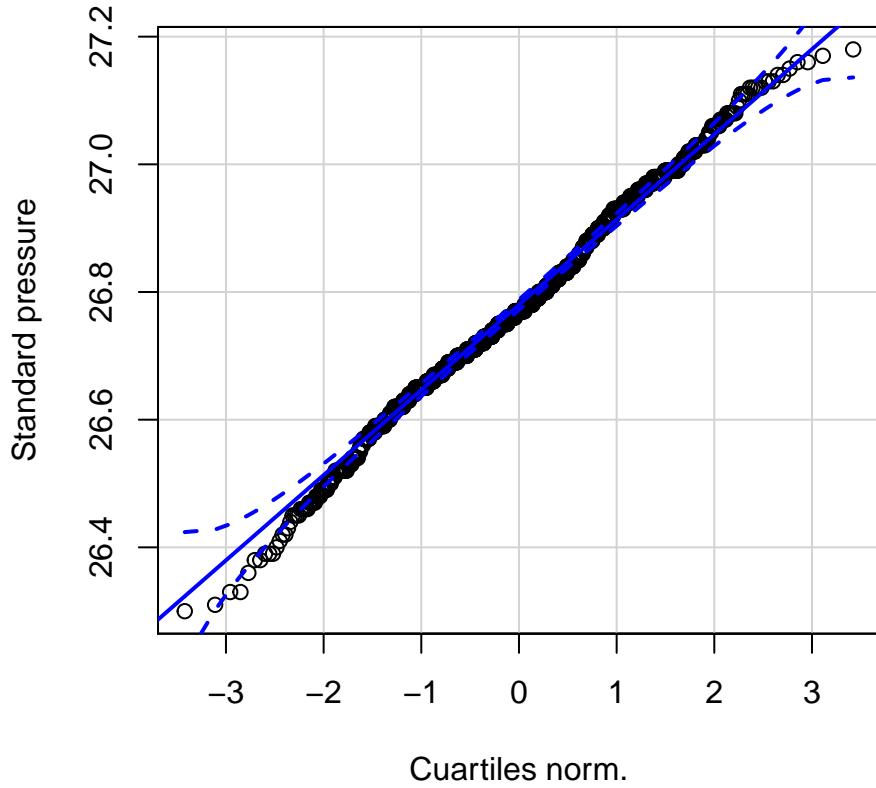


Figura 3: Gráfico Q-Q de la variable Standard_pressure

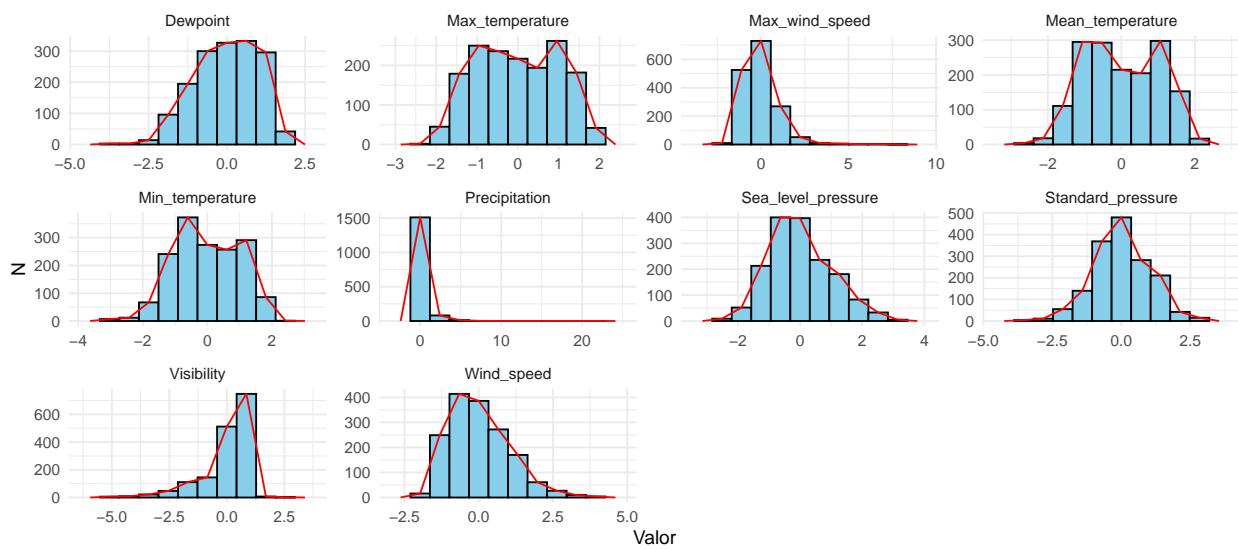


Figura 4: Histogramas de las variables estandarizadas

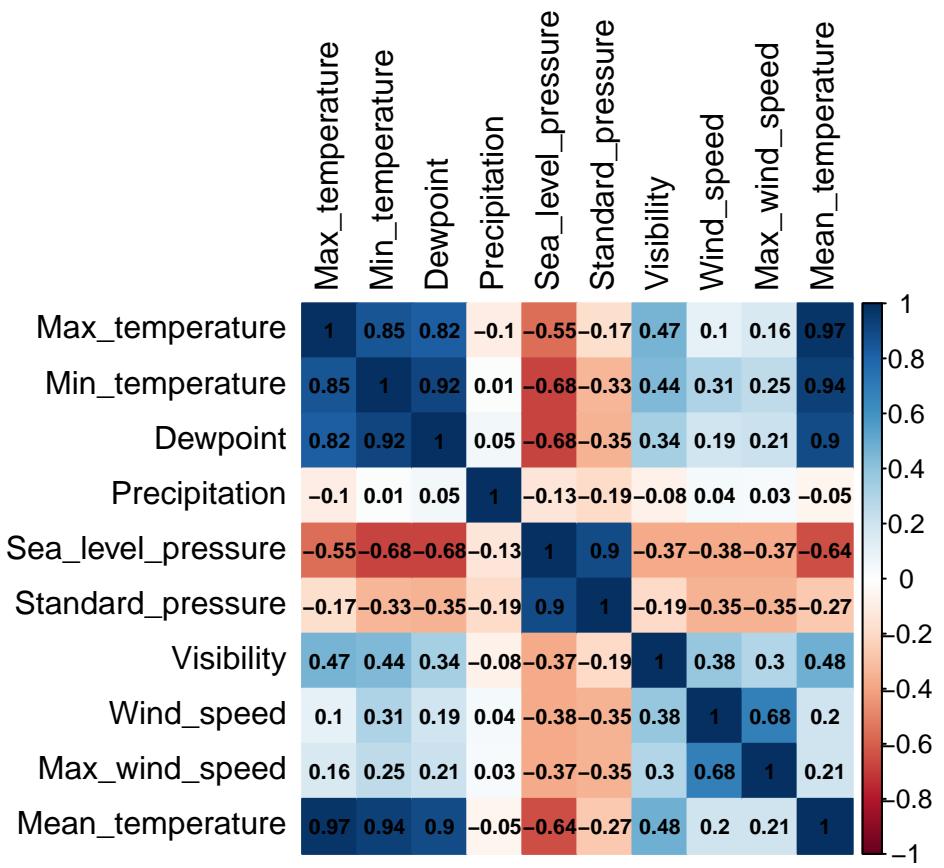


Figura 5: Correlación entre las variables del dataset

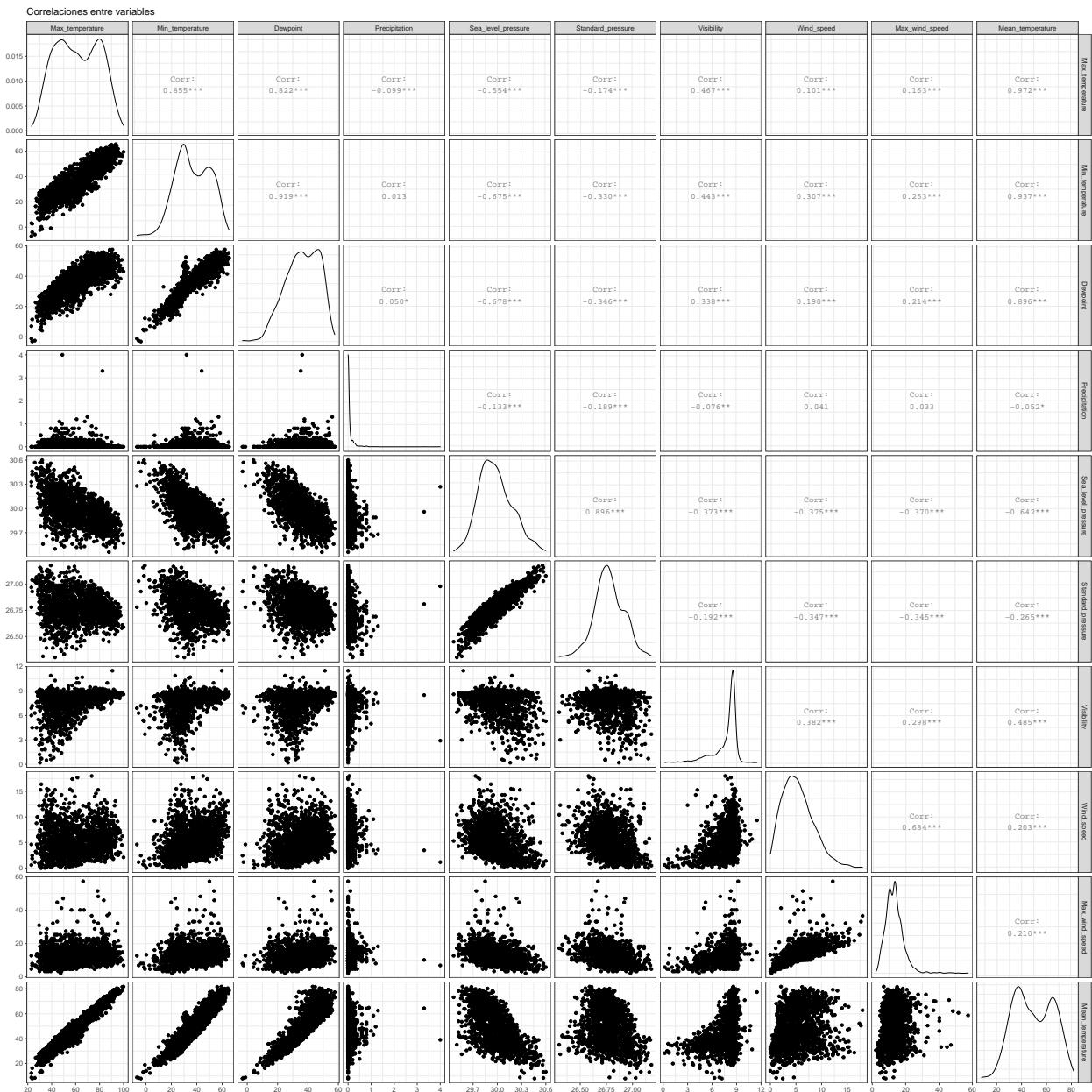


Figura 6: Relaciones entre las variables del dataset

media que podemos obtener a partir de las temperaturas máxima y mínima no será exactamente igual, pero sí un valor muy aproximado a la temperatura media real.

- El punto de rocío (*Dewpoint*) también presenta una severa correlación positiva con la temperatura media, de una forma similar a las temperaturas máxima y mínima.
- Por otro lado, *Sea_level_pressure* (y *Standard_pressure* en menor medida) están relacionadas negativamente con la temperatura media. Nos podría interesar estudiar esta correlación en detalle.
- La visibilidad, o *Visibility*, presenta una correlación positiva con la temperatura media. Como ya hemos hipotetizado, los días más calidos podrían ser aquellos con una mayor incidencia de luz solar en la superficie, lo que supondría este crecimiento de las temperaturas.
- Finalmente, resulta llamativo el cómo la variable *Precipitation* presenta una gran independencia del resto y, en general, parece aportarnos poca información.

De esta forma, las variables: *Max_temperature*, *Min_temperature*, *Dewpoint*, y posiblemente *Sea_level_pressure* y *Visibility* podrían ser las más apropiadas a la hora de estimar la temperatura media (*Mean_temperature*). No obstante, de cara a simplificar el modelo de regresión múltiple, de entre las variables *Max_temperature*, *Min_temperature* y *Dewpoint* deberíamos optar por mantener únicamente una de ellas, ya que están altamente correladas entre sí y añadirían redundancia.

La influencia de las diferentes variables sobre el modelo será algo con lo que se experimentará en el apartado de regresión, el cuál será abordado sobre el conocimiento extraído de este análisis previo.

1.1.5. Hipótesis

Tratemos ahora de dar respuesta a las preguntas planteadas al principio de este análisis:

¿Los días en los que se registran mayores temperaturas la visibilidad es mayor? Esto podría deberse a una mayor incidencia de la luz solar en la superficie.

Si observamos la Figura 7, observamos una alta concentración de días con una visibilidad de entre 7 y 9. Si atendemos a la *Escala Internacional de Visibilidad en el Aire*² (que se intuye es la referencia tomada en el dataset para medir la visibilidad), dichos niveles se corresponden con atmósfera diáfana con niveles de visibilidad *bueno* (7), *muy bueno* (8) y *excelente* (9), respectivamente. De esta información podemos extraer que la mayoría de días en Ankara la visibilidad es buena y que los días cálidos tienden a presentar una alta visibilidad por lo general, respondiendo así a la pregunta planteada.

Sin embargo, podemos extraer aun más conclusiones: si atendemos a los días con baja visibilidad (niebla), normalmente se trata de días con bajas temperaturas, como cabría esperar. No obstante, no todos los días con bajas temperaturas son días con niebla. De hecho, la mayor cantidad de días con bajas temperaturas son días con una alta visibilidad.

Es por esto por lo que podemos concluir en que mientras **una baja visibilidad tiende a asociarse a días fríos, en los días despejados la visibilidad es un mal predictor de la temperatura**. No obstante, y a modo de conclusión, de la Figura 7 podemos observar que la distribución de los puntos podría corresponderse con una función cuadrática o incluso exponencial, algo que consideraremos a la hora de construir nuestro modelo de regresión.

¿En los días con más precipitaciones cuál tiende a ser la temperatura media? ¿Las tormentas son más comunes con altas o bajas temperaturas?

Lo primero que podemos observar en la relación entre la temperatura media y las precipitaciones (ver Figura 8) es la presencia de dos posibles *outliers*, tal y como se muestran en la Figura 9. Podemos ver en detalle estas observaciones en la Tabla 5, donde se observa que realmente no son días con valores anómalos en sus

²Véase <https://www.tiempo.com/ram/1041/meteorologa-viila-visibilidad-y-los-factores-meteorológicos-que-influyen-en-ella/> y <https://www.semanticscholar.org/paper/Analysis-of-the-atmospheric-visibility-Restoration-Deshpande/662237bb893d2b50a728751880f20cf1f8225ae>

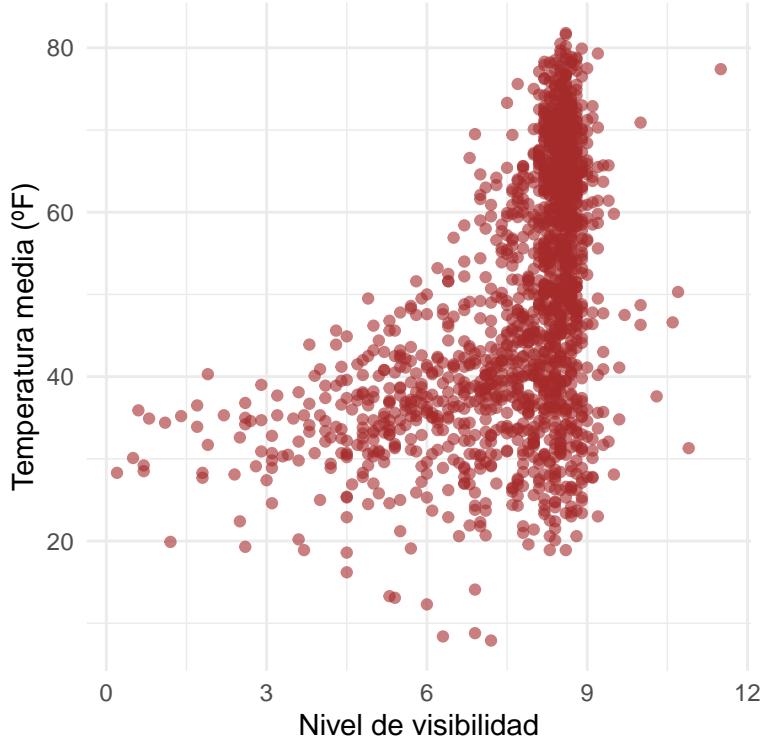


Figura 7: Relación entre temperatura media y visibilidad

Tabla 5: Días con precipitaciones anómalas

Max_temperature	Min_temperature	Dewpoint	Precipitation	Sea_level_pressure	Standard_pressure	Visibility	Wind_speed	Max_wind_speed	Mean_temperature
48.9	32.0	36.2	4.0	30.27	26.98	2.9	1.15	6.9	39.0
82.4	43.9	35.3	3.3	29.96	26.81	8.5	3.45	10.2	64.4

predictores. Ya que no disponemos de información suficiente como para catalogarlos como errores de medición y pudiendo esa cantidad atípica de precipitaciones deberse a otros factores no contemplados en el dataset (por ejemplo, concentración de nubes), se ha considerado directamente no tenerlos en cuenta.

Una vez aclarado el tratamiento de las observaciones atípicas, encontramos una distribución de las precipitaciones más parecida a una variable categórica que a una continua (ver Figura 10). Esto parece deberse a que en la medición de las precipitaciones no se emplean valores continuos sino discretos dentro, posiblemente, de la escala de un pluviómetro.

Sin hacer un estudio más profundo de la variable y ante las limitaciones de cómo se encuentra medida en nuestro conjunto de datos, **no podemos confirmar que las precipitaciones y la temperatura media estén estrechamente relacionadas**. Esto la convierte en una variable descartable de nuestro modelo.

¿Cómo afecta la presión a la temperatura media?

A primera vista, tanto en el caso de la presión estándar (Figura 11) como de la presión a nivel del mar (Figura 12) observamos cierta correlación negativa, ya descubierta en secciones previas.

Especialmente en el caso a nivel del mar, observamos que a menor presión mayor es la temperatura media; esto podría deberse al siguiente fenómeno³:

³https://es.wikipedia.org/wiki/Presi%C3%B3n_atmosf%C3%A9rica

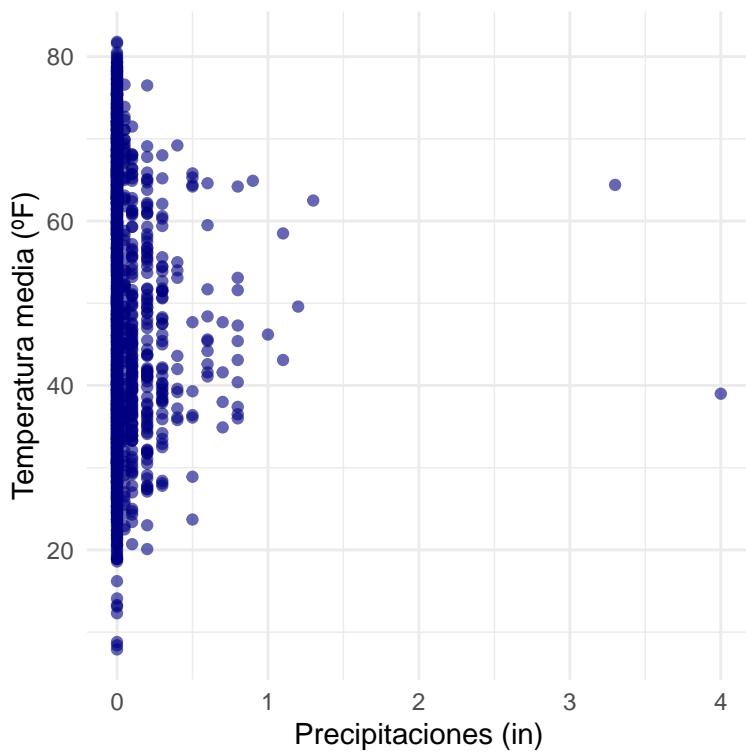


Figura 8: Relación entre la temperatura media y las precipitaciones

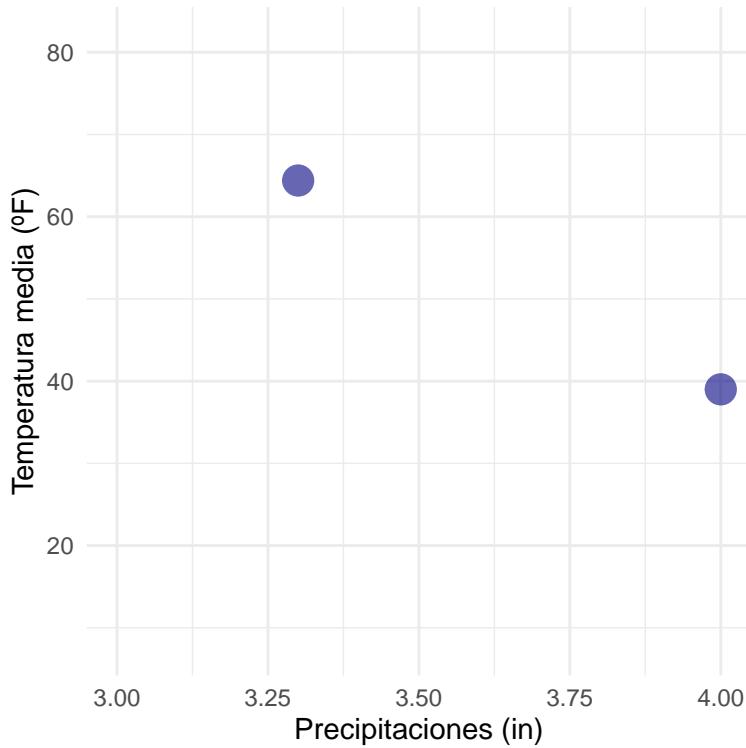


Figura 9: Valores anómalos en la relación temperatura media - precipitaciones

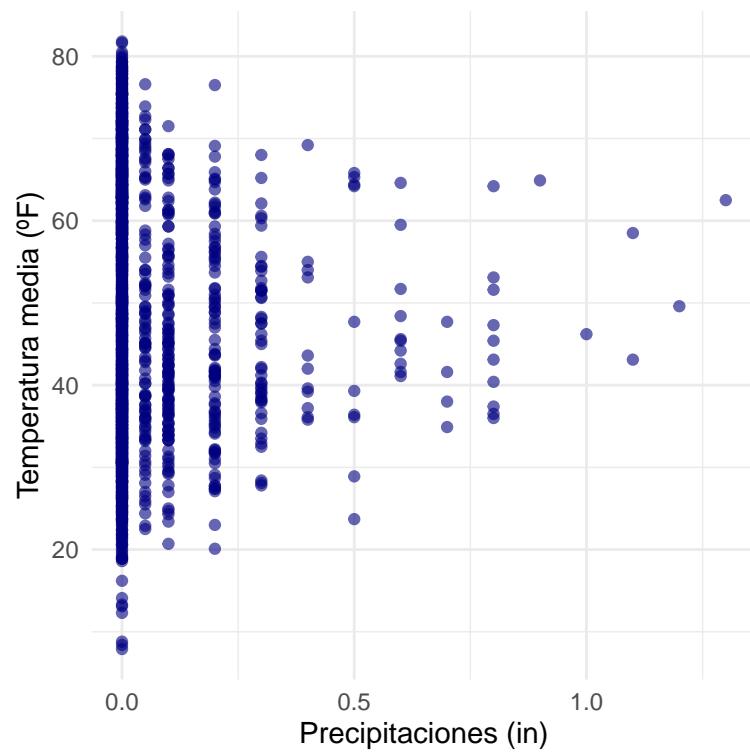


Figura 10: Relación entre temperatura media y las precipitaciones (sin valores anómalos)

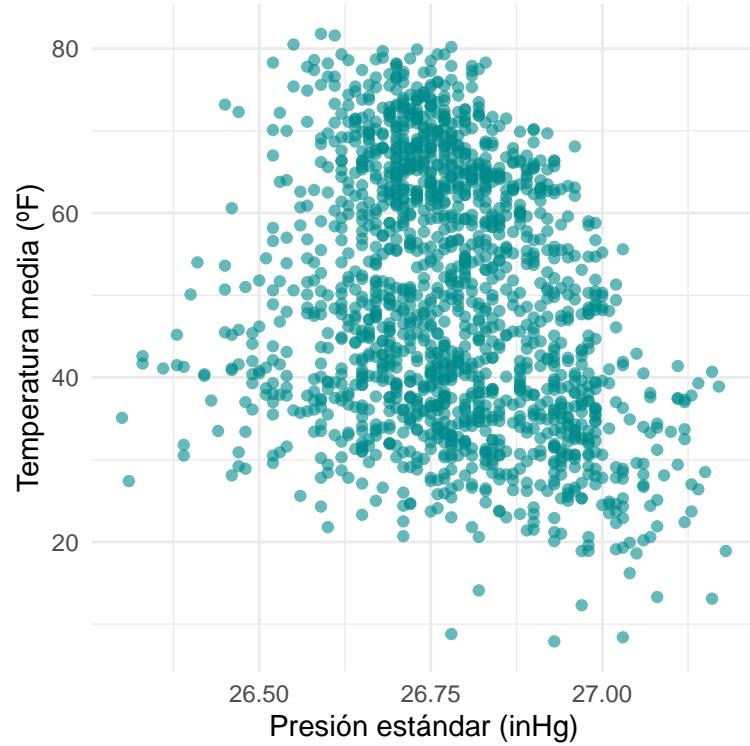


Figura 11: Relación entre temperatura media y la presión estándar

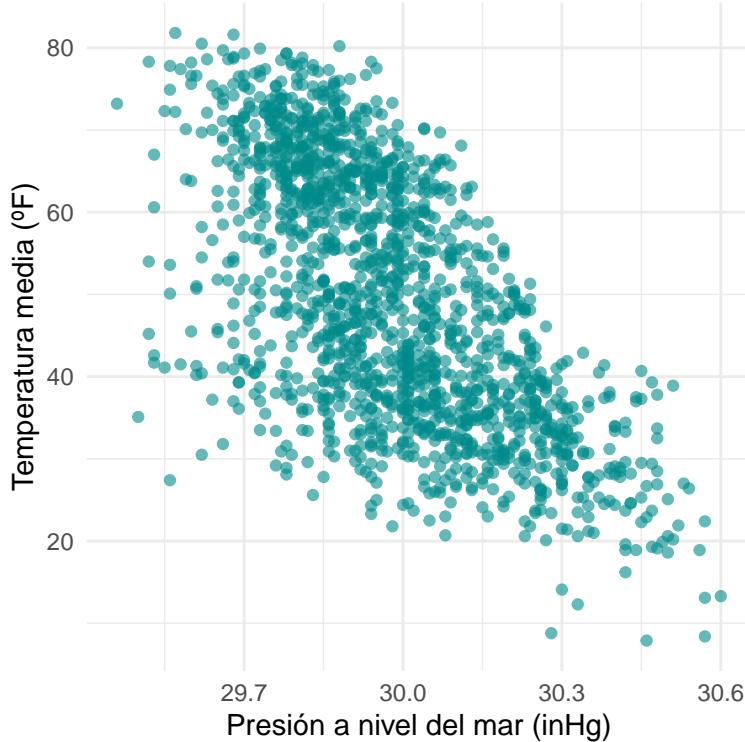


Figura 12: Relación entre temperatura media y la presión a nivel del mar

- A bajas temperaturas, el aire se contrae, aumentando su densidad y, por lo tanto, descendiendo, de tal forma que aumenta la presión en la superficie.
- Si, por el contrario, las temperaturas son altas, el aire asciende y la presión disminuye.

Independientemente de la interpretación meteorológica de esta relación, observamos que **la presión muestra una correlación significativa con la temperatura media**, por lo que ambas variables serían buenas candidatas a formar parte de nuestro modelo. No obstante, de cara a simplificar nuestro modelo de regresión múltiple y evitar variables que redundan en la misma información, seguramente *Sea_level_pressure* sería la variable que mejor representa la relación entre presión y temperatura.

¿Cómo afecta el viento a la temperatura media?

Finalmente, la relación entre viento y temperatura media mostrada en la Figura 13 es poco esclarecedora. Tal vez pueda entreverse una velocidad del viento ínfimamente mayor para temperaturas altas, pero definitivamente no es un variable que vaya a resultar de gran utilidad en nuestro modelo. Afirmaremos, pues, que **a partir de los datos proporcionados, no existe relación directa entre la temperatura media y la velocidad del viento**.

Estudiado el conjunto de datos destinados a regresión, pasemos ahora a abordar el dataset empleado en el problema de clasificación.

1.2. Dataset *newthyroid*

El dataset empleado para nuestro problema de clasificación es *newthyroid*. Este conjunto de datos es una de los múltiples conjuntos de datos sobre Tiroides disponibles en el UC Irvine Machine Learning Repository

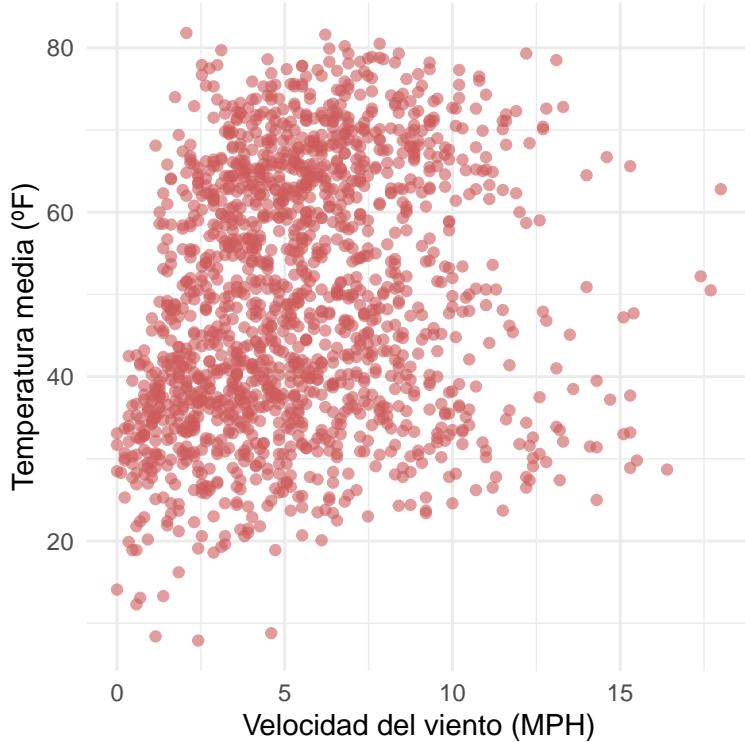


Figura 13: Relación entre temperatura media y la velocidad del viento

(UCI)⁴. Nuestro objetivo será detectar si un determinado paciente es normal (1) o si padece hipertiroidismo (2) o hipotiroidismo (3). Cada tipo de paciente es una clase a identificar (atributo *Class*). En la Tabla 6 se muestran las primeras filas del dataset.

1.2.1. Estructura y variables

Comencemos viendo la estructura del conjunto de datos e información relevante sobre el mismo:

- Contamos con **215 filas** (observaciones) y **6 columnas** (5 atributos y 1 clase a predecir).
- Con respecto al tipo de datos, la columna *T3resin* está compuesta por valores numéricos enteros, mientras que *Thyroxin*, *Triiodothyronine*, *Thyroidstimulating* y *TSH_value* son variables numéricas con decimales. Finalmente, el atributo *Class* (clase a predecir) ha sido convertido en un factor (variable categórica), ya que inicialmente fue cargada como un atributo numérico entero.
- No existen *missing values (NAs)* ni registros duplicados.
- En principio, no contamos con variables compuestas ni redundantes, aunque trataremos de asegurarnos de esto último tratando de conocer a fondo las variables del dataset.

Las variables con las que trabajaremos son las siguientes⁵:

- *T3resin*: captación de resina T3 (también llamada “captación T3” o “T3RU”). Es un análisis de sangre realizado como parte de una evaluación de la función tiroidea.
- *Triiodothyronine*: la triyodotironina (o T3), es una hormona tiroidea. Afecta a casi todos los procesos fisiológicos en el cuerpo, incluyendo crecimiento y desarrollo, metabolismo, temperatura corporal y ritmo cardíaco.

⁴<https://archive.ics.uci.edu/ml/index.php>

⁵Se han tomado como referencia las siguientes fuentes: <https://es.wikipedia.org/wiki/Triyodotironina>, <https://es.wikipedia.org/wiki/Tirotropina>, <https://www.healthline.com/health/tsh#results>

Tabla 6: Primeras filas del dataset newthyroid

T3resin	Thyroxin	Triiodothyronine	Thyroidstimulating	TSH_value	Class
107	10.1	2.2	0.9	2.7	1
113	9.9	3.1	2.0	5.9	1
127	12.9	2.4	1.4	0.6	1
109	5.3	1.6	1.4	1.5	1
105	7.3	1.5	1.5	-0.1	1
105	6.1	2.1	1.4	7.0	1
110	10.4	1.6	1.6	2.7	1
114	9.9	2.4	1.5	5.7	1
106	9.4	2.2	1.5	0.0	1
107	13.0	1.1	0.9	3.1	1

Tabla 7: Niveles de TSH medios para cada tipo de paciente

Class	Mean TSH	Median TSH
Normal	2.516667	2.20
Hipertiroidico	-0.020000	0.00
Hipotiroidico	17.533333	12.05

- *Thyroxin*: tiroxina (o T4). Es otra hormona tiroidea que se mide con el mismo propósito que la T3.
- *Thyroidstimulating*: la tirotropina, hormona estimulante de la tiroideas, hormona tiroestimulante u hormona tirotrópica (abreviada también TSH, del inglés Thyroid-Stimulating Hormone) es una hormona producida por la hipófisis que regula la producción de hormonas tiroideas por la glándula tiroideas.
- *TSH_value*: el rango normal de los niveles de TSH es de 0.4 a 4.0. A priori, se trata de un atributo realmente importante a la hora de diagnosticar la tiroideas si nos basamos en conocimiento experto:
 - Un valor por encima del rango normal generalmente indica que la tiroideas está poco activa. **Esto indica hipotiroidismo**. Cuando la tiroideas no produce suficientes hormonas, la glándula pituitaria libera más TSH para intentar estimularla.
 - Un valor por debajo del rango normal significa que la tiroideas está hiperactiva. **Esto indica hipertiroidismo**. Cuando la tiroideas produce demasiadas hormonas, la glándula pituitaria libera menos TSH. Esto puede corroborarse claramente en nuestro conjunto de datos (lo abordaremos más adelante).
- *Class*: tipo de paciente: *normal* (1), *hipertiróidico* (2) o *hipotiróidico* (3).

Si bien se trata de un campo de conocimiento muy especializado, a partir de la información recabada de las diferentes fuentes podemos hacernos una idea de las variables que emplearemos para clasificar. Por ejemplo, *TSH_value* parece ser un muy buen predictor del tipo de paciente. De hecho, a primera vista podemos ver que el razonamiento sobre la relación entre los niveles de TSH y el hipotiroidismo/hipertiroidismo es acertado y se corresponde con las observaciones del dataset, tal y como se muestra en la Tabla 7 y la Figura 14.

Por otro lado, una vez interpretados los datos, podemos plantearnos algunas preguntas iniciales, tales como:

- ¿Hay alguna diferencia entre *T3resin* y *Triiodothyronine*? ¿Ambas son mediciones relacionadas con la hormona T3 o reflejan información diferente?
- De la misma forma, ¿existen diferencias entre *Thyroidstimulating* y *TSH_value*?

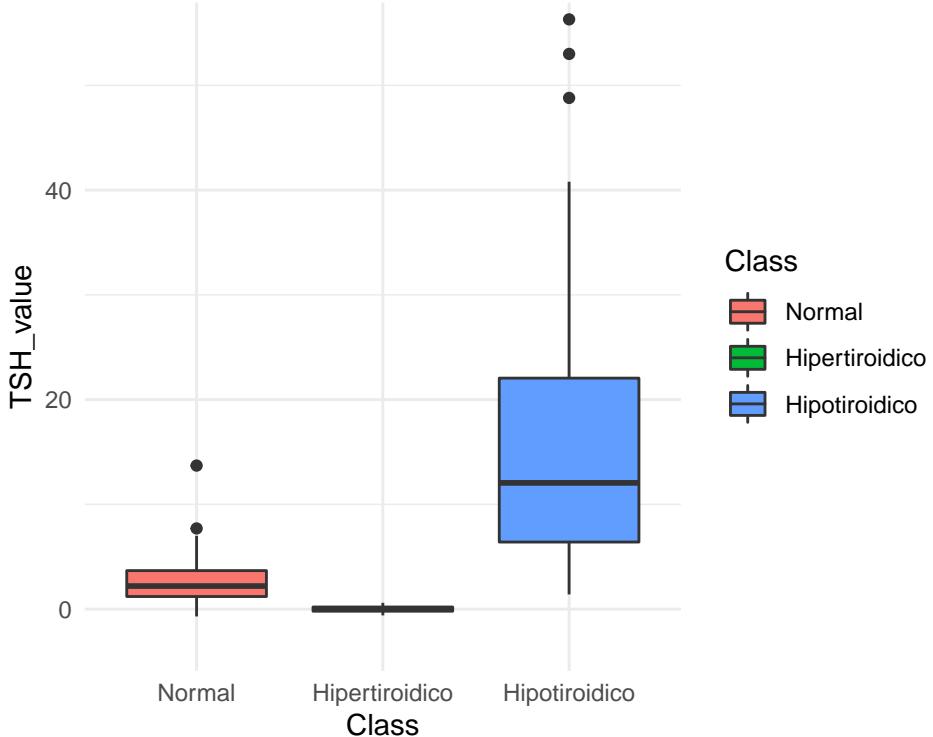


Figura 14: Niveles de TSH por clase

De nuevo, abordaremos estas cuestiones más adelante.

1.2.2. Balanceo de clases

Si atendemos a la variable *Class*, podemos observar un claro desbalanceo de clases, con un total de **150** observaciones para la *Clase 1* frente a **35** para la *Clase 2* y **30** para la *Clase 3* (ver Figura 15). Debido que el empleo de *oversampling* y/o *undersampling* escapa de los objetivos de la asignatura en la que se enmarca este trabajo, nos conformaremos con realizar una división entre train y test que respete las proporciones de cada una de las clases.

1.2.3. Distribución de los datos

Estudiemos la distribución de los datos a partir de los boxplots e histogramas de las Figuras 16 y 17, respectivamente.

- De los boxplots observamos que las variables *Thyroidstimulating* y *TSH_value* siguen una distribución muy similar, lo cual reafirma la teoría de que representan información similar sobre la hormona TSH, aunque tal vez medida de forma distinta. A su vez, podemos afirmar que las observaciones que superan el tercer cuartil deberán corresponderse con casos de hipotiroidismo.
- Por otro lado, los histogramas revelan un desplazamiento en la distribución de las variables referentes al TSH hacia la izquierda, algo comprensible ahora que sabemos que el índice de TSH de una persona sin tiroides se encuentra entre 0.4 y 4, siendo casos con TSH superiores sinónimos de hipotiroidismo. Esto también nos revela que la desviación con respecto a la media de los primeros es mayor, como podemos ver en la Tabla 8, en relación con los hipertíroidicos.

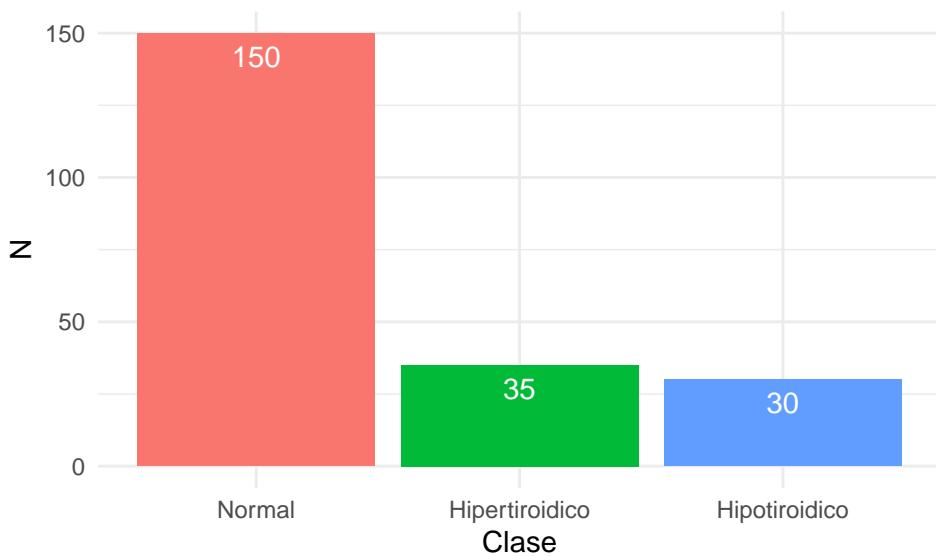


Figura 15: Frecuencia de cada una de las clases

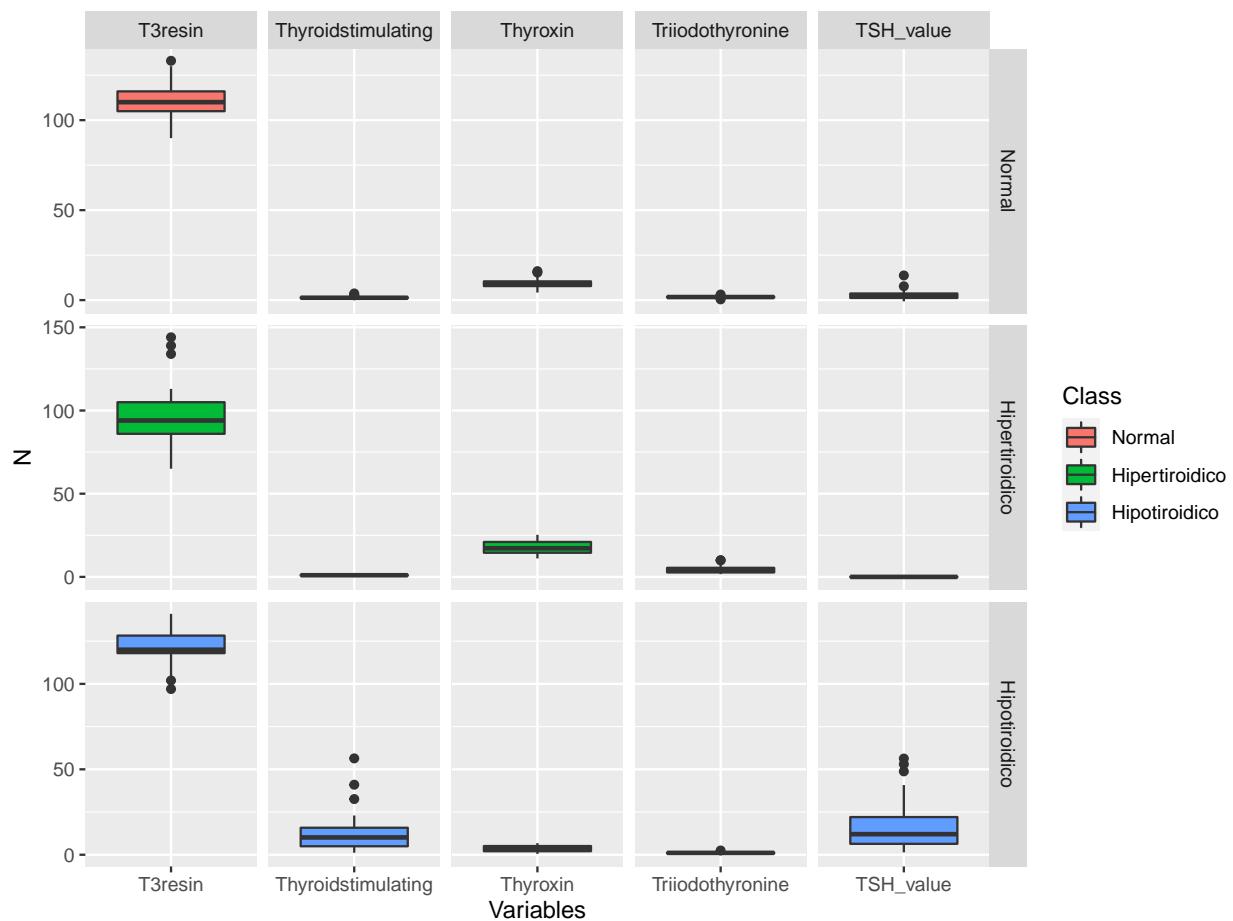


Figura 16: Distribución de los datos: boxplots

Tabla 8: Desviación típica de los niveles de TSH por clase

Clase	TSH SD
Normal	1.9754675
Hipertiroidico	0.2698584
Hipotiroidico	15.5062909

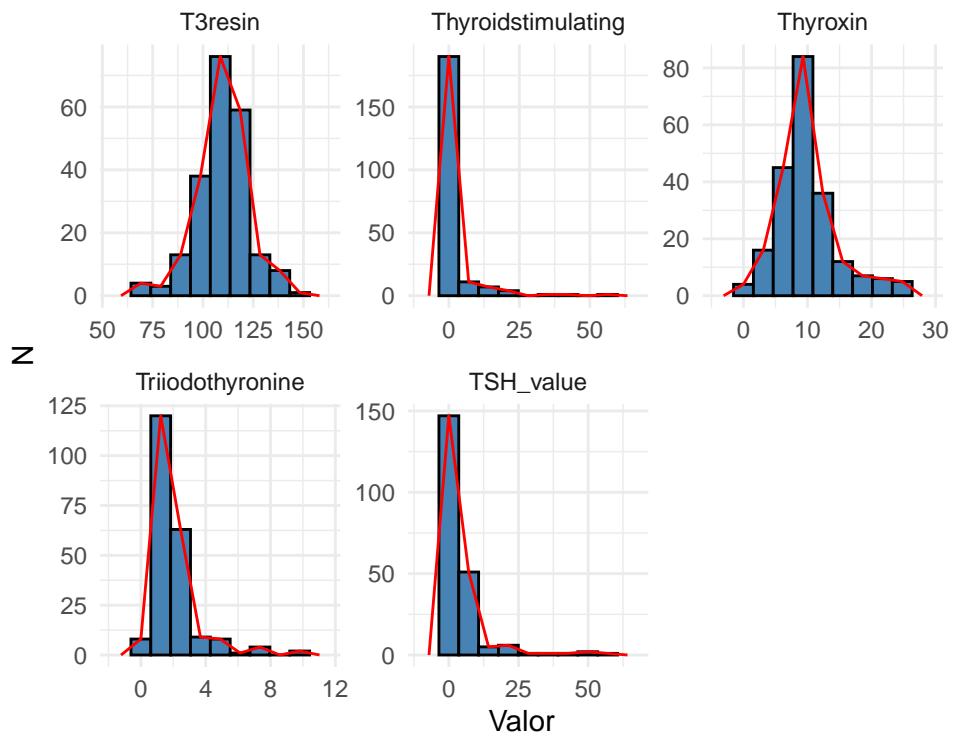


Figura 17: Distribución de los datos: histogramas

Tabla 9: Asimetría de las variables

	skewness
T3resin	-0.5023979
Thyroxin	1.0248080
Triiodothyronine	2.9873885
Thyroidstimulating	5.4538899
TSH_value	4.2436496

Tabla 10: Curtosis de las variables

	kurtosis
T3resin	4.528494
Thyroxin	4.568800
Triiodothyronine	14.224752
Thyroidstimulating	38.970422
TSH_value	23.261734

Finalmente, los valores de *skewness* y *kurtosis* son los representados en las Tablas 9 y 10, complementando la información reflejada en los gráficos. Podemos ver que las variables con menor asimetría son *T3resin* y *Thyroxin*, así como aquellas que menor curtosis presentan, lo cual las convierte en candidatas a ser consideradas normales.

Vista la distribución de las variables, estudiemos pues su normalidad.

1.2.4. Normalidad

Tras aplicar el test de *Shapiro-Wilk* para comprobar la normalidad de nuestras variables (ver Tabla 11), observamos que para ningun p-value se cumple $p\text{-value} > 0,05$, por lo que no podemos asumir la normalidad de ninguna de nuestras variables.

- Con el fin de mejorar la normalidad de nuestros datos, se probó a aplicar las transformaciones de *Box-Cox* y *Yeo-Johnson* en el preprocesamiento de los datos. Los resultados no dieron lugar a cambios significativos con respecto a la normalidad de los datos (ver Tablas 12 y 13), por lo que se optó por no mantener estas transformaciones sobre el conjunto de datos original.

Finalmente, tras observar que ninguna de nuestras variables puede hacerse corresponder con una distribución normal, aplicaremos las operaciones de escalado y normalización de cara a mejorar el desempeño de los algoritmos de clasificación a emplear. Los resultados se reflejan en los histogramas de la Figura 18.

Tabla 11: Normalidad de las variables

vars	statistic	p_value	sample
T3resin	0.9672727	7.024721e-05	215
Thyroxin	0.9266804	7.140882e-09	215
Triiodothyronine	0.6933600	1.427736e-19	215
Thyroidstimulating	0.3582617	6.345000e-27	215
TSH_value	0.4928164	1.909088e-24	215

Tabla 12: Normalidad de las variables tras aplicar Box-Cox

vars	statistic	p_value	sample
T3resin	0.9803607	4.294244e-03	215
Thyroxin	0.9697516	1.442548e-04	215
Triiodothyronine	0.9430391	1.819066e-07	215
Thyroidstimulating	0.8875871	1.398945e-11	215
TSH_value	0.4928164	1.909088e-24	215

Tabla 13: Normalidad de las variables tras aplicar Yeo-Johnson

vars	statistic	p_value	sample
T3resin	0.9803509	4.279723e-03	215
Thyroxin	0.9712386	2.248992e-04	215
Triiodothyronine	0.9637388	2.626407e-05	215
Thyroidstimulating	0.9538355	2.095494e-06	215
TSH_value	0.9849578	2.209906e-02	215

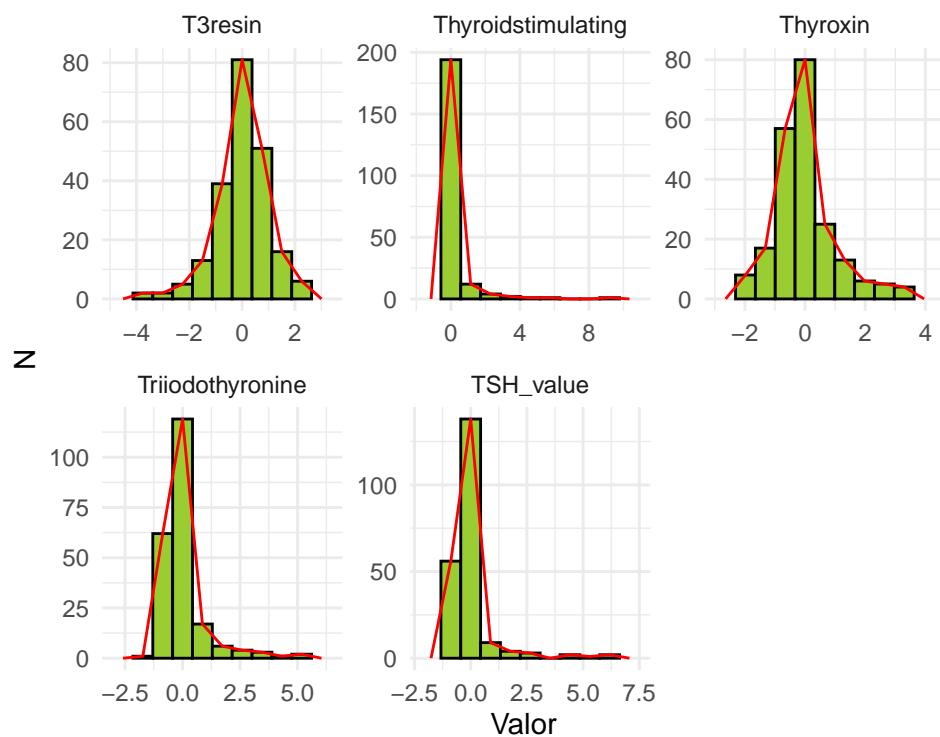


Figura 18: Histogramas de las variables estandarizadas y normalizadas

1.2.5. Correlación

Estudiemos ahora la correlación entre variables, ilustrada en las Figuras 19 y 20. Se trata de un proceso de especial relevancia en el que trataremos de descartar variables correlacionadas que añadan redundancia a nuestros modelos de clasificación.

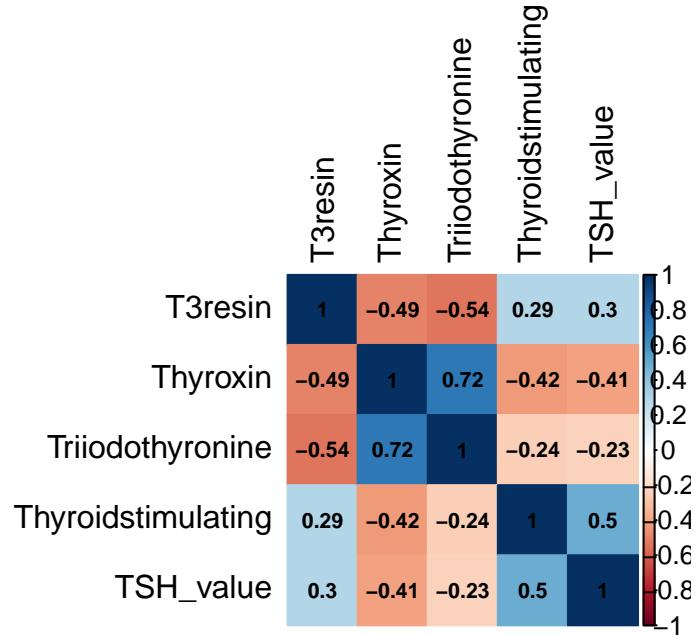


Figura 19: Correlación entre las variables del dataset

De los resultados obtenidos podemos extraer las siguientes conclusiones, las cuales dan respuesta a las hipótesis planteadas al inicio de esta sección:

- Las variables *Triiodothyronine* (T3) y *Thyroxin* (T4) guardan una alta correlación negativa.
- Las relaciones entre *TSH_value* y *Thyroidstimulating* (TSH), así como entre *T3resin* y *Triiodothyronine* (T3), no son tan estrechas como se hipotetizó en un principio, aunque igualmente son significativas y deberán tenerse en cuenta en la parte de clasificación.

1.2.6. Conclusiones

A partir de la exploración del dataset *newthyroid* y de información médica recabada a partir de diferentes fuentes, observamos que las variables relacionadas con los índices de TSH (*TSH_value* y *Thyroidstimulating*) son las más prometedoras a la hora de clasificar a los pacientes, por lo que al menos una de ellas será elegida para llevar a cabo el diagnóstico de tiroides.

Por otro lado, las variables referentes a las hormonas T3 (*Triiodothyronine*) y T4 (*Thyroxin*) son relevantes, aunque en menor medida. Tras observar su correlación, puede que empleando solamente una de ellas la clasificación sea adecuada.

Finalmente, sin conocimiento de un experto, no nos es posible conocer la relevancia de la variable *T3resin* y en qué se diferencia su medición de la de T3, lo que abre la puerta a experimentar cuál de ambos predictores representa mejor la presencia de la hormona T3 en los pacientes.



Figura 20: Relaciones entre las variables del dataset

Tabla 14: Resultados de cada modelo de regresión

	R.squared	Adj.R.squared
Model 1: Max_temperature	0.9455987	0.9455648
Model 2: Min_temperature	0.8775994	0.8775232
Model 3: Dewpoint	0.8036188	0.8034966
Model 4: Sea_level_pressure	0.4122018	0.4118360
Model 5: Visibility	0.2349831	0.2345070

2. Regresión

En esta sección se abordará el conjunto de tareas correspondientes a la parte de **regresión** sobre el dataset *wankara*. Nótese como la única modificación que se ha realizado sobre el dataset en la Sección 1 ha sido la estandarización de las variables (recomendable para *k-NN* y regresión múltiple), aunque se ha decidido trabajar con el dataset original para evitar inconsistencias con los ficheros 5-fcv proporcionados.

Veamos pues, paso a paso, la solución propuesta para cada uno de los enunciados.

2.1. Regresión lineal

Utilizar el algoritmo de regresión lineal simple sobre cada regresor (variable de entrada) para obtener los modelos correspondientes. Si el *dataset_R* asignado incluye más de 5 regresores, seleccione de manera justificada los 5 que considere más relevantes. Una vez obtenidos los modelos, elegir el que considere más adecuado para su conjunto de datos según las medidas de calidad conocidas.

Para abordar esta tarea, ya que contamos con más de 5 regresores, tomaremos como base las conclusiones extraídas del análisis exploratorio de los datos, donde se ha observado que las variables de mayor influencia en la predicción del atributo *Mean_temperature* son: *Max_temperature*, *Min_temperature*, *Dewpoint*, *Sea_level_pressure* y *Visibility*, tal y como se muestra en la Figura 5.

Una vez elegidas las variables a emplear, estudiemos los resultados obtenidos para cada modelo (Figura 21). Para compararlos, tendremos en cuenta los siguientes valores, resumidos en la Tabla 14:

- *Multiple R-squared* o R^2 , que nos muestra en qué proporción la variación de X explica la variación de Y , por lo que cuanto mayor sea, mejor.
- *Adjusted R-squared*, similar al anterior, excepto porque en este caso el valor de R^2 está escalado teniendo en cuenta número de variables del modelo. En este caso la diferencia no será significativa con respecto a R^2 , ya que solamente tenemos una variable.

Así pues, comparando los modelos obtenidos observamos que **el modelo que mejor se ajusta a los datos reales es el *Modelo 1*, el cual toma como X el atributo *Max_temperature* ($R^2 = 94,56\%$)**.

2.2. Regresión lineal múltiple

Utilizar el algoritmo para regresión lineal múltiple. Justificar adecuadamente si el modelo obtenido aporta mejoras respecto al modelo elegido en el paso anterior (en este apartado tenga también en cuenta la consideración de posibles interacciones y no linealidad).

Vamos a seguir un enfoque descendente (*stepwise backward*) para tratar de obtener un modelo de regresión múltiple que mejore los resultados del modelo de regresión simple previamente elegido. Veamos, paso por

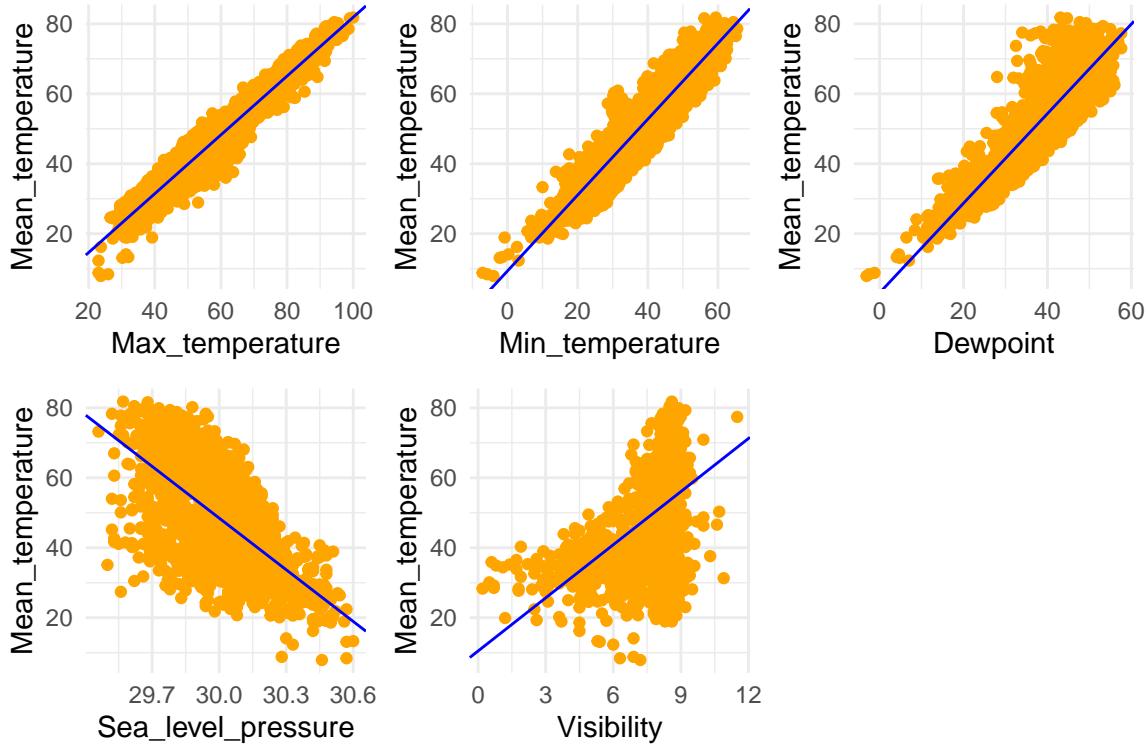


Figura 21: Representación de los modelos de regresión lineal

paso, el procedimiento llevado a cabo:

1. Partimos de un modelo con todas las variables del dataset, del cual observamos un excelente valor para R^2 y $\text{Adj. } R^2$. Podemos ver que con un 98,98 % de R^2 se describe casi a la perfección el comportamiento de la variable *Mean_temperature*.

```
m1 <- lm(Mean_temperature ~ ., data=wankara)
summary(m1)
```

2. Aunque ya hemos obtenido un modelo considerablemente mejor que el de regresión lineal simple, tratemos de mejorarla si es posible. Para ello, vamos a eliminar del conjunto de predictores la variable *Precipitation*, la cual, como ya adelantamos en la sección dedicada al análisis exploratorio de los datos, apenas era relevante. También podemos observar su baja relevancia si observamos el valor de $Pr(> |t|)$, el cuál nos indica que la variable es poco significativa.

```
m2 <- lm(Mean_temperature ~ . - Precipitation, data=wankara)
summary(m2)
```

Así pues, tras eliminar *Precipitation* del conjunto de predictores obtenemos un nuevo modelo en el que apenas se producen variaciones ($R^2 = 98,98\%$), más allá de una mejora de un 0,01 % en el valor del error residual estándar (*Residual standard error*).

3. Tras observar que ahora todas las variables que forman parte del modelo son significativas, una observación de especial relevancia descubierta durante el análisis exploratorio de los datos (ver Sección 1.1.5) es que la distribución de la visibilidad con respecto a la temperatura media podría corresponderse con una función cuadrática. Tratemos pues de añadir no-linealidad al modelo elevando al cuadrado la variable *Visibility*.

```
m3 <- lm(Mean_temperature ~ . - Precipitation + I(Visibility^2), data=wankara)
summary(m3)
```

Como podemos observar, los resultados mejoran aun más: hemos pasado de un $Adj.R^2 = 98,98\%$ a un $Adj.R^2 = 98,99\%$, lo que supone una mejora de un 0,01% y un modelo resultante casi perfecto.

- Partiendo de un modelo tan bueno como base, tratemos de simplificarlo a costa de reducir levemente su precisión. Partiendo primero de las variables menos correlacionadas con *Mean_temperature*, eliminaremos inicialmente la variable *Wind_speed*, obteniendo un 98,95% de R^2 .

```
m4 <- lm(Mean_temperature ~ . - Precipitation - Wind_speed + I(Visibility^2),
          data=wankara)
summary(m4)
```

- Observamos ahora que la variable *Max_wind_speed* deja de ser significativa, lo que nos lleva a descartarla, manteniendo un R^2 del 98,95%. Nos mantenemos aproximadamente en un 98% de R^2 tras haber quitado dos variables, lo cual es bueno porque nuestro modelo se ha simplificado a costa de perder muy poca precisión.

```
m5 <- lm(Mean_temperature ~ . -
          Precipitation - Wind_speed - Max_wind_speed + I(Visibility^2),
          data=wankara)
summary(m5)
```

- Siguiendo el orden de variables menos correlacionadas con la variable a predecir, eliminemos *Standard_pressure* de la lista de candidatas. De nuevo, nos mantenemos alrededor del 98% (98,92%) de R^2 con un modelo mucho más simple.

```
m6 <- lm(Mean_temperature ~ . -
          Precipitation - Wind_speed -
          Max_wind_speed - Standard_pressure + I(Visibility^2),
          data=wankara)
summary(m6)
```

Si siguiésemos el proceso de eliminación de variables, realmente podríamos observar que con sólo la temperatura mínima (*Min_temperature*) y la temperatura máxima (*Max_temperature*) la temperatura media (*Mean_temperature*) se predice muy bien, obteniendo un 98,7% de R^2 .

```
m7 <- lm(Mean_temperature ~ Max_temperature + Min_temperature, data=wankara)
summary(m7)
```

También observamos que añadiendo interacción entre ambas variables los resultados no varían (98,7% de R^2):

```
m8 <- lm(Mean_temperature ~ Max_temperature * Min_temperature, data=wankara)
summary(m8)
```

No obstante, finalmente **nos quedaremos con el modelo obtenido en la iteración 6**, ya que es el único que no desciende del 98,9% de R^2 , lo cual lo convierte un modelo casi perfecto y bastante simple. También debemos tener en cuenta que obtener la temperatura media empleando únicamente las variables *Max_temperature* y *Min_temperature* es algo que cualquier persona podría conseguir aplicando la media y obteniendo un error aproximadamente similar al del modelo. Sin embargo, es cuando otras variables entran en juego (visibilidad, presión, etc.) cuando de verdad el modelo cobra valor, de ahí que se haya decidido por no eliminar todas estas variables y detener aquí el proceso.

Tabla 15: MSE para k-NN y regresión lineal múltiple sobre train y test (5-fcv)

	MSE.train	MSE.test
k-NN: todas las variables	2.733698	6.773059
k-NN: mejores variables lm	1.470406	3.310527
k-NN: max y min temp.	1.661257	2.970584
lm: mejores variables	2.575251	2.619007

Tabla 16: Comparativa entre LM (R+) y k-NN (R-) aplicando el test de Wilcoxon

	R+	R-	p-value
LM vs. k-NN	78	93	0.7660294

2.3. Regresión mediante *k-NN*

Aplicar el algoritmo k-NN para regresión no paramétrica.

En este apartado aplicaremos el algoritmo de *k-NN* para regresión utilizando las particiones 5-fcv proporcionadas. Tras esto, comprobaremos si los resultados mejoran con las mismas particiones en comparación con el mejor modelo de regresión lineal múltiple obtenido en el apartado anterior.

Concretamente, vamos a evaluar *k-NN* de tres formas diferentes:

- Utilizando todas las variables del dataset.
- Empleando las variables del mejor modelo obtenido para regresión lineal múltiple.
- Utilizando únicamente las variables *Max_temperature* y *Min_temperature*.

De los resultados resumidos en la Tabla 15 observamos que el modelo con menor error cuadrático medio (*MSE*) sobre los datos de *test* es el modelo de regresión lineal que emplea el conjunto de variables seleccionado en el apartado anterior. Por otro lado, si estudiamos los resultados de *k-NN* observamos que *k-NN* con las mejores variables tiene un mejor rendimiento con los datos de *train* pero no generaliza bien para los datos de *test*. También podemos ver que el *k-NN* con mejores resultados para los datos de *test* es el que trabaja únicamente con las variables *Max_temperature* y *Min_temperature*. Finalmente, *k-NN* utilizando todas las variables es el caso en el que peor rendimiento obtenemos tanto para *train* como para *test*.

2.4. Comparativa

Comparar los resultados de los dos algoritmos de regresión múltiple entre sí, y adicionalmente mediante comparativas múltiples con un tercero (el modelo de regresión M5', cuyos resultados ya están incluidos en las tablas de resultados disponibles).

Primero compararemos regresión lineal múltiple (*LM*) con *k-NN* aplicando el test de Wilcoxon. Los resultados están reflejados en la Tabla 16.

- De la tabla podemos observar que **no hay diferencias significativas entre ambos modelos** (sólo hay un $(1 - 0,7660) * 100 = 23,4\%$ de confianza en que sean distintos).

Ahora compararemos *LM*, *k-NN* y *M5'* aplicando Friedman y Holm como post-hoc (Tabla 17).

- Tras aplicar el test de Friedman ($\tilde{\chi}^2 = 8,4444$, $p = 0,01467$) observamos que al menos uno de los métodos de regresión difiere del resto (rechazamos H_0), por lo que aplicamos Holm como post-hoc.

Tabla 17: Comparativa entre LM, k-NN y M5' aplicando Holm

	LM	k-NN
k-NN	0.5798416	NA
M5'	0.0805435	0.1077118

Tabla 18: Resultados obtenidos tras aplicar k-NN con 10-fcv

	Acc. CV train	Acc. CV test
KNN (K=1)	1.0000000	0.9627706
KNN (K=3)	0.9731264	0.9538961
KNN (K=5)	0.9627878	0.9443723
KNN (K=7)	0.9498718	0.9348485

- El post-hoc nos muestra que existen diferencias significativas a favor de M5' ($M5 \text{ vs } LM = 0,081$ y $M5 \text{ vs } KNN = 0,108$, con aproximadamente un 90 % de confianza), mientras que los otros se pueden considerar equivalentes.

3. Clasificación

Esta última sección estará dedicada a las tareas de clasificación propuestas sobre el dataset *newthyroid*. Toda la información relativa al estudio de las variables a emplear, así como de las clases a identificar se encuentra detallada en la Sección 1, dedicada al análisis exploratorio de los datos.

Veamos pues, paso por paso, la resolución de los diferentes problemas planteados.

3.1. Clasificación mediante *k-NN*

Utilizar el algoritmo k-NN probando con diferentes valores de k . Elegir el que considere más adecuado para su conjunto de datos. Analice qué ocurre en los valores de precisión en training y test con los diferentes valores de k .

Para abordar esta tarea, se ha aplicado *k-NN* con todas las variables y comparado los resultados obtenidos sobre el conjunto de datos 10-fcv proporcionado empleando diferentes valores de k , concretamente: 1, 3, 5 y 7. Los resultados se muestran en la Tabla 18 y en la Figura 22, donde podemos observar que los niveles de *accuracy* para *train* y *test* disminuyen a medida que aumentamos el valor de k . Por tanto, el mejor valor de k para este caso es $k = 1$ ($Accuracy_{train} = 1$; $Accuracy_{test} = 0,96$).

3.2. Clasificación mediante *LDA*

Utilizar el algoritmo LDA para clasificar. No olvide comprobar las asunciones.

Abordemos el mismo problema de clasificación empleando LDA. Antes de poderlo aplicar, es necesario tener en cuenta los siguientes requisitos:

1. **Las observaciones provienen de una muestra aleatoria:** asumiremos que es así.

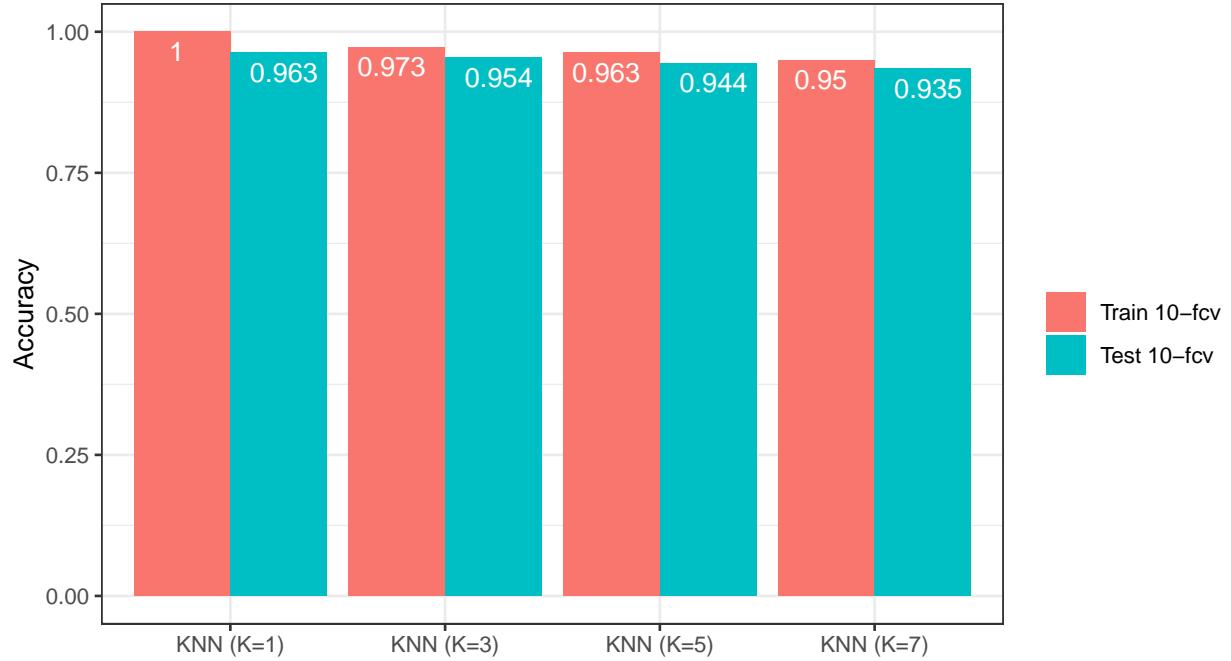


Figura 22: accuracy de k-NN en train y test para diferentes valores de k

Tabla 19: Varianzas por variable y clase

	Normal	Hipertoroidico	Hipotiroidico
T3resin	65.6072036	352.0336134	122.2862069
Thyroxin	4.1988049	17.3131429	3.0848276
Triiodothyronine	0.2259244	5.0812269	0.3086092
Thyroidstimulating	0.2475727	0.1602017	153.4478621
TSH_value	3.9024720	0.0728235	240.4450575

2. **Todo predictor tiene una distribución normal:** como pudimos ver en la Sección 1.2.4, no podemos asumir la normalidad de ninguno de los predictores de nuestro dataset (ni siquiera tras aplicar transformaciones como *Box-Cox* y *Yeo-Johnson*). Esto puede empeorar de forma notable los resultados obtenidos; no obstante, trataremos de aplicar LDA pasando por alto este requisito, aunque siendo conscientes de que no conseguiremos un clasificador óptimo⁶.
3. **Todas las clases comparten la misma matriz de covarianza (la matriz de covarianzas es homogénea en todos los grupos):** de los resultados expresados en las Tablas 19 y 20, observamos que, tras aplicar el Test de Levene, no podemos asumir la homeogeneidad de las varianzas. De nuevo, aplicaremos LDA a pesar de que no se cumpla esta asunción, a costa de obtener peores resultados.

Antes de aplicar LDA, también se puede considerar como preconditionamiento que los predictores empleados sean independientes y que contemos con más observaciones que predictores. También es aconsejable que los predictores estén centrados y escalados, eliminando aquellos con varianza cercana a 0.

⁶La condición de normalidad supone una pérdida de precisión al aplicar LDA pero no hace inviable este método de clasificación. Véase: Li, Tao & Zhu, Shenghuo & Ogihara, Mitsunori. (2006). Using discriminant analysis for multi-class classification: An experimental investigation. Knowledge and Information Systems. 10. 453-472. 10.1007/s10115-006-0013-y.

Tabla 20: Resultados tras aplicar el Test de Levene

Variable	p.value
T3resin	7.17426864138572e-07
Thyroxin	5.57893956250375e-10
Triiodothyronine	4.49827176856079e-20
Thyroidstimulating	2.87843705709032e-20
TSH_value	6.40620165275579e-23

Tabla 21: Varianzas de los predictores

	varianza
T3resin	172.802782
Thyroxin	22.065212
Triiodothyronine	2.014941
Thyroidstimulating	37.430299
TSH_value	65.133270

- Con respecto a la independencia de los predictores, probaremos a aplicar LDA considerando tanto todas las variables como tan sólo un subconjunto de ellas con una baja correlación entre sí. De esta forma, podremos estudiar cómo afecta la independencia de las variables a los resultados obtenidos.
- Contamos con muchas más observaciones (215) que predictores (6).
- El escalado y la centralidad de los predictores ya se tuvo en cuenta al abordar el análisis exploratorio de los datos (ver Sección 1). Se aplicará dicho preprocesamiento antes de utilizar el modelo sobre los 10-fcv.
- La varianza de ninguno de los predictores es cercana a 0, tal y como se muestra en la Tabla 21.

Estudiados los prerrequisitos, apliquemos LDA. De los resultados resumidos en la Tabla 22 y la Figura 23 podemos observar que el valor de accuracy obtenido es superior tanto para *train* como para *test* en el caso en el que eliminamos del conjunto de predictores aquellas variables que dan lugar a correlaciones indeseadas (en este caso, la variable *Thyroidstimulating*, muy relacionada con *TSH_value*, como ya vimos en la sección 1.2.5). De esta forma, logramos mejorar ligeramente el modelo.

3.3. Clasificación mediante QDA

Utilizar el algoritmo QDA para clasificar. No olvide comprobar las asunciones.

Abordemos ahora el mismo problema utilizando QDA. Se trata de un tipo de clasificador que funciona mejor que LDA cuando las varianzas entre las diferentes clases son significativamente diferentes, como es el caso (véase la Tabla 19), por lo que a priori podemos esperar mejores resultados. Recordemos las asunciones previas a aplicar QDA:

Tabla 22: Resultados obtenidos tras aplicar LDA con 10-fcv

	Acc. CV train	Acc. CV test
LDA: todas las variables	0.9183457	0.9164502
LDA: variables seleccionadas	0.9266145	0.9209957

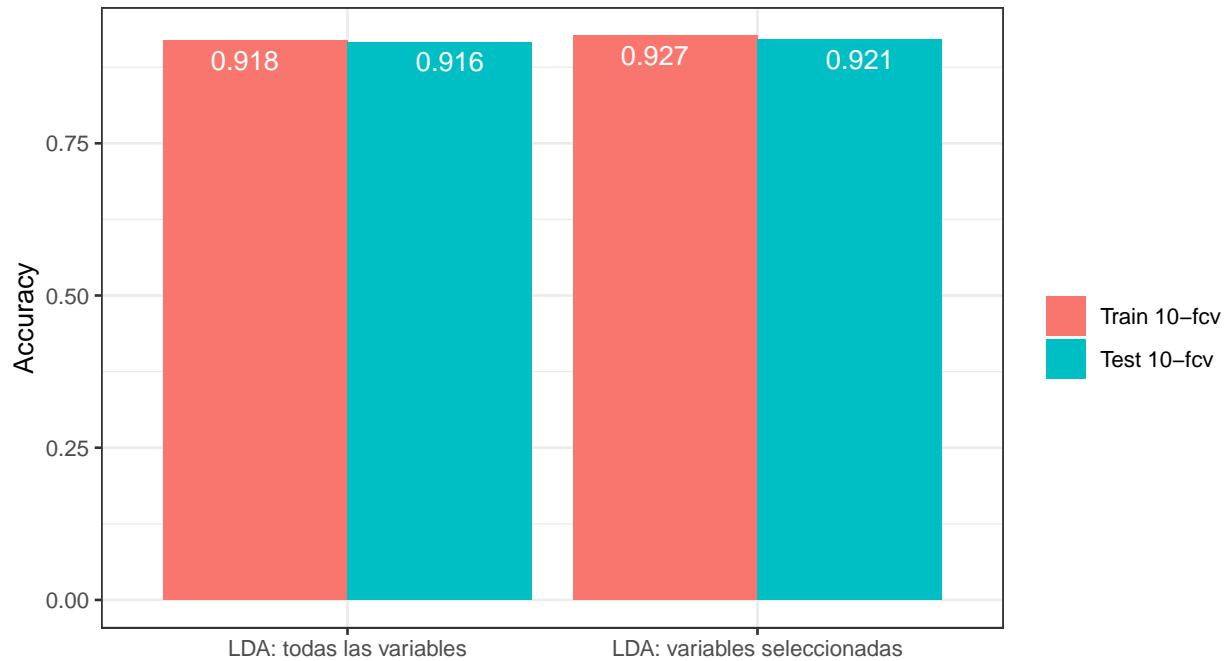


Figura 23: accuracy de LDA en train y test

Tabla 23: Resultados obtenidos tras aplicar QDA con 10-fcv

	Acc. CV train	Acc. CV test
QDA: todas las variables	0.9700283	0.9629870
QDA: variables seleccionadas	0.9741574	0.9766234

- **El número de predictores es menor que el número de observaciones para cada clase:** esta condición se cumple.
- **Los predictores de cada clase deben ser independientes.** Al igual que en el caso anterior, probaremos el algoritmo con todos los predictores y con un subconjunto de variables linealmente independientes para comparar la mejora.

Como podemos observar en la Tabla 23 y la Figura 24, de nuevo el uso de un subconjunto de variables independientes mejora la precisión del modelo frente al caso en que utilizamos todas las variables. También cabe destacar que el valor de *accuracy* obtenido empleando QDA es notablemente mejor que el de LDA: si comparamos ambos métodos para el caso en el que utilizamos las mejores variables identificadas, QDA ofrece una precisión del 97,7% para el conjunto de *test*, frente al 92,1% obtenido por LDA. Esta mejora podría deberse a que QDA es independiente de la condición de homocedasticidad que LDA requiere y que en este caso no se cumple.

3.4. Comparativa

Comparar los resultados de los tres algoritmos.

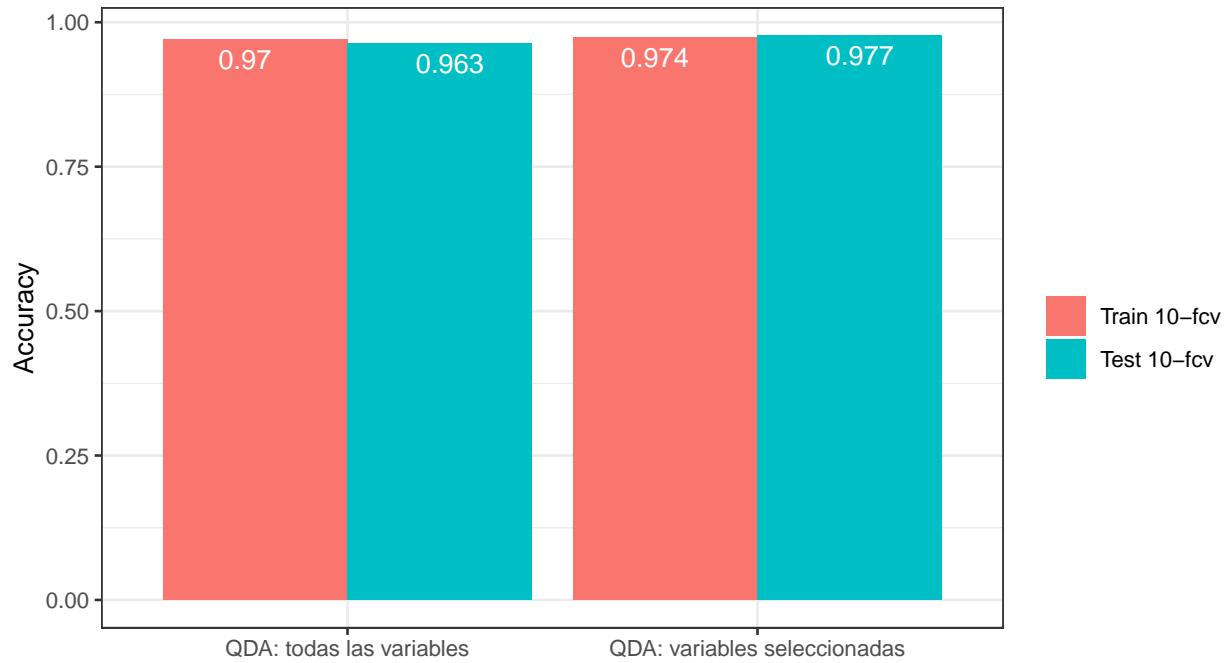


Figura 24: accuracy de QDA en train y test

Una vez estudiados los tres algoritmos de clasificación, compararemos sus resultados. Tras aplicar el test de Friedman ($\tilde{\chi}^2 = 0,7$, $p = 0,7047$) no podemos rechazar la hipótesis nula y, por tanto, ninguno de los métodos de clasificación empleados (*k*-NN, LDA, QDA) se considera significativamente diferente del resto.

Así damos por concluida esta sección, habiendo estudiado en detalle diferentes métodos de clasificación sobre el dataset *newthyroid*. Como conclusión, podemos observar que una buena elección de variables y tener en cuenta las asunciones necesarias para cada tipo de algoritmo hará que generalmente mejore el rendimiento obtenido. Por ejemplo, para *LDA* y *QDA* hemos visto cómo eliminar variables correlacionadas mejora el rendimiento de los clasificadores. Finalmente, para el caso de *k*-NN hemos evaluado su rendimiento para diferentes valores de *k*, observando su notable influencia en la precisión del modelo.