Una exploración sobre algoritmos de clasificación en Machine Learning: Competición en Kaggle - Vol.2

Se presenta esta memoria como requisito para obtener

la titulación del

Máster en Matemáticas

redactada por

Pablo Domínguez Balbás

Bajo la supervisión como tutor de

Dr. Víctor Blanco Izquierdo

y como cotutor de

Dr. Pedro A. García Sánchez

UNIVERSIDAD DE GRANADA

Universidad de Granada
10 de julio de 2022

Análisis y modelado de datos

Pablo Domínguez

2022-07-08

Índice

Índice de tablas

Índice de figuras

Planteamiento del problema a abordar

En el mundo actual existe la necesidad de diversificar y divulgar el conocimiento sobre la ciencia del dato. Es por este motivo por el que entidades como Kaggle se encargan de recopilar bases de datos y proponer problemas relativos a los mismos, con vistas de motivar la resolución de problemas de *aprendizaje automático* cada vez más complicados.

Entre estos, nos encontramos con un conjunto de datos obtenidos a partir de mediciones meteorológicas realizadas por el gobierno de Australia[1]. Estos datos, recogidos en distintas localidades, se han capturado realizando mediciones diarias de temperatura, lluvia, evaporación, sol, viento, humedad etc.

En la referencia mencionada advierten que el control de calidad aplicado a la captura de estos datos ha sido limitado, por lo que es posible que existan imprecisiones debidas a datos faltantes, valores acumulados tras varios datos faltantes o errores de varios tipos. Es por este motivo que empezaremos nuestro estudio realizando una revisión de la calidad y estructura del dato. Tras este proceso, construiremos una serie de variables que transformarán el problema y la estructura de datos para que puedan aplicarse los modelos de clasificación supervisada planteados.

Partiendo de la base de datos procesada, la segmentaremos para aplicar varios modelados diferentes por cada zona climática[4]. Finalmente, compararemos los modelos, los ensamblaremos y presentaremos unos resultados de la precisión del modelo final.

Con esta aplicación práctica de los modelos teóricos abordados en el capítulo anterior buscamos reflejar la capacidad de herramientas matemáticas abstractas a la hora de resolver situaciones que pueden tener un gran beneficio en varios ámbitos, tales como sociales, económicos o medioambientales.

Para ello, nos basaremos en las pautas definidas en Everitt et al. (2011) a la hora de abordar un problema de clasificación supervisada, esto es:

1. Definir un conjunto de datos sobre los que modelar el problema de clasificación.
2. Seleccionar un conjunto de *variables predictoras* relevante para el problema.
3. Tratar incorrecciones o valores faltantes.
4. Estandarizar variables.
5. Creación de nuevas variables.
6. Modelado del problema de clasificación con diferentes algoritmos.
7. Presentar, visualizar, comparar e interpretar los resultados.
8. Conclusiones.

Origen de los datos y variable objetivo

El buró de metereología australiano coordina una serie de estaciones metereológicas locales repartidas a lo largo del territorio. De esta manera, recopila y reporta datos sobre mediciones meteorológicas. En nuestro caso, tenemos información de 49 ciudades repartida a lo largo de unos 8 años.

Tabla 1: Muestra de algunas filas y columnas del conjunto de datos

Date	Location	MinTemp	MaxTemp	Rainfall	—	Temp9am	Temp3pm	RainToday	RainTomorrow
2008-12-01	Albury	13.4	22.9	0.6	—	16.9	21.8	No	No
2008-12-02	Albury	7.4	25.1	0.0	—	17.2	24.3	No	No
2008-12-03	Albury	12.9	25.7	0.0	—	21.0	23.2	No	No
2008-12-04	Albury	9.2	28.0	0.0	—	18.1	26.5	No	No
2008-12-05	Albury	17.5	32.3	1.0	—	17.8	29.7	No	No
2008-12-06	Albury	14.6	29.7	0.2	—	20.6	28.9	No	No

[1]Notes about Daily Weather Observations - Australian Goverment (2004)

Destacar que hemos importado las variables tipo *string* como *factor*. En la tabla 2, presentamos el tipo de cada variable.

Tabla 2: Tipo de datos de las variables

(a) Tipos de las once primeras variables

	Data_types
Date	Date
Location	factor
MinTemp	numeric
MaxTemp	numeric
Rainfall	numeric
Evaporation	numeric
Sunshine	numeric
WindGustDir	factor
WindGustSpeed	numeric
WindDir9am	factor
WindDir3pm	factor

(b) Tipos de las doce variables restantes

	Data_types
WindSpeed9am	numeric
WindSpeed3pm	numeric
Humidity9am	numeric
Humidity3pm	numeric
Pressure9am	numeric
Pressure3pm	numeric
Cloud9am	numeric
Cloud3pm	numeric
Temp9am	numeric
Temp3pm	numeric
RainToday	factor
RainTomorrow	factor

De esta manera, según se refleja en Australian Goverment (2004), el conjunto de datos con el que vamos a trabajar cuenta con las siguientes variables.

- *Date*: Fecha en la que se realizó la medición, en formato *AAAA-MM-DD*.
- *Location*: Localización donde se realizó la medición.
- *MinTemp*: Temperatura mínima alcanzada, medida en grados celsius [2].
- *MaxTemp*: Temperatura máxima alcanzada, medida en grados celsius[2].
- *Rainfall*: Cantidad total de precipitación, medida en milímetros[2].
- *Evaporation*: Evaporación en milímetros sobre un *tanque evaporimétrico clase "A"*[2].
- *Sunshine*: Tiempo en horas de alto nivel de luminosidad solar[3].
- *WindGustDir*: Dirección general del viento racheado a lo largo del día[3], tomando una de las 16 direcciones posibles del viento.
- *WindGustSpeed*: Velocidad global del viento racheado a lo largo del día[3], medido en kilómetros por hora.
- *WindDir9am*, *WindDir3pm*: Dirección promedio del viento en los 10 minutos previos a la hora indicada en cada variable.
- *WindSpeed9am*, *WindSpeed3pm*: Velocidad promedio del viento en los 10 minutos previos a la hora indicada en cada variable, medida en kilómetros por hora.
- *Humidity9am*, *Humidity3pm*: Humedad relativa, medida en tanto porciento, a las horas indicadas en cada variable.
- *Pressure9am*, *Pressure3pm*: Presión atmosférica medida a nivel del mar a las horas indicadas en cada variable, usando el hectopascal como unidad de medida.
- *Cloud9am*, *Cloud3pm*: Fracción del cielo cubierta por nubes a las horas indicadas, medida en *octas de cielo*.
- *Temp9am*, *Temp3pm*: Temperatura en grados celsius medida a las horas indicadas.
- *RainToday*, *RainTomorrow*: Variables binarias indicando si ha habido o no lluvia ese día.

De este modo, *RainTomorrow* será nuestra variable objetivo en los clasificadores. Con ello, desarrollaremos estrategias buscando que los modelos puedan decidir lo mejor posible si, dadas las observaciones de un día e información agregada de días anteriores, lloverá o no al día siguiente.

[2]En un rango de 24 horas desde las 9am.
[3]En un rango de 24 horas desde las 12am.

Información de los datos y filtrado de datos

A continuación, buscamos comprobar si la serie temporal está completa y sin días repetidos para cada valor de la variable *Location*. En la tabla 3, se muestran las variables conteniendo la siguiente información.

- *min_fec*, *max_fec*: Fecha donde comienzan y terminan los datos en la serie temporal.
- *obs*: Número de observaciones total en cada serie temporal.
- *rep*: Booleano que representa si existen fechas repetidas dentro de la serie temporal, esto es, *FALSE* indica ausencia de fechas repetidas.
- *n_diff_dates*: Cantidad de fechas faltantes dentro de la serie temporal.
- *top_diff_date*: Última fecha faltante en la serie temporal.
- *range_free*: Cantidad de datos sin fechas faltates medidos hacia atrás desde el último registro. Es decir, longitud de la serie temporal continua más reciente.
- *range*: Rango, medido en dias, de la serie temporal. Esto es, número de días entre *max_fec* y *min_fec*.

Tabla 3: Continuidad de las series temporales

	min_fec	max_fec	obs	rep	n_diff_dates	top_diff_date	range_free	range
Adelaide	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
Albany	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Albury	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
AliceSprings	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
BadgerysCreek	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Ballarat	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Bendigo	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Brisbane	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
Cairns	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Canberra	2007-11-01	2017-06-25	3436	FALSE	89	2013-02-28	1579	3525 days
Cobar	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
CoffsHarbour	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Dartmoor	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Darwin	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
GoldCoast	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Hobart	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
Katherine	2013-03-01	2017-06-25	1578	FALSE	0	NA	NA	1578 days
Launceston	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Melbourne	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
MelbourneAirport	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Mildura	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days

Moree	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
MountGambier	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
MountGinini	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Newcastle	2008-12-01	2017-06-24	3039	FALSE	89	2013-02-28	1578	3128 days
Nhil	2013-03-01	2017-06-25	1578	FALSE	0	NA	NA	1578 days
NorahHead	2009-01-01	2017-06-25	3004	FALSE	94	2013-12-31	1273	3098 days
NorfolkIsland	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Nuriootpa	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
PearceRAAF	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Penrith	2008-12-01	2017-06-25	3039	FALSE	90	2013-02-28	1579	3129 days
Perth	2008-07-01	2017-06-25	3193	FALSE	89	2013-02-28	1579	3282 days
PerthAirport	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Portland	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Richmond	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Sale	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
SalmonGums	2009-01-01	2017-06-25	3001	FALSE	97	2013-02-28	1579	3098 days
Sydney	2008-02-01	2017-06-25	3344	FALSE	89	2013-02-28	1579	3433 days
SydneyAirport	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Townsville	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Tuggeranong	2008-12-01	2017-06-25	3039	FALSE	90	2013-02-28	1579	3129 days
Uluru	2013-03-01	2017-06-25	1578	FALSE	0	NA	NA	1578 days
WaggaWagga	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Walpole	2009-01-01	2017-06-25	3006	FALSE	92	2013-02-28	1579	3098 days
Watsonia	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Williamtown	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Witchcliffe	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days
Wollongong	2008-12-01	2017-06-25	3040	FALSE	89	2013-02-28	1579	3129 days
Woomera	2009-01-01	2017-06-25	3009	FALSE	89	2013-02-28	1579	3098 days

Como podemos observar en el análisis anterior, vemos que la localización *NorahHead* tiene el último hueco en su serie temporal el *2013-12-31*. Es por este motivo que vamos a filtrar a partir de este valor con el fin de asegurarnos que no existan fechas faltantes dentro de las series temporales. Se puede observar además que el resto de localizaciones tienen como último salto en su serie temporal la fecha *2013-02-28*, y que las localizaciones de *Katherine*, *Nhil* y *Uluru* comienzan su serie temporal el *2013-03-01*. Esto es, que cuando comenzó a haber registros en estas tres localizaciones dejó de haber errores de grabado de datos en la serie temporal en el resto de localizaciones (salvo en *NorahHead*). Este suceso dentro del proceso de grabación de datos podría ser indicativo de una reestructuración de los mecanismos de captación de los datos climáticos por parte del buró de meteorología australiano.

Finalmente, filtrando desde el *2013-12-31* hasta el *2017-06-24*, que es el último dato registrado en todas las series, se puede observar en la tabla 4 que ya no existen discontinuidades en ninguna serie temporal.

Tabla 4: Muestra de series temporales desde el 2014-01-01 hasta el 2017-06-24

	min_fec	max_fec	obs	rep	n_diff_dates	top_diff_date	range_free	range
Adelaide	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Albany	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Albury	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
AliceSprings	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
BadgerysCreek	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Ballarat	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Bendigo	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Brisbane	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Cairns	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Canberra	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
Cobar	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days
CoffsHarbour	2014-01-01	2017-06-24	1271	FALSE	0	NA	NA	1271 days

Agrupación por zona climática

Con el objetivo de modelar con mayor precisión los datos, vamos a realizar una segmentación de la información asignando la zona climática[4] a la que pertenece cada localización. Estas zonas, representadas en la figura 1, están descritas de la siguiente manera.

- *Zona 1*: Veranos húmedos y cálidos, con inviernos cálidos.
- *Zona 2*: Veranos húmedos y templados, con inviernos suaves.
- *Zona 3*: Veranos secos y cálidos, con inviernos cálidos.
- *Zona 4*: Veranos secos y cálidos, con inviernos frescos.
- *Zona 5*: Clima templado.
- *Zona 6*: Clima suave.
- *Zona 7*: Clima fresco.
- *Zona 8*: Clima alpino.

Una estrategia alternativa que planteamos como propuesta de mejora en este punto consistiría en realizar un clustering no supervisado, de manera que se segmenten las localizaciones en 8 categorías no equitativas. A continuación, compararíamos esta categorización no supervisada contra la categorización por zona climática para comprobar la similitud de ambas asignaciones.

En cualquier caso, en el planteamiento que proponemos hemos eliminado aquellas localizaciones donde no ha sido posible determinar la zona climática a la que pertenecen. En concreto, se han eliminado *NorfolkIsland*, *Portland*, *Richmond*, *Walpole* y *Williamtown*. A continuación se presentan en la tabla 5 la cantidad total de observaciones junto con la cantidad de ciudades por cada zona climática.

Llegados a este punto, estamos en disposición de crear un conjunto de datos para cada zona climática, a los cuales someteremos a una limpieza y procesado de datos. Es importante aplicar este procedimiento por separado para cada zona climática ya que, entre otras técnicas, imputaremos datos faltantes. Por tanto, es necesario que la información con la que sustituimos estos valores sea coherente con la subdivisión por zonas climáticas del conjunto de datos original. Finalmente, tenemos el conjunto de datos segmentado en 8 zonas disjuntas, que abordaremos y modelaremos en paralelo.

Calidad y limpieza de los datos

A continuación, nos disponemos a tratar los aspectos relativos a la calidad del dato. Esto es, tratamiento de valores faltates o *missings*, valores atípicos u *outliers*, normalización de los datos y balanceo de los datos.

[4]Desarolladas en el National Construction Code por el Buró de Metereología del Gobierno de Australia.

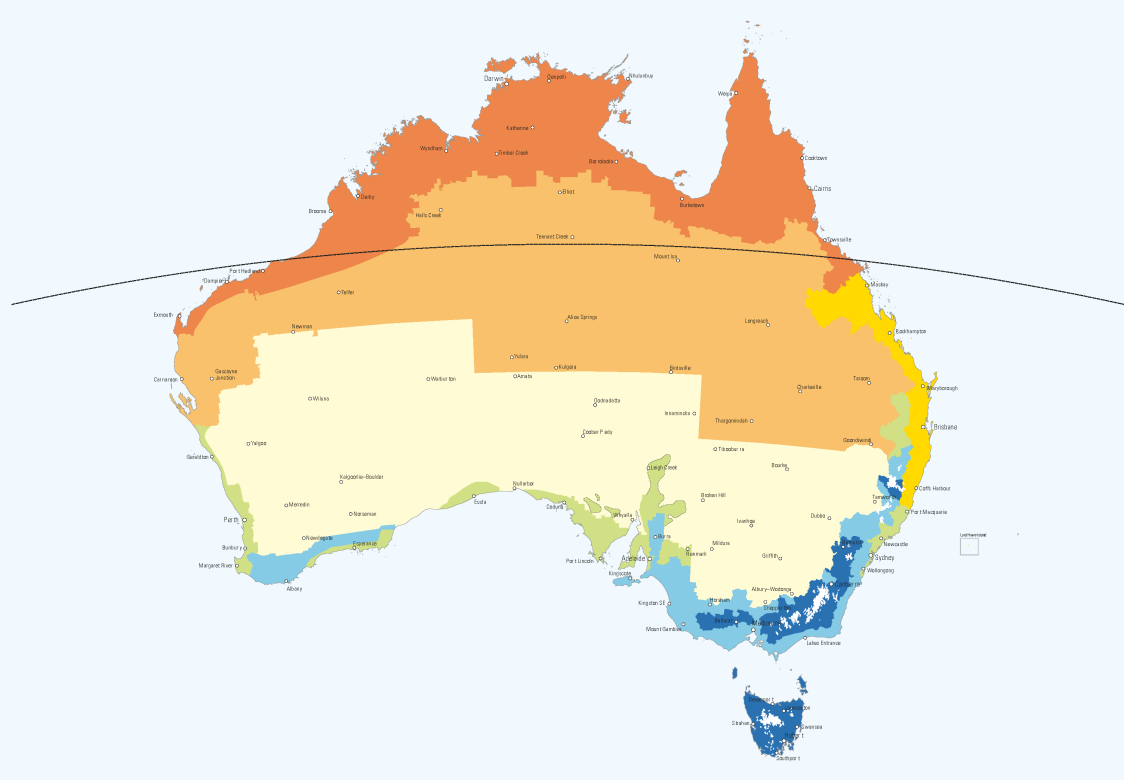

Figura 1: Mapa de las zonas climáticas de Australia - Fuente: Australian Building Codes Board

Tabla 5: Cantidad de observaciones y localizaciones por zona climática

	n_obs	n_locations
Zona 1	5084	4
Zona 2	3813	3
Zona 3	2542	2
Zona 4	8897	7
Zona 5	13981	11
Zona 6	13981	11
Zona 7	6355	5
Zona 8	1271	1

Missings

En la tabla 6 que se muestra abajo se refleja el porcentaje en tanto por ciento de valores faltantes en cada variable por zona climática.

Podemos observar que existen múltiples variables que tienen un *100 %* de valores faltantes para la zona climática 8. Además, en estas mismas variables existe una gran cantidad de valores faltantes en el resto de zonas climáticas, por lo que prescindiremos de ellas por su baja calidad del dato. Las variables de las que prescindiremos aplicando este criterio son *Evaporation*, *Sunshine*, *Pressure9am*, *Pressure3pm*, *Cloud9am* y *Cloud3pm*.

Una vez eliminadas dichas variables, resta por tratar el resto de columnas que presentan valores faltantes. Fijándonos de nuevo en la tabla 6 vemos que, a lo más, hay un *15 %* de valores faltantes para la variable *Humidity3pm* en la zona climática 1, mientras que el resto de zonas climáticas tiene menos valores faltantes

Tabla 6: Porcentaje de valores faltantes por cada zona y variable.

	Location	Date	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
zona1	0	0	0.7867821	0.6294256	1.4162077	9.677419	42.58458	1.5932337
zona2	0	0	0.5245214	0.8130081	3.3044847	64.096512	65.59140	12.0902177
zona3	0	0	1.4162077	0.2753737	2.4783635	57.553108	68.29268	2.3210071
zona4	0	0	0.1573564	0.1236372	2.4053052	44.745420	66.78656	0.9441385
zona5	0	0	1.1873257	0.9012231	1.6522423	61.376153	53.84450	9.9849796
zona6	0	0	4.0698090	4.0340462	4.7135398	42.743724	45.45455	11.0149489
zona7	0	0	0.0786782	0.0944138	0.5979544	80.062943	80.09441	1.2745869
zona8	0	0	0.0000000	0.0000000	0.3147128	100.000000	100.00000	0.5507474

	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
zona1	1.4948859	1.9276161	0.3933910	0.0590087	0.1573564	1.1408340	15.0472069
zona2	12.0902177	10.1757147	8.7070548	8.5759245	8.2612116	0.7605560	0.7343299
zona3	2.3210071	6.4909520	1.3375295	1.3768686	1.1801731	0.6294256	0.1966955
zona4	0.9441385	4.6420142	0.8317410	0.5507474	0.5395077	1.5510846	1.5623244
zona5	9.9778271	9.1481296	6.5374437	0.8225449	5.3858808	1.2159359	5.7148988
zona6	10.9005078	6.1726629	6.1082898	0.6294256	5.3930334	5.4431014	10.0207424
zona7	1.2745869	10.3068450	0.7395751	0.0786782	0.1101495	0.2517703	0.1888277
zona8	0.5507474	0.8654603	1.0228167	0.4720692	0.3933910	0.3933910	0.0786782

	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
zona1	0.0393391	0.0983478	8.064516	23.50511	0.6490952	13.9653816	1.4162077	1.4358773
zona2	7.9727249	7.9989510	54.287962	53.13402	0.4196171	0.4458432	3.3044847	3.3044847
zona3	0.3933910	0.1966955	48.505114	44.64988	0.5507474	0.1573564	2.4783635	2.4783635
zona4	0.3596718	0.3371923	37.034956	36.57413	0.1685962	0.1348769	2.4053052	2.4053052
zona5	18.3892425	18.3248695	45.619054	50.21815	0.8511551	5.3715757	1.6522423	1.6379372
zona6	13.6685502	13.6900079	40.547886	49.21679	3.9339103	8.6259924	4.7135398	4.7063872
zona7	18.2848151	18.2533438	46.766326	46.24705	0.0629426	0.0944138	0.5979544	0.5979544
zona8	100.0000000	100.0000000	100.000000	100.00000	0.2360346	0.0000000	0.3147128	0.3147128

en todas sus variables. Con el objetivo de presentar a los modelos datos limpios con los que puedan trabajar, vamos a imputar de diferentes maneras los datos faltantes según el tipo de dato de cada columna:

- En las columnas numéricas, sustuiremos en cada variable los valores faltantes por el valor promedio de dicha variable en cada zona.
- En las columnas categóricas, reemplazaremos los valores faltantes por el valor modal de cada variable en cada zona. Destacar que si el valor modal es NA, lo sustituiremos por el siguiente valor categórico más frecuente dentro de la zona climática.

La decisión del uso de estos estadísticos para imputar los valores faltantes de cada variable queda respaldada por el hecho de que el uso de estos estadísticos minimiza el impacto de la imputación sobre los propios estadísticos. Esto es, sustituir valores faltantes por la media no modifica la media de la distribución subyacente.

Vemos finalmente en la tabla 7 que tras el proceso de imputación detallado anteriormente ya no contamos con ningún valor faltante en ninguna variable.

Outliers

Mostramos a continuación una serie de diagramas de cajas y violín para cada variable separado por valor de la variable objetivo, para cada zona climática.

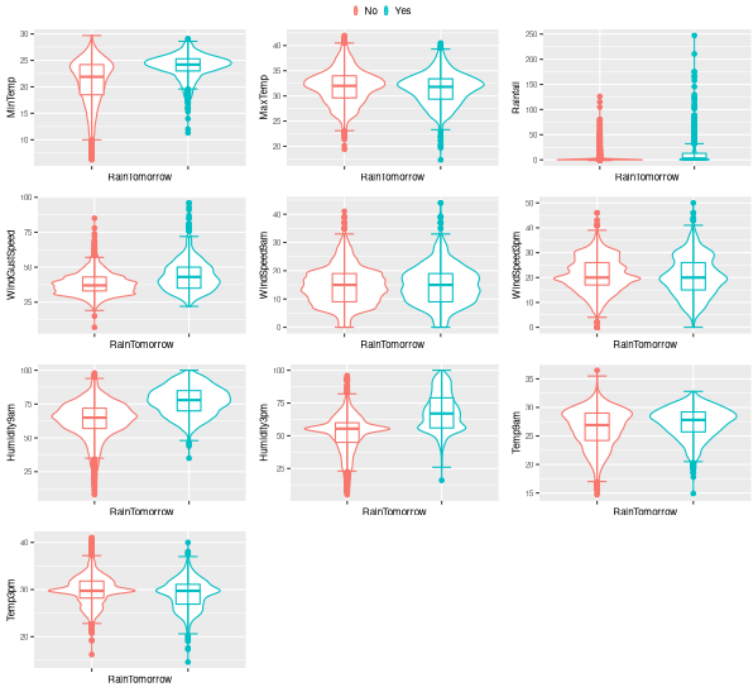

Figura 2: Boxplots y violin plots de las variables de la Zona 1, según el valor de 'RainTomorrow'

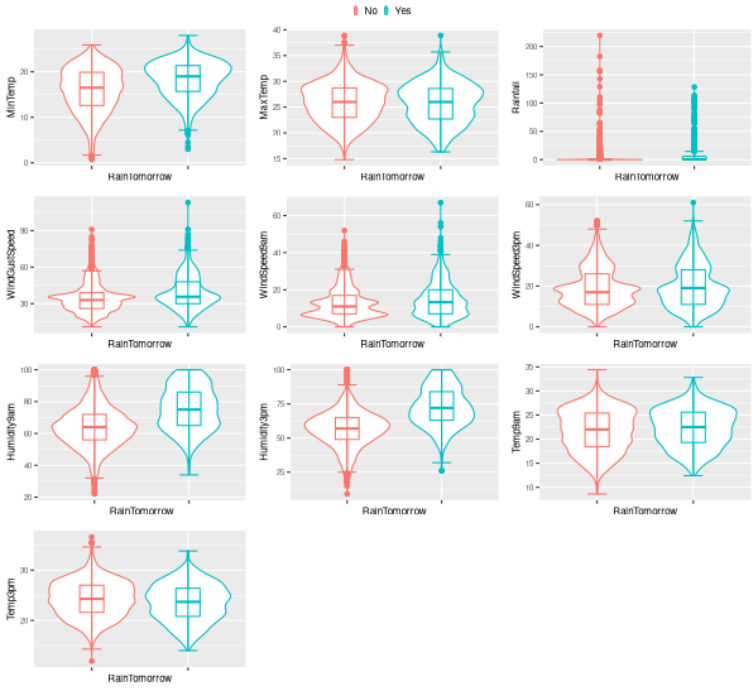

Figura 3: Boxplots y violin plots de las variables de la Zona 2, según el valor de 'RainTomorrow'

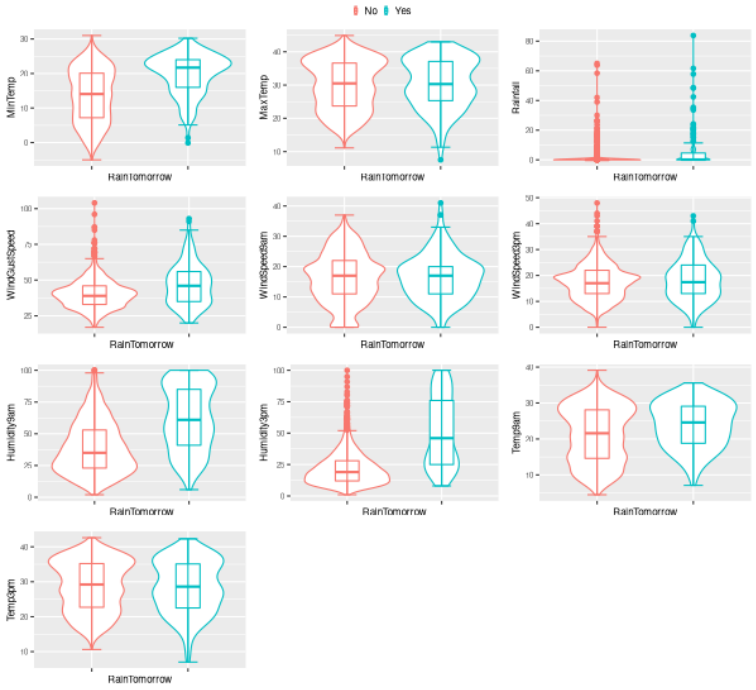

Figura 4: Boxplots y violin plots de las variables de la Zona 3, según el valor de 'RainTomorrow'

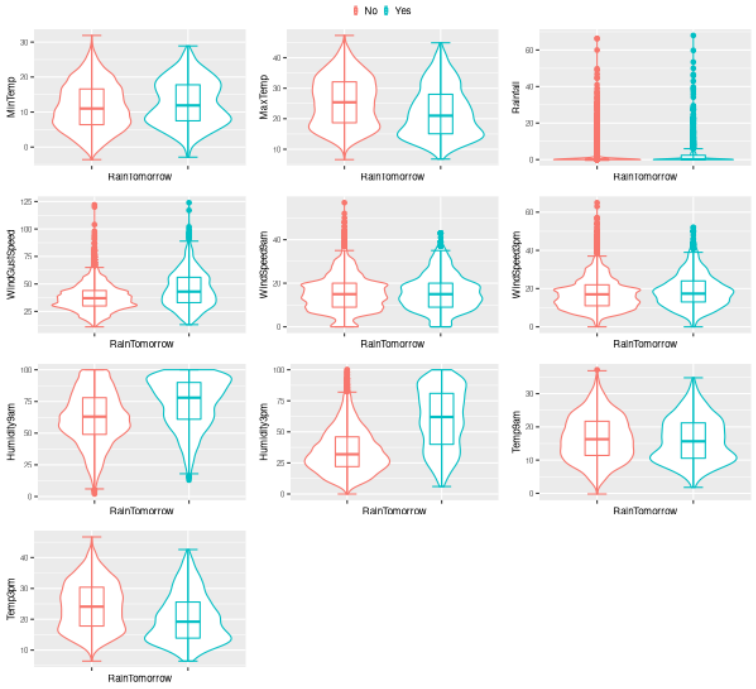

Figura 5: Boxplots y violin plots de las variables de la Zona 4, según el valor de 'RainTomorrow'

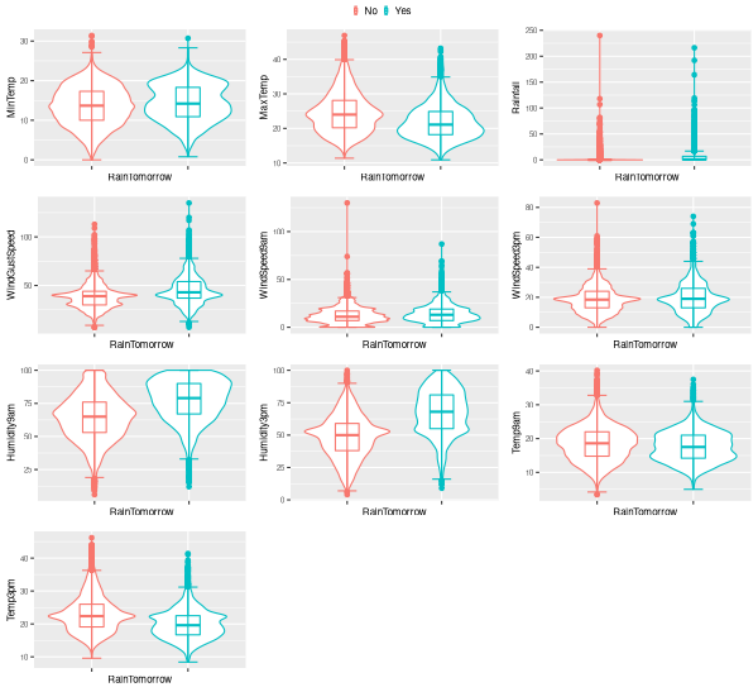

Figura 6: Boxplots y violin plots de las variables de la Zona 5, según el valor de 'RainTomorrow'

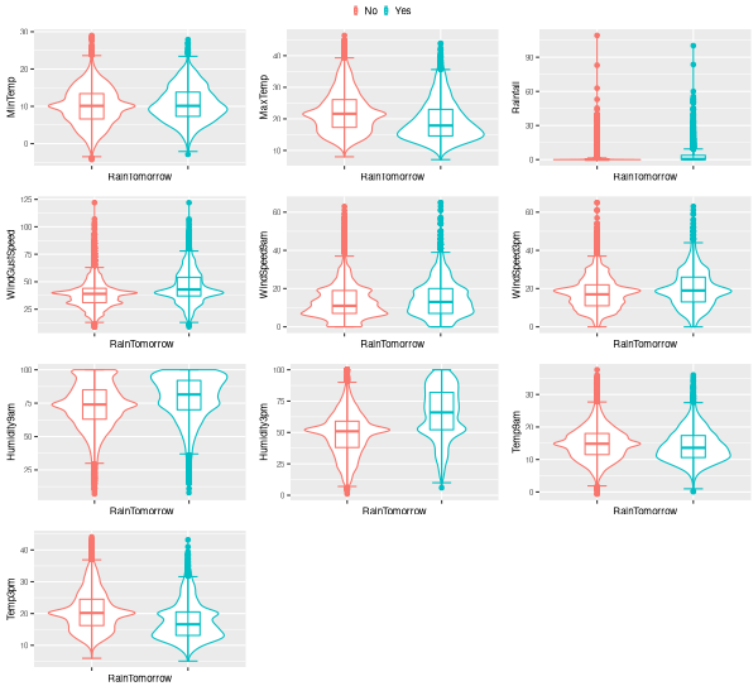

Figura 7: Boxplots y violin plots de las variables de la Zona 6, según el valor de 'RainTomorrow'

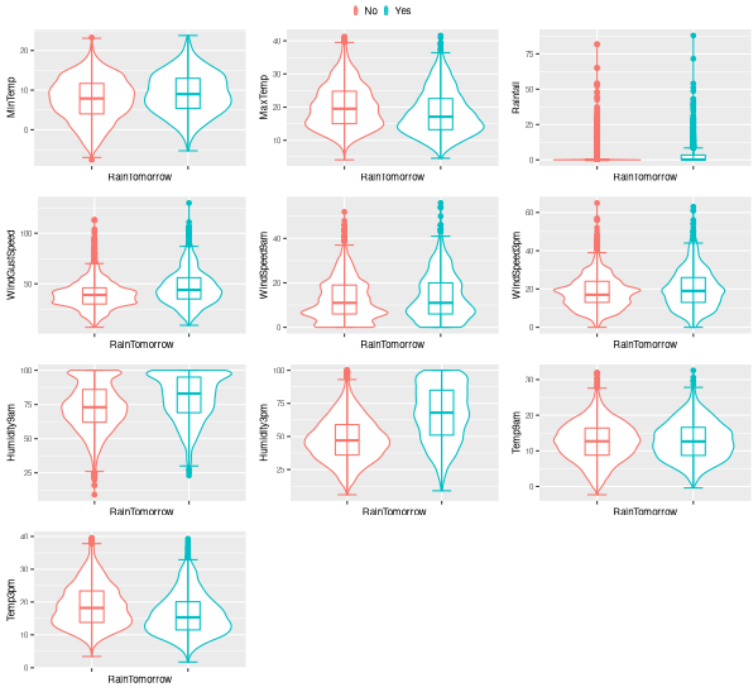

Figura 8: Boxplots y violin plots de las variables de la Zona 7, según el valor de 'RainTomorrow'

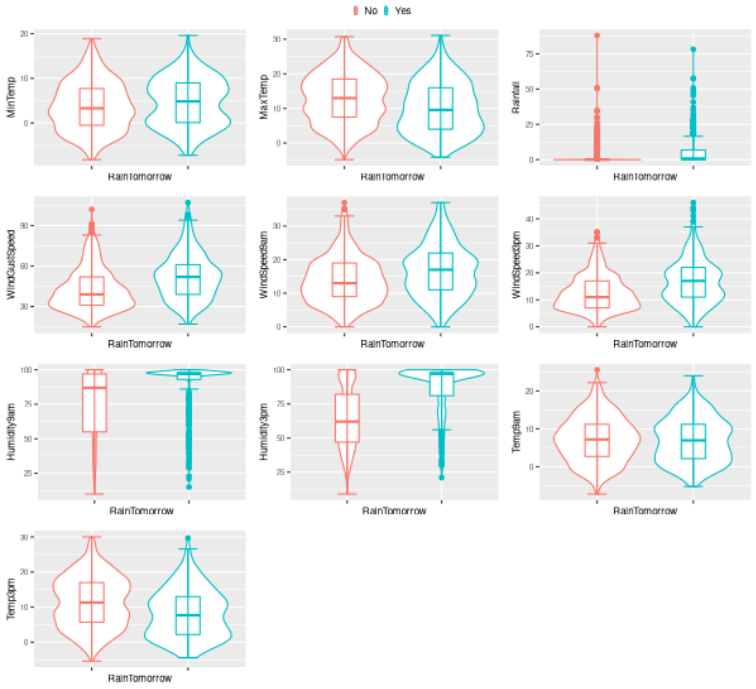

Figura 9: Boxplots y violin plots de las variables de la Zona 8, según el valor de 'RainTomorrow'

Tabla 7: Muestra de variables con valores faltantes imputados.

	Rainfall	WindGustDir	WindGustSpeed	WindDir3pm	WindSpeed3pm	Humidity3pm	Temp3pm	RainTomorrow
zona1	0	0	0	0	0	0	0	0
zona2	0	0	0	0	0	0	0	0
zona3	0	0	0	0	0	0	0	0
zona4	0	0	0	0	0	0	0	0
zona5	0	0	0	0	0	0	0	0
zona6	0	0	0	0	0	0	0	0
zona7	0	0	0	0	0	0	0	0
zona8	0	0	0	0	0	0	0	0

Debido a la naturaleza caótica del tiempo, no vamos a tratar los outliers como valores atípicos porque consideramos que son de valor para el entrenamiento de los modelos y no introducen demasiado ruido. Nos limitaremos entonces a realizar una normalización de los datos numéricos junto con una selección de variables aplicando por un algoritmo de *ranking* de variables tipo *step*.

No obstante, si que es posible en este punto estimar qué variables van a ser más relevantes para los modelos en cada zona. En aquellas variables en donde exista una gran diferencia visual entre el caso `RainTomorrow=YES` y el caso `RainTomorrow=NO` tendremos que las variables aleatorias subyacentes en estos casos son diferentes, y por tanto aportarán más información al modelo. De este modo, estas variables son `WindGustSpeed`, `WindGustSpeed9am`, `MinTemp`, `WindGustSpeed`, `Humidity9am` y `Humidity3pm`, dependiendo de si nos encontramos en una u otra zona climática.

Finalmente, en la tabla 8 comprobamos si la variable `RainTomorrow` tiene categorías balanceadas en caza zona. Se puede observar que existe cierta desproporción ya que hay valores inferiores al 30 de casos de lluvia en todas las zonas, especialmente la zona 3 con sólamente un 7 de valores positivos. Si tras un primer modelado no consiguiéramos unos resultados satisfactorios, una técnica que podría aplicarse, conocida como *downsampling*, consiste en seleccionar un subconjunto de entrenamiento donde los casos de la variable objetivo estén más proporcionados.

Tabla 8: Índice de lluvia por zona climática

	perct_rain_tomorrow
Zona1	22.639654
Zona2	23.603462
Zona3	7.435091
Zona4	13.217939
Zona5	23.295902
Zona6	21.443388
Zona7	21.117231
Zona8	26.278521

Transformación de las variables categóricas

Las variables categóricas que disponemos son *Location*, *WindGustDir*, *WindDir9am*, *WindDir3pm*, *RainToday* y *RainTomorrow*. Para la variable *Location* vamos a aplicar técnicas de one hot encoding con las que transformaremos estas variables en numéricas. Las variables *RainToday* y *RainTomorrow* por ser binarias no necesitan ninguna transformación más que cambiar su tipo de `TRUE` y `FALSE` a unos y ceros (tipo numérico). Por otro lado, para las variables categóricas relativas a la dirección del viento, vamos a aplicar una transformación de forma que combinemos esta información con las variables enteras relativas a la intensidad del viento, obteniendo dos variables numéricas que expresan el valor del coseno y del seno del vector viento.

Del proceso de *one hot encoding* podemos estacar que en la zona 8 se ha eliminado la variable `Location` ya que esta sólamente tenía un único valor, por lo que no aporta información al modelado para esta zona. Además, se han creado variables numéricas del tipo `Location.Katherine` que indican si la observación pertenece o no a dicha localización. De esta manera, tras el proceso de *one hot encoding*, explotamos cada variable categórica con n niveles en n variables binarias, conocidas como *dummy variables*.

Tras el tratamiento de esas variables categóricas, vamos a asignar un valor en radianes a cada dirección del viento según se muestran en la figura 10[5]. Después, calcularemos el coseno (componente este-oeste) y el seno (componente norte-sur) de dicha dirección y los multiplicaremos por sus respectivas variables de intensidad del viento. Con esta transformación, condensamos todas las variables relativas a la dirección e intensidad del viento en la coordenada x e y del vector viento.

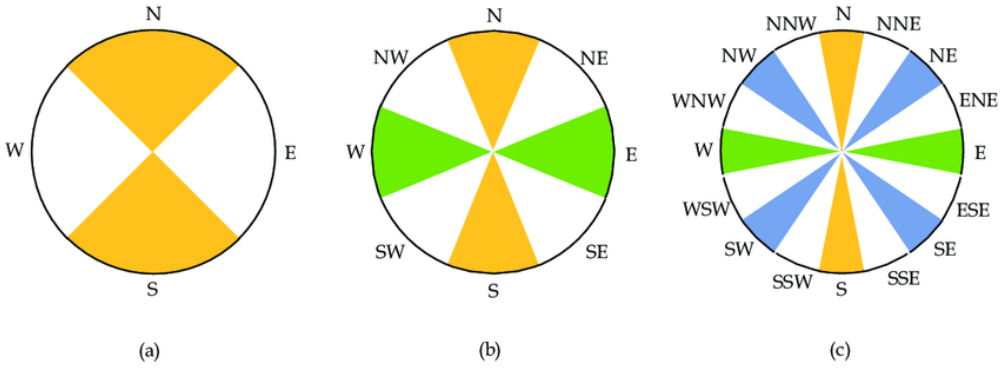

Figura 10: 16 direcciones del viento

Tabla 9: Conversión de variables de viento

Date	MinTemp	MaxTemp	Rainfall	Humidity9am	Humidity3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Location.MountGinini	WindGustDir_x	WindGustDir_y	WindDir9am_x	WindDir9am_y	WindDir3pm_x	WindDir3pm_y
2014-01-01	7.1	16.3	0.0	64	45	12.0	14.5	0	0	1	-70	0	-12.01043	4.974885	-12.01043	-4.974885
2014-01-02	9.3	16.9	0.2	95	36	13.1	14.8	0	0	1	-78	0	-20.32535	-8.419035	-22.17311	-9.184402
2014-01-03	10.1	21.6	0.0	97	58	12.9	19.3	0	0	1	-61	0	-15.00000	0.000000	-22.17311	9.184402
2014-01-04	4.2	16.5	0.0	52	35	8.3	15.3	0	0	1	-50	0	-12.01043	-4.974885	-22.17311	-9.184402
2014-01-05	4.8	18.0	0.0	62	33	9.6	16.7	0	0	1	-65	0	-17.00000	0.000000	-24.02087	9.949769
2014-01-06	2.1	15.5	0.0	54	30	6.6	13.9	0	0	1	-76	0	-13.85819	-5.740252	-13.85819	-5.740252

Con este proceso, hemos transformado las variables categóricas `WindGustDir`, `WindDir9am` y `WindDir3pm` en dos componentes numéricas, combinándolas respectivamente con `WindGustSpeed`, `WindSpeed9am` y `WindSpeed3pm`.

Normalización de datos.

Como útima etapa del tratamiento de datos, nos disponemos a aplicar una normalización estándar sobre las variables numéricas que disponemos que no sean tipo *one hot*. El objetivo de este proceso es que todas tengan el mismo peso en la etapa de modelado, además de ser clave ya que vamos a trabajar con modelos

[5]Fuente - Yannopoulos (2011)

basados en distancias. Vemos en la tabla 10 como el valor de todas las variables numéricas ha pasado a estar entre -1 y 1, salvo las de tipo *one hot* que mantendremos en $[0, 1]$.

Tabla 10: Variables normalizadas a 0-1 – Zona 3

Date	MinTemp	MaxTemp	Rainfall	Humidity9am	Humidity3pm	Temp9am	Temp3pm	RainToday	RainTomorrow	Location.AliceSprings	Location.Uluru	WindGustDir_x	WindGustDir_y	WindDir9am_x	WindDir9am_y	WindDir3pm_x	WindDir3pm_y
2014-01-01	1.7412780	1.8469692	-0.1796094	-1.6081750	-1.0852394	2.2795558	1.8782571	0	0	1	0	-0.5892848	2.0130743	-0.1280432	2.1449425	-0.5070789	1.5410655
2014-01-02	1.2100051	1.9808611	-0.1796094	-0.9586381	-0.7225000	2.1376252	1.8920452	0	0	1	0	-1.9527091	-1.2581917	-2.5965255	0.1721940	-1.7407151	-1.0735842
2014-01-03	1.8838147	1.0703960	-0.1796094	-0.5410787	-0.6620434	1.2989445	1.0509708	0	0	1	0	-1.4228158	-1.9643652	-0.8859926	-2.3513278	-1.2961070	-1.7978773
2014-01-04	0.9638054	0.4411039	-0.1796094	-0.8658471	-0.4806737	0.6538055	0.4580823	0	0	1	0	0.9941816	-0.4882094	0.8964719	-1.7494668	0.4453622	-0.1681480
2014-01-05	0.3288695	0.7356662	-0.1796094	-1.4225930	-1.1456959	0.7699305	0.7338444	0	0	1	0	0.4964050	-0.9559425	0.3235369	-1.1317901	1.3978032	-0.5938396
2014-01-06	0.1604171	0.4411039	-0.1796094	-0.8194517	-0.6620434	0.5376805	0.4994466	0	0	1	0	0.9003825	-1.3955778	1.6924347	-0.9792294	1.1751523	-1.5576307

Representación de los datos, tratamiento de la variable `tiempo` y selección de variables

Representación de las series temporales

Nos disponemos a continuación a visualizar las series temporales de las variables numéricas de la zona 1. El objetivo de esta visualización es apoyar la descomposición que haremos más tarde de las series temporales, aportándo variables tipo ventana con información de estadísticos previos.

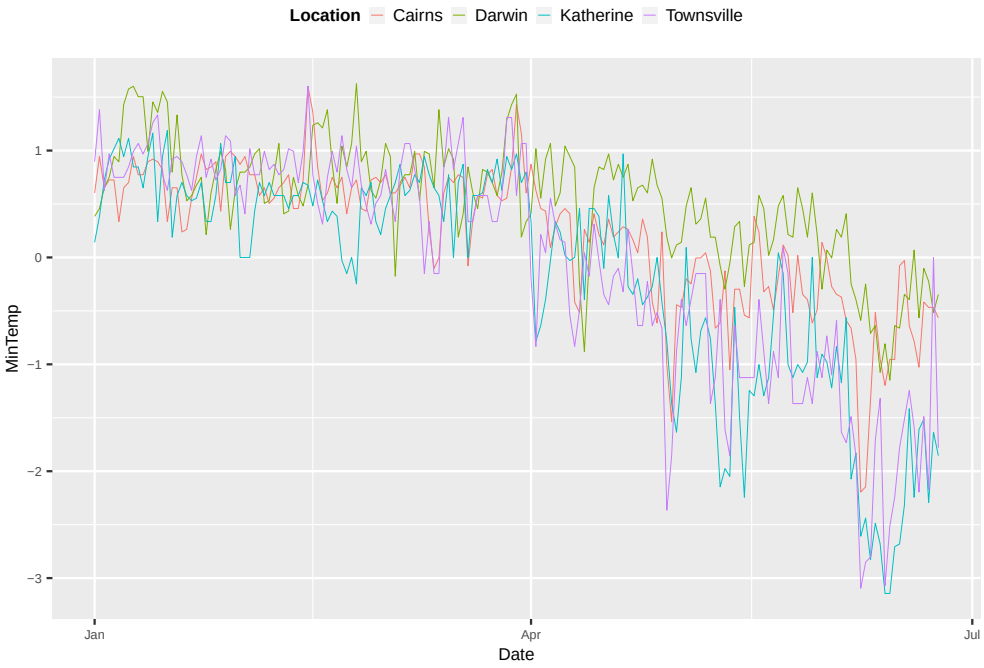

Figura 11: Serie temporal de la variable 'MinTemp' en la zona 1, según el valor de 'Location'

Figura 12: Serie temporal de la variable 'MaxTemp' en la zona 1, según el valor de 'Location'

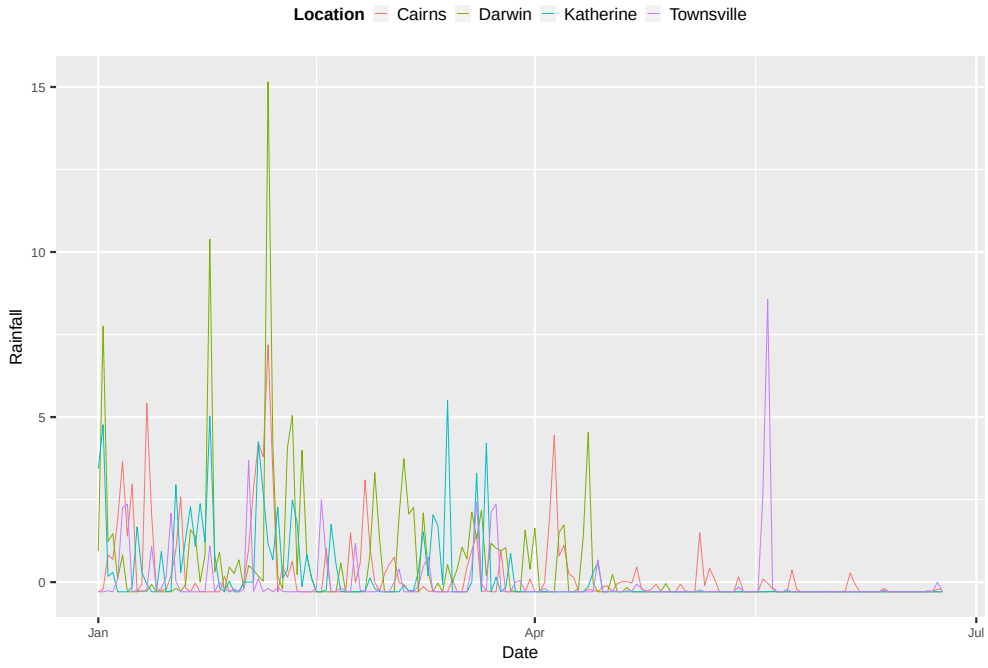

Figura 13: Serie temporal de la variable 'Rainfall' en la zona 1, según el valor de 'Location'

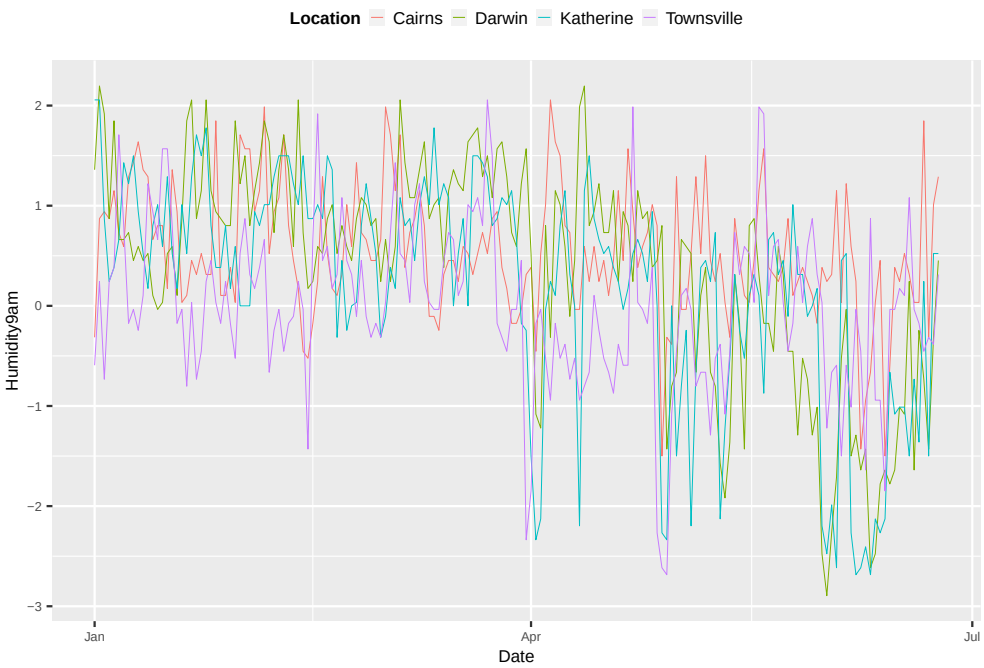

Figura 14: Serie temporal de la variable 'Humidity9am' en la zona 1, según el valor de 'Location'

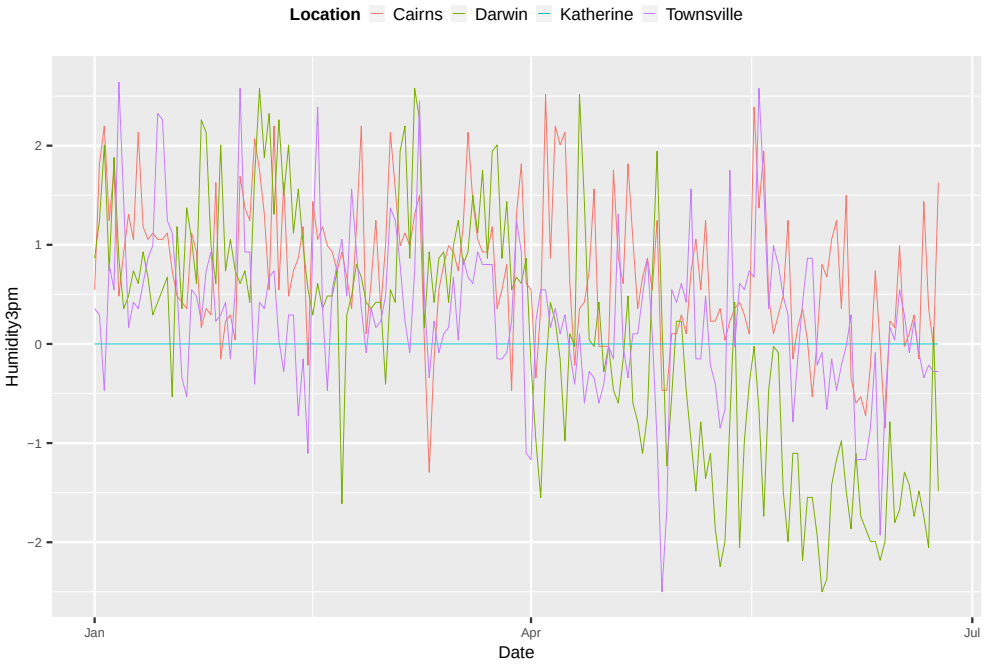

Figura 15: Serie temporal de la variable 'Humidity3pm' en la zona 1, según el valor de 'Location'

Figura 16: Serie temporal de la variable ‘Temp9am‘ en la zona 1, según el valor de ‘Location‘

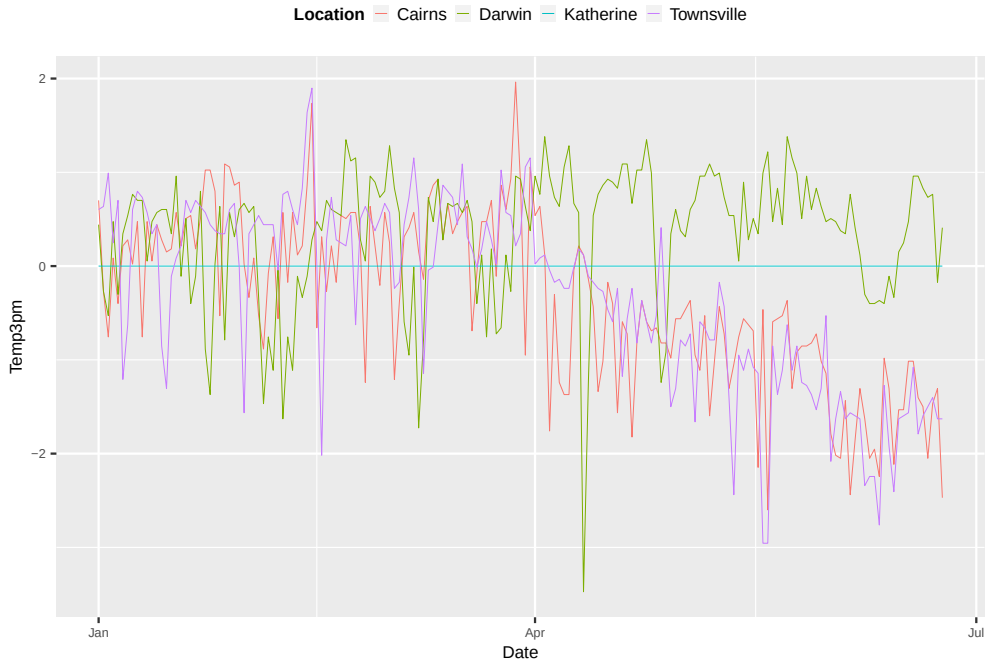

Figura 17: Serie temporal de la variable ‘Temp3pm‘ en la zona 1, según el valor de ‘Location‘

Figura 18: Serie temporal de la variable 'WindGustDir_x' en la zona 1, según el valor de 'Location'

Figura 19: Serie temporal de la variable 'WindGustDir_y' en la zona 1, según el valor de 'Location'

Figura 20: Serie temporal de la variable 'WindDir9am_x' en la zona 1, según el valor de 'Location'

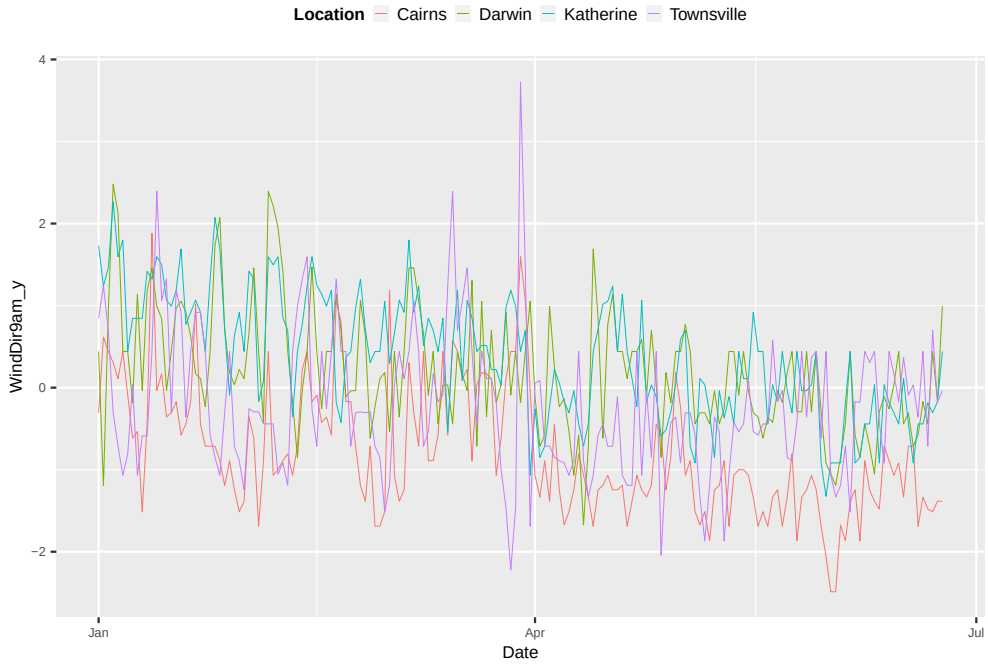

Figura 21: Serie temporal de la variable 'WindDir9am_y' en la zona 1, según el valor de 'Location'

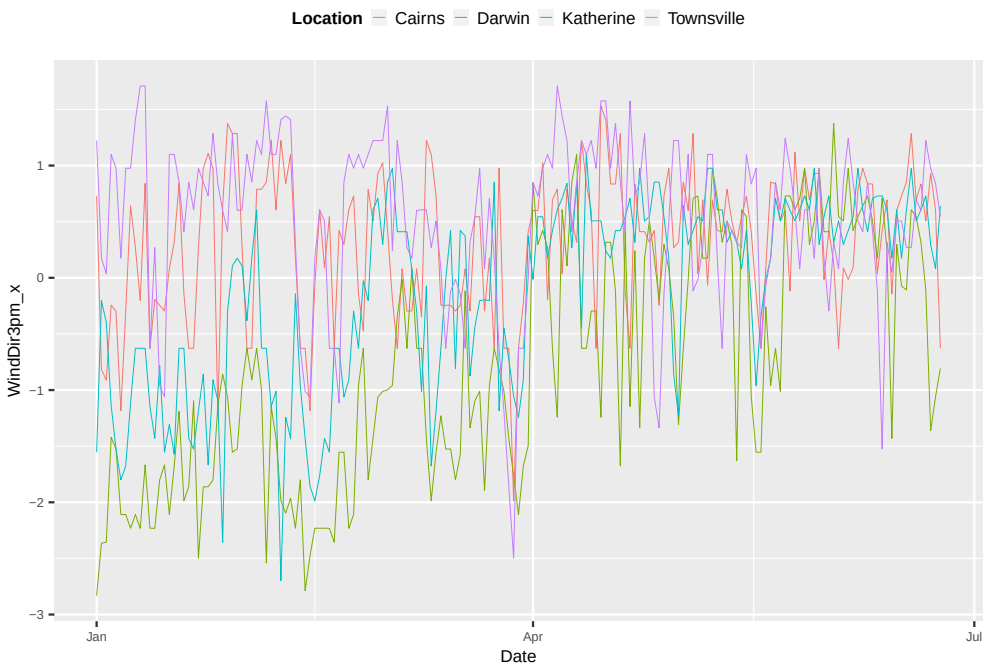

Figura 22: Serie temporal de la variable 'WindDir3pm_x' en la zona 1, según el valor de 'Location'

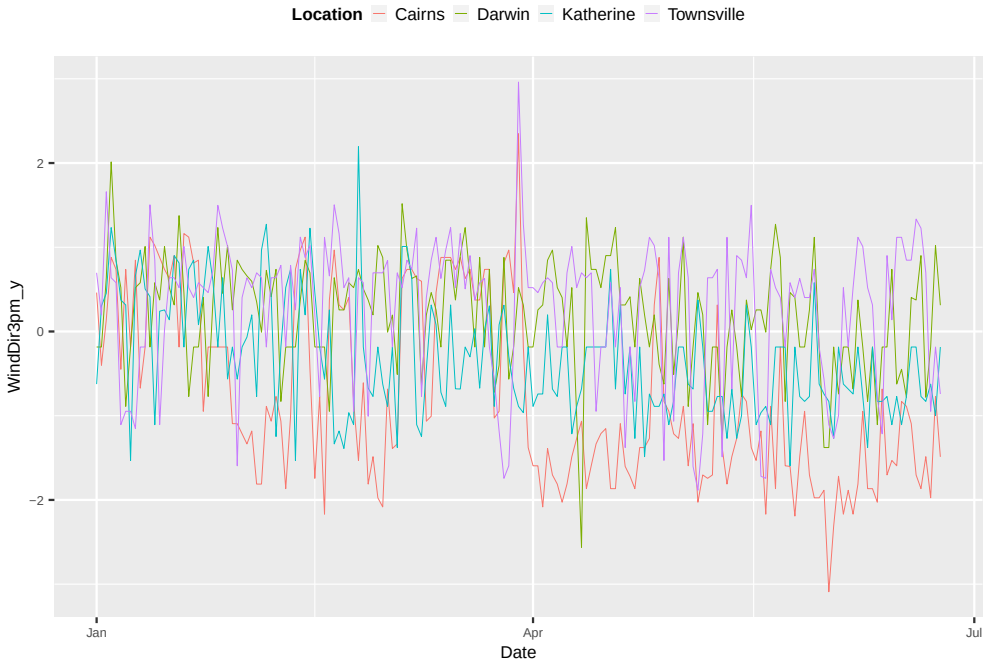

Figura 23: Serie temporal de la variable 'WindDir3pm_y' en la zona 1, según el valor de 'Location'

Creación de variables ventana

Siguiendo lo propuesto en Hällman (2017), vamos a transformar la variable *Date* en una serie de variables temporales tipo ventana. La idea subyacente a esta técnica es que si para cada variable de un conjunto de datos añadimos columnas con información relativa al pasado, entonces podemos prescindir de la variable del tiempo ya que esta queda representada por un subconjunto de las variables ventana.

De esta manera, para cada variable del conjunto de datos hemos creado otra variable conteniendo el promedio de los últimos cinco días. A esta técnica se la conoce como *rolling windows*, y facilita que podamos tratar una serie temporal como un conjunto de datos estático en el tiempo.

Tabla 11: Muestra de algunas variables ventana de la zona 1

WindGustDir_x_mean5	Temp9am_mean5	WindDir3pm_x_mean5	Humidity9am_mean5	WindGustDir_y_mean5
0.1669130	-0.1620371	0.3843359	0.2823871	-0.5555522
0.1669130	-0.1620371	0.3843359	0.2823871	-0.5555522
0.1669130	-0.1620371	0.3843359	0.2823871	-0.5555522
0.1669130	-0.1620371	0.3843359	0.2823871	-0.5555522
-0.2750914	1.1360396	-0.2907757	-0.0353920	1.2319710
-0.1989804	1.1483498	-0.2068851	0.0482835	1.1789291

Feature selection

Como paso final del procesamiento de datos previo a la etapa de modelado, nos disponemos a realizar una selección de variables. El objetivo de dicho procedimiento es reducir la dimensionalidad de los conjuntos de datos con los que estamos trabajando al tiempo que se intenta condensar la información que aportan las variables explicativas en un subconjunto de las mismas.

Con esta idea en mente, para cada zona vamos a realizar una selección de variables siguiendo el algoritmo *step* descrito en Venables and Ripley (2002). Este algoritmo realiza una regresión sencilla, añadiendo variables a un modelo al tiempo que calcula el aporte de cada variable sobre la predicción siguiendo el *criterio de información de Akaike*. Con esto, en cada iteración se añaden o eliminan variables hasta que se obtiene una selección de las variables más relevantes para la predicción.

Modelado

En la fase de modelado, nos disponemos a aplicar los tres modelos explicados en teoría, junto con modelos tipo *Decission Trees* y *Random Forest*. A continuación, vamos a detallar el proceso de modelado para la zona climática 7, que dispone de cinco localizaciones diferentes y, como vimos en Agrupación por zona climática.

Train-test split

El primero de los pasos del proceso de modelado, partiendo de los datos procesados anteriormente, consiste en dividir el conjunto de datos en dos subconjuntos. Uno de ellos, que llamaremos *train*, lo usaremos para entrenar a cada uno de los modelos y el otro subconjunto, al que llamaremos *test*, lo usaremos para comprobar la precisión de nuestros modelos.

Así pues, se puede comprobar en la siguiente celda de código como se ha obtenido un subconjunto aleatorio que contiene tres cuartas partes del original como conjunto *train*, y el restante se ha tomado como conjunto *test*. Además, vemos que esta selección mantiene la proporción entre los casos de la variable objetivo *RainTomorrow* para ambos conjuntos.

```r
## 75% of the sample size
smp_size <- floor(0.75 * nrow(main_data))

#
set.seed(123)
train_ind <- sample(seq_len(nrow(main_data)), size = smp_size)

train <- main_data[train_ind, ]
test <- main_data[-train_ind, ]


#train$RainTomorrow%>% table()
#test$RainTomorrow%>% table()
freq_df <- data.frame(train$RainTomorrow %>% table(),test$RainTomorrow %>% table())
freq_df <- freq_df[,c("Freq","Freq.1")]
rownames(freq_df) <- c("0","1")
colnames(freq_df) <- c("Freq_train","Freq_test")
my_cap <- "Frecuencia de los niveles de la variable RainTomorrow en los conjuntos train y test"
freq_df %>%  kbl(., booktabs = T,
                 caption = my_cap) %>% kable_styling(latex_options = c("striped","hold_position"))
```

Tabla 12: Frecuencia de los niveles de la variable RainTomorrow en los conjuntos train y test

	Freq_train	Freq_test
0	3762	1251
1	1004	338

Modelado inicial - Zona 7

En un primer modelado, vamos a aplicar los datos de la zona 7 sobre cinco familias de modelos, y comparar sus resultados. En concreto, los modelos que hemos aplicado son:

- Regresión logística, sin ningún parámetro adicional.
- Suport Vector Machines. En este caso, hemos creado modelos con *kernels* lineal, polinomial y sigmoidal. En todos los casos hemos aplicado una óptimización de hiperparámetros siguiendo potencias de 2, tanto positivas como negativas. Además, hemos aplicado validación cruzada con cinco pliegues para estimar la precisión del modelo.
- K-Nearest Neighbours. En el caso de este modelo, hemos aplicado una optimización del hiperparámetro k, haciéndolo variar entre $k = 3$ y $k = 45$. Además, hemos aplicado el mismo tipo de validación cruzada que para el *SVM*.
- Árboles de decisión y Random forest. Para los árboles de decisión hemos decidido dejar los hiperparámetros por defecto, y en el caso de Random forest hemos construido el modelo a partir de 50 árboles.

De esta manera, obtenemos las siguientes matrices de confusión para cada modelo.

Como vemos, se puede obtener la precisión o *accuracy* de cada modelo sumando los porcentajes de la diagonal principal de las respectivas matrices de confusión. Con esto, concluimos que para la zona 7 el modelo que más eficiente resulta para la predicción es el *random forest* con 50 árboles, dando una precisión del 83,6.

Aplicando la misma estrategia sobre el resto de zonas, obtenemos los modelos más precisos por cada zona.

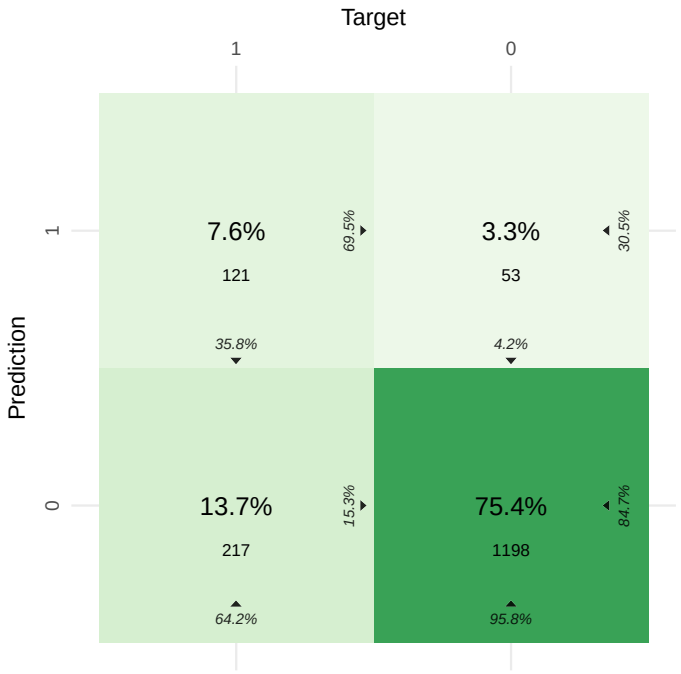

Figura 24: Matriz de confusión regresión logística

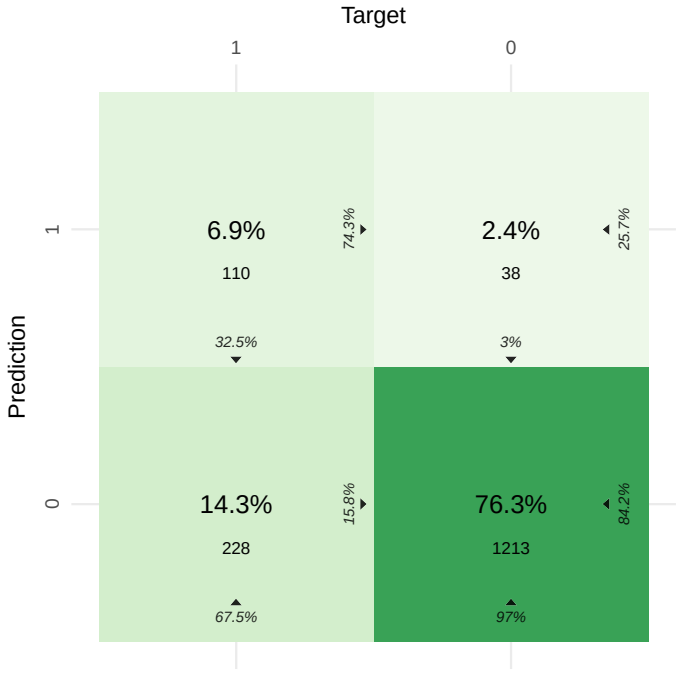

Figura 25: Matriz de confusión svm con kernel lineal

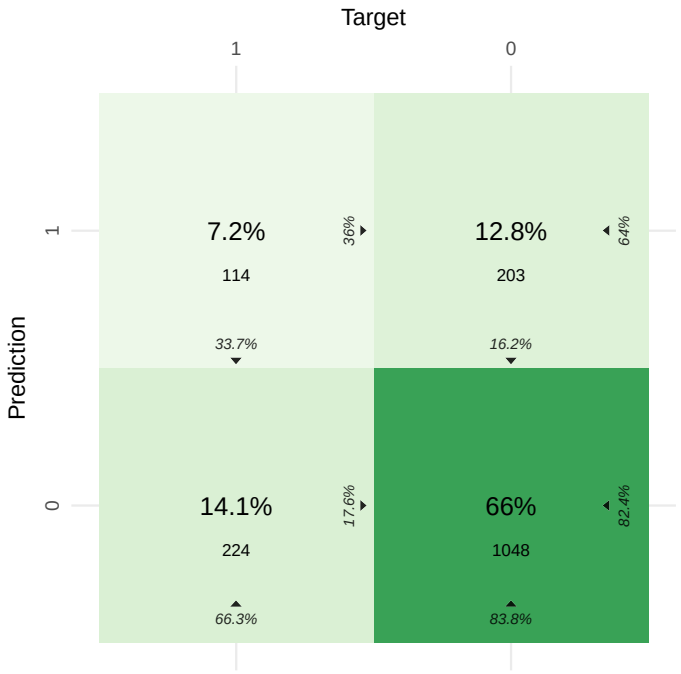

Figura 26: Matriz de confusión svm con kernel sigmoidal

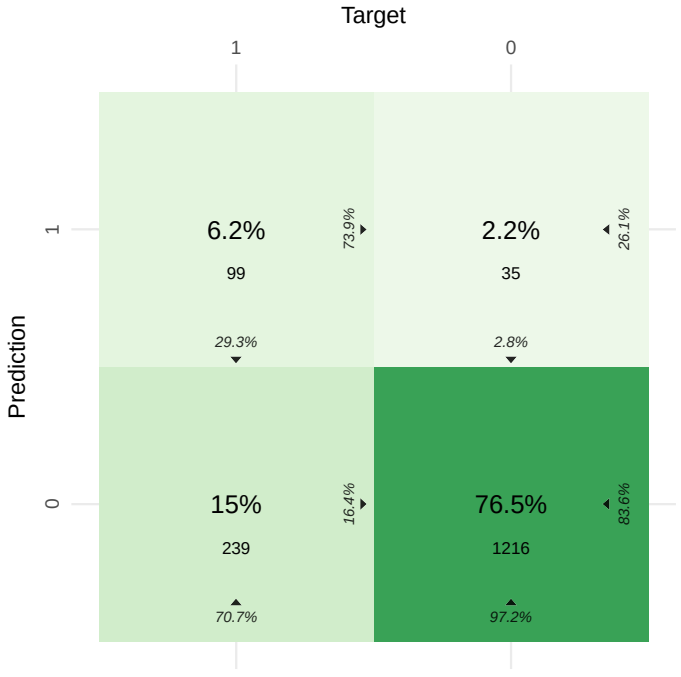

Figura 27: Matriz de confusión knn.

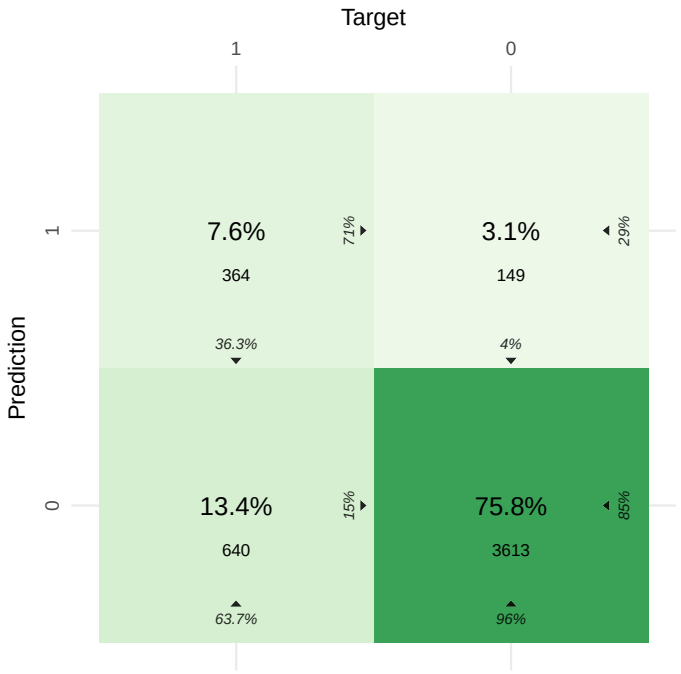

Figura 28: Matriz de confusión del árbol de decisión

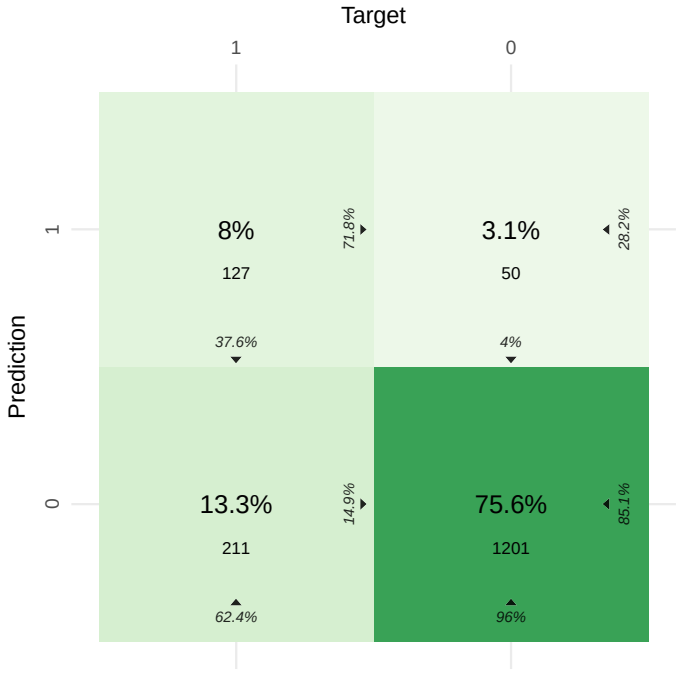

Figura 29: Matriz de confusión del random forest

- Para la zona 1, el modelo *SVM* con kernel polinomial de grado 3 y un coste de 1 obtiene una precisión del 85,5.
- Para la zona 2, el modelo *SVM* con kernel lineal y un coste de 4 obtiene una precisión del 97,06.
- Para la zona 4, el modelo *SVM* con kernel polinomial de grado 3 y un coste de 1 obtiene una precisión del 91,37.
- Para la zona 5, el modelo *SVM* con kernel polinomial de grado 3 y un coste de 2 obtiene una precisión del 84,92.
- Para la zona 6, el modelo *SVM* con kernel polinomial de grado 3 y un coste de 16 obtiene una precisión del 84,75.
- Para la zona 8, el modelo de *regresión logística* obtiene una precisión del 96,22.

Como vemos, en aquellas zonas donde hay menos localizaciones agrupadas los modelos logran alcanzar mayor precisión. Además, tenemos que el modelo de *SVM* con kernel polinomial ha obtenido resultados dominantes. Por esto, una posible propuesta de mejora a nuestro proceso de modelado sería realizar un modelo por localización, aplicando *SVM* polinomial junto con una búsqueda intensiva de hiperparámetros.

Conclusiones

Con la realización de esta memoria se ha conseguido de forma satisfactoria aplicar conocimientos teóricos a un caso práctico. Como se ha visto, la combinación de conocimientos teóricos específicos junto con una aplicación de dichos conocimientos ha permitido abordar un problema con estrategias que sin los primeros no habría sido posible. De este modo, gracias al trabajo de revisión bibliográfica junto con la construcción en el lenguaje de programación R de modelos de clasificación supervisada, se ha logrado reflejar la capacidad de las matemáticas para analizar, complementar y transformar el mundo moderno.

Referencias

Australian Goverment. 2004. “Notes about Daily Weather Observations.” Bureau of Meteorology. http://www.bom.gov.au/climate/dwo/IDCJDW0000.pdf.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. “Cluster Analysis.” In, 5th ed., 260–62. John Wiley & Sons.

Hällman, Ludvig. 2017. “The Rolling Window Method: Precisions of Financial Forecasting.” TRITA-MAT-e. Master’s thesis, KTH, Mathematical Statistics; KTH, Mathematical Statistics.

Venables, Bill, and B Ripley. 2002. “Modern Applied Statistics with s.” In *Springer*. https://doi.org/10.1007/b97626.

Yannopoulos, Panayotis. 2011. “Quick and Economic Spatial Assessment of Urban Air Quality.” In. https://doi.org/10.5772/22425.