FIUBA
CEAI - Análisis de Datos
Trabajo final integrador

Introducción:

En el presente trabajo se realiza un completo desarrollo de un modelo de aprendizaje automático supervisado aplicado sobre un dataset descargado de Kaggel con datos de distintas estaciones meteorológicas de Australia con el objetivo de predecir si lloverá el día siguiente.

1 - Análisis exploratorio inicial

Del estudio del dataset se desprenden las siguiente conclusiones:

- El dataset cuenta con 145460 muestras, 22 features y una salida ("Rain Tomorrow").
- Solo las feature 'Date' y 'Location' están completas por lo que es necesario hacer ingeniería de features, estudiando cada caso.
- Detectamos 1 variable compuesta:
  - Date: Fecha de observación codificada en Año-Mes-Dia
- Detectamos 7 variables categóricas:
  - Location: Nombre propio de la ciudad
  - WindGustDir: Dirección de la rafaga de viento más fuerte del día, codificado con los puntos cardinales.
  - WindDir9am: Dirección del viento a las 9 am, codificado con los puntos cardinales.
  - WindDir3pm: Dirección del viento a las 3 pm, codificado con los puntos cardinales.
  - Cloud9am: Fracción del cielo obstruido a las 9 am, medido en oktas.
  - Cloud3pm: Fracción del cielo obstruido a las 3 pm, medido en oktas.
  - RainToday: Indicación booleana de Sí/No
- Detectamos 14 variables numéricas tipo _flotante de 64 bits_:
  - MinTemp: Temperatura minima en grados celsius
  - MaxTemp: Temperatura máxima en grados celsius
  - Rainfall: Lluvia caida durante el dia en mm
  - Evaporation: Evaporacion de la bandeja (Pan evaporation).
  - Sunshine: Número de horas de sol durante el día.
  - WindGustSpeed: Velocidad de la rafaga de viento más fuerte del día en km/h.
  - WindSpeed9am: Velocidad del viento promedio durante las 8:50 am y las 9 am en km/h.
  - WindSpeed3pm: Velocidad del viento promedio durante las 2:50 pm y las 3 pm en km/h.
  - Humidity9am: Humedad en % a las 9 am.
  - Humidity3pm: Humedad en % a las 3 pm.
  - Pressure9am: Presión atmosférica a las 9 am en hpa.
  - Pressure3pm: Presión atmosférica a las 3 pm en hpa.
  - Temp9am: Temperatura a las 9 am en grados celsius.
  - Temp3pm: Temperatura a las 3 pm en grados celsius.
- Una salida tipo categórica:
  - RainTomorrow: Indicación booleana de Sí/No

Variables numéricas

Para el análisis de distribución analizaremos en grupos de variables que considero variables semejantes, de cada variable se realizó un histograma y un diagrama de Box-Whiskers para su análisis. Se agruparon los 14 features en 5 grupos:

- Temperatura: 'Temp9am', 'MinTemp', 'Temp3pm', 'MaxTemp'
  - Las distribuciones de las temperaturas responden en gran medida a una distribución normal por lo que en principio no es necesario realizar ninguna transformación sobre ellas.

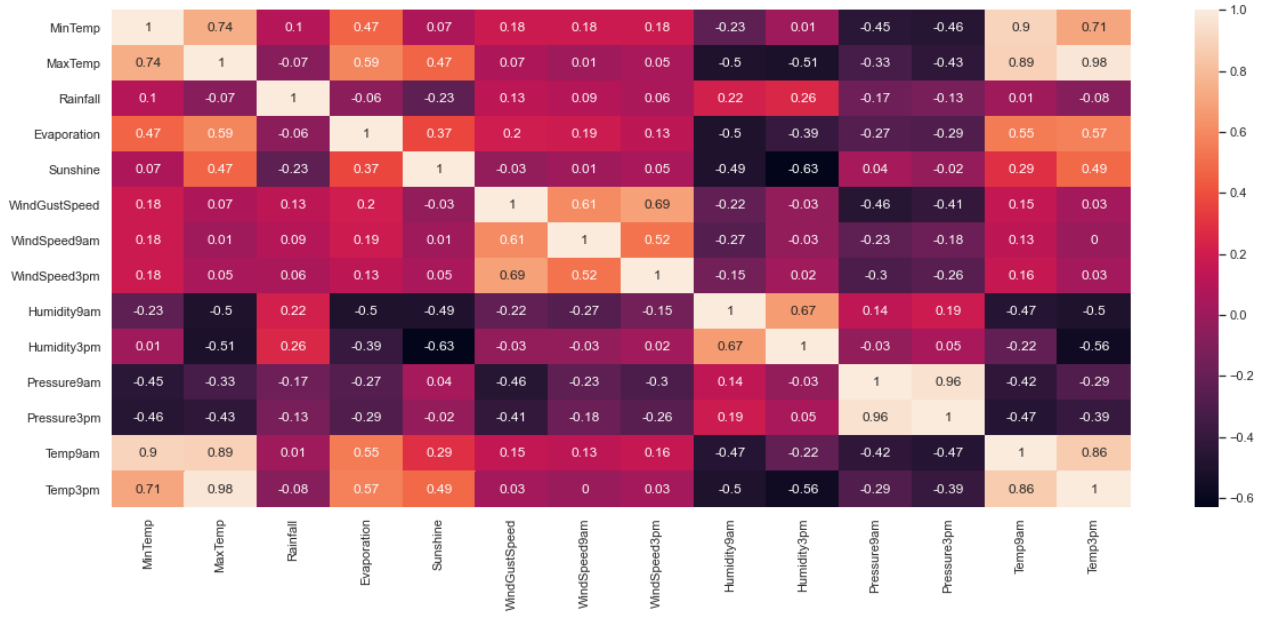
- Humedad: 'Humidity9am', 'Humidity3pm'
    - Las distribuciones de la humedad tienen una marcada oblicuidad hacia uno de sus lados.
- Presión: 'Pressure9am', 'Pressure3pm'
    - Las distribuciones de las presiones responden en gran medida a una distribución normal por lo que en principio no es necesario realizar ninguna transformación sobre ellas
- Viento: 'WindGustSpeed', 'WindSpeed9am', 'WindSpeed3pm'
    - Las velocidades del viento presentan distribuciones con oblicuidades y largas colas hacia valores altos debido a la gran cantidad de outliers que presentan.
- Otros: 'Sunshine','Rainfall', 'Evaporation'
    - Las horas de sol presenta una distribución muy distinta a una distribución normal, lo cual en una primera instancia hay que tenerlo en cuenta para analizar una transformación de Box-Cox o Yeo-Johnson, al igual que las distribuciones de evaporación y lluvia que su distribución se asemeja a una logarítmica.

Correlación entre variables numéricas

De la matriz de correlaciones por el método de Pearson podemos obtener algunas relaciones interesantes que valen la pena remarcar que se enumeran a continuación:

- La temperatura mínima está correlacionada con la temperatura a las 9 am.
- La temperatura máxima está correlacionada con la temperatura a las 3 pm.
- Hay una obvia relación entre temperatura mínima y máxima.
- La evaporación está correlacionada con la temperatura máxima de manera positiva mientras que con la humedad de manera negativa.
- Las horas de sol tienen una relación negativa con la humedad.
- Las temperaturas tienen relaciones negativas con la humedad y las presiones medidas a las 9 am y a las 3 pm.

Estas relaciones mostradas desde la correlación, responden a relaciones lineales por el método utilizado.



Variables categóricas

Se analizarán las variables categóricas y su cardinalidad. A partir de los resultados hallados se propondrán distintas modificaciones en los features. Como bien ya se mencionó, el dataset cuenta con 7 variables categóricas:

	Location	WindGustDir	WindDir9am	WindDir3pm	Cloud9am	Cloud3pm	RainToday
0	Albury	W	W	WNW	8.0	NaN	No
1	Albury	WNW	NNW	WSW	NaN	NaN	No
2	Albury	WSW	W	WSW	NaN	2.0	No

Variable	Location	WindGustDir	WindDir9am	WindDir3pm	Cloud9am	Cloud3pm	RainToday
Cant Etiquetas	49	16	16	16	10	10	2

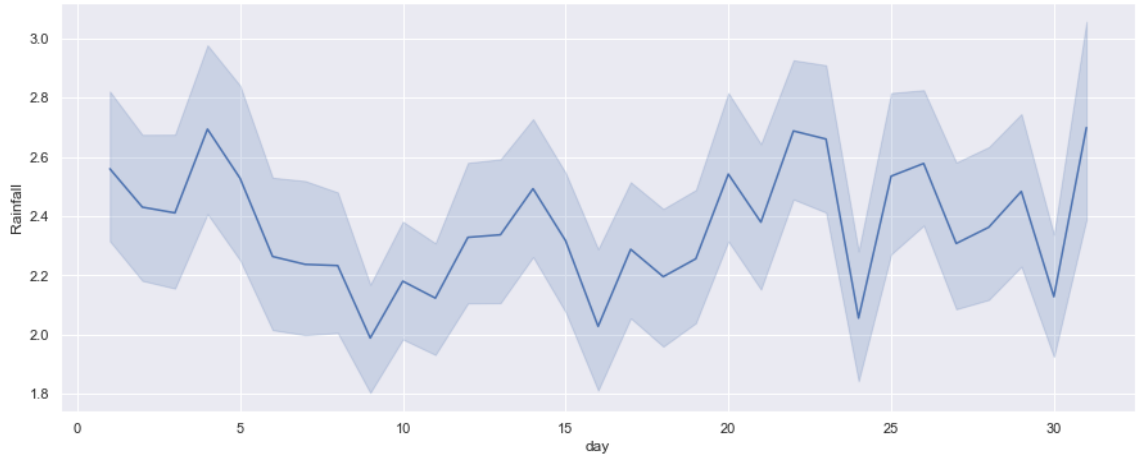
Se grafican las distintas variables para analizar la distribución de cada una de ellas, concluyendo que:

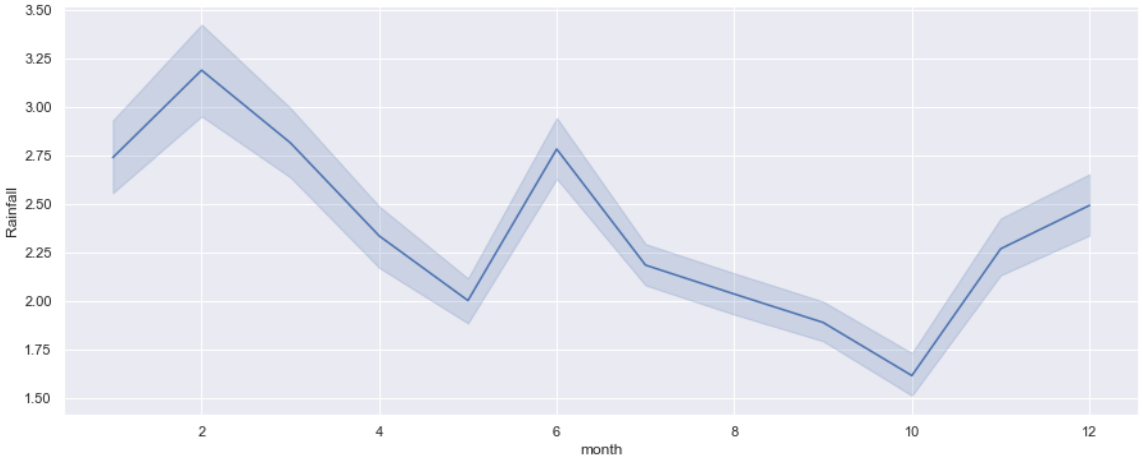
- La variable de Location está bastante bien balanceada con excepción de tres ciudades que poseen menor cantidad de datos. Esta también presenta una cantidad mucho mayor de etiquetas que el resto, lo que empeora la relación señal/ruido y puede generar algunos problemas a la hora de entrenar nuestro modelo como overfitting. Se propone para estas variables utilizar Label Encoding
- Las features de dirección de viento se encuentran bien balanceadas en 16 valores que indican los puntos cardinales por lo que no presenta etiquetas pocos frecuentes. Se propone para estas variables codificarlas en grados según la coordenada.
- La medición de cielo cubierto presenta una distribución bimodal con 8 etiquetas, debido a su medición en oktas. En principio, al contar con valores discretos y numéricos que hacen referencia al porcentaje de cielo cubierto no haría falta codificar.
- Finalmente tenemos la feature de si llovió el día en cuestión que se encuentra desbalanceada lo cual es un punto que debemos tener en cuenta. Se propone para esta variable utilizar una codificación de One Hot Encoding o Dummy Encoding

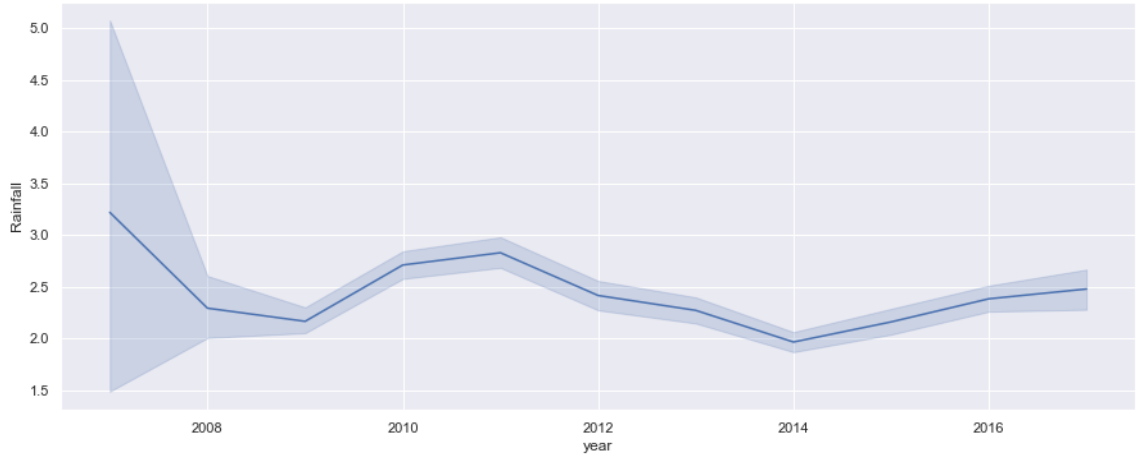
Variables compuestas

Como variable compuesta nos encontramos con la variable de fecha, compuesta por el año, el mes y el día. Analizaremos qué variables de año, mes y día conviene mantener. Se grafica para su análisis los mm de lluvia caídos vs día, mes y año.

Como el objetivo es predecir la posibilidad de que llueva al día siguiente, centraremos la información de la fecha en el dato del mes. Esto es debido a que el año en sí, no me genera un aporte sustancial a la posibilidad de que llueva en un determinado día y tampoco lo hace el día del mes en que llovió. Por otro lado, el dato del mes como podemos observar me da una noción de las épocas más lluviosas, lo cual puede resultar un dato de gran importancia para el modelo







Variable de salida

Al igual que lo visto en la feature de RainToday esta variable de salida se encuentra totalmente desbalanceada. Para esta variable de salida se utilizará One Hot Encoding o dummy encoding los valores 'Yes' se reemplazarán por 1 y los valores 'No' se reemplazará por 0.

2 - Esquema de validación de resultados

En esta etapa se divide el dataset en dos, un 80% de las muestras pertenecerán al set de entrenamiento y el 20% para el set de test.

3 - Limpieza y preparación de datos / ingeniería de features

Codificación de variables categóricas

Codificaremos algunas variables categóricas por una representación numérica.

- 'Location': Label/integer encoding, se codifican las 49 ciudades numéricamente con valores de 0 a 48
- 'WindGustDir','WindDir9am' y 'WindDir3pm': se codifica de la siguiente manera para que queden valores entre 0 y 1 y mantener la continuidad de la coordenada cardinal.
    - E -> (sin(0°)+1)/2
    - ENE -> (sin(22.5°)+1)/2
    - NE -> (sin(45°)+1)/2
    - NNE -> (sin(67.5°)+1)/2
    - N -> (sin(90°)+1)/2
    - NNW -> (sin(112.5°)+1)/2
    - NW -> (sin(135°)+1)/2
    - WNW -> (sin(157.5°)+1)/2
    - W -> (sin(180°)+1)/2
    - WSW -> (sin(202.5°)+1)/2
    - SW -> (sin(225°)+1)/2
    - SSW -> (sin(247.5°)+1)/2
    - S -> (sin(270°)+1)/2
    - SSE -> (sin(292.5°)+1)/2
    - SE -> (sin(315°)+1)/2
    - ESE -> (sin(337.5°)+1)/2
- 'RainToday' y 'RainTomorrow': One Hot Encoding - Se reemplazará el valor 'Yes' por '1' y el valor 'No' por '0'

Análisis de datos inválidos

Se calculan los datos faltantes o inválidos de cada variable obteniendo lo siguientes valores:

Cantidad de muestras: 145460

Sunshine:	Datos no NaN: 75625	Datos Nan: 69835	En%: 48 %
Evaporation:	Datos no NaN: 82670	Datos Nan: 62790	En%: 43.17 %
Cloud3pm:	Datos no NaN: 86102	Datos Nan: 59358	En%: 40.8 %
Cloud9am:	Datos no NaN: 89572	Datos Nan: 55888	En%: 38.42 %
Pressure9am:	Datos no NaN: 130395	Datos Nan: 15065	En%: 10.35 %
Pressure3pm:	Datos no NaN: 130432	Datos Nan: 15028	En%: 10.33 %
WindDir9am:	Datos no NaN: 134894	Datos Nan: 10566	En%: 7.26 %
WindGustDir:	Datos no NaN: 135134	Datos Nan: 10326	En%: 7.09 %
WindGustSpeed:	Datos no NaN: 135197	Datos Nan: 10263	En%: 7.05 %
Humidity3pm:	Datos no NaN: 140953	Datos Nan: 4507	En%: 3.09 %
WindDir3pm:	Datos no NaN: 141232	Datos Nan: 4228	En%: 2.9 %
Temp3pm:	Datos no NaN: 141851	Datos Nan: 3609	En%: 2.48 %
RainTomorrow:	Datos no NaN: 142193	Datos Nan: 3267	En%: 2.24 %
RainToday:	Datos no NaN: 142199	Datos Nan: 3261	En%: 2.24 %
Rainfall:	Datos no NaN: 142199	Datos Nan: 3261	En%: 2.24 %
WindSpeed3pm:	Datos no NaN: 142398	Datos Nan: 3062	En%: 2.1 %
Humidity9am:	Datos no NaN: 142806	Datos Nan: 2654	En%: 1.82 %
WindSpeed9am:	Datos no NaN: 143693	Datos Nan: 1767	En%: 1.21 %
Temp9am:	Datos no NaN: 143693	Datos Nan: 1767	En%: 1.21 %
MinTemp:	Datos no NaN: 143975	Datos Nan: 1485	En%: 1.02 %
MaxTemp:	Datos no NaN: 144199	Datos Nan: 1261	En%: 0.86 %
Location:	Datos no NaN: 145460	Datos Nan: 0	En%: 0.0 %
Date:	Datos no NaN: 145460	Datos Nan: 0	En%: 0.0 %

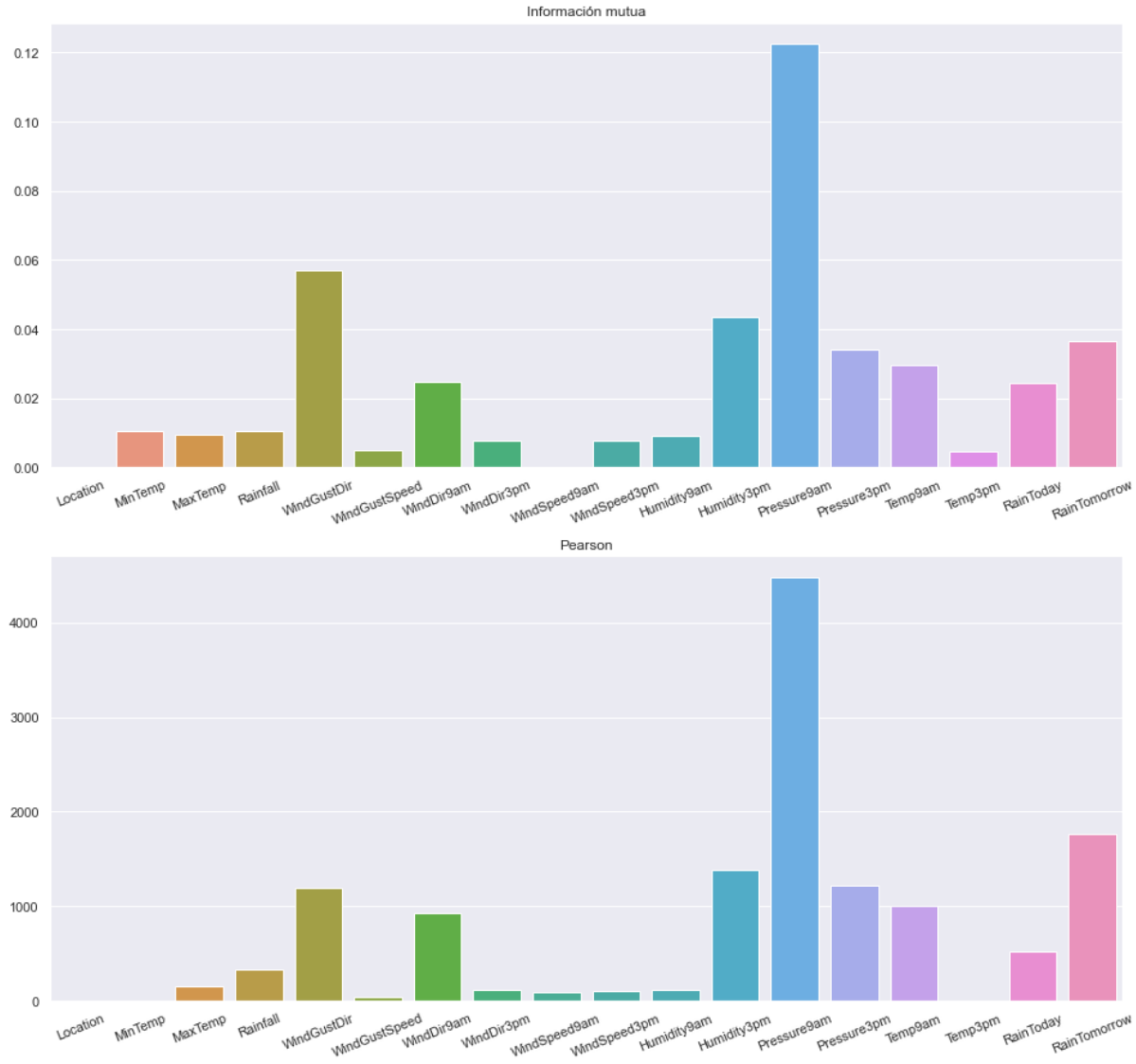
A primera vista hay que notar que hay un 2.25% de las muestras que no poseen datos de salida. Por lo tanto, estas muestras no son de utilidad para el entrenamiento o test de nuestro modelo y deben ser removidas.

Las features 'Sunshine', 'Evaporation', 'Cloud3pm' y 'Cloud9am' serán excluidas debido a su alto porcentaje de valores NaN.

Se procede con el armado de 2 datasets para armar dos sets de entrenamientos y sus respectivos set de test.

Caso 1

En un primer caso se eliminarán todas las muestras que tengan al menos un caso no válido. Esto nos deja con el 38.79 % del dataset original, manteniendo 56420 de las 145460 muestras originales.
Se grafican por método de pearson y por información mutua las relaciones entre las features y la salida:

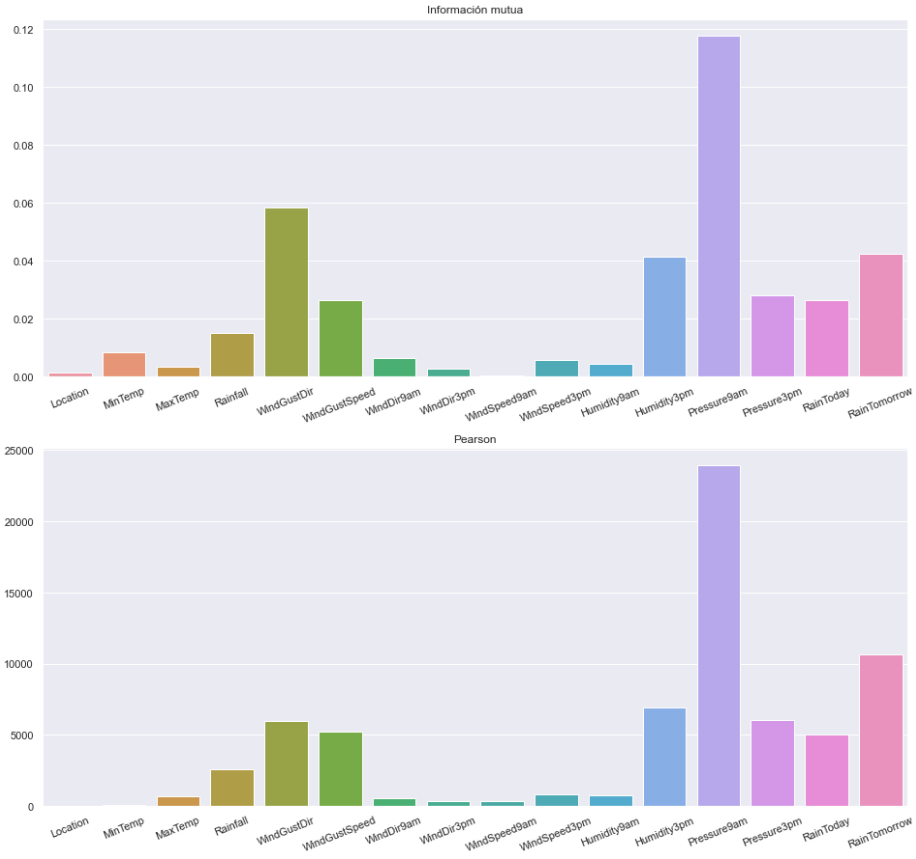


Caso 2

En este segundo caso utilizaremos varias técnicas para subsanar los datos no válidos de cada feature.
Pressure9am: Datos Nan: 14014 En%: 9.85 %

Pressure3pm:	Datos Nan: 13981	En%: 9.83 %
WindDir9am:	Datos Nan: 10013	En%: 7.04 %
WindGustDir:	Datos Nan: 9330	En%: 6.56 %
WindGustSpeed:	Datos Nan: 9270	En%: 6.51 %
WindDir3pm:	Datos Nan: 3778	En%: 2.65 %
Humidity3pm:	Datos Nan: 3610	En%: 2.53 %
Temp3pm:	Datos Nan: 2726	En%: 1.91 %
WindSpeed3pm:	Datos Nan: 2630	En%: 1.84 %
Humidity9am:	Datos Nan: 1774	En%: 1.24 %
RainToday:	Datos Nan: 1406	En%: 0.98 %
Rainfall:	Datos Nan: 1406	En%: 0.98 %
WindSpeed9am:	Datos Nan: 1348	En%: 0.94 %
Temp9am:	Datos Nan: 904	En%: 0.63 %
MinTemp:	Datos Nan: 637	En%: 0.44 %
MaxTemp:	Datos Nan: 322	En%: 0.22 %

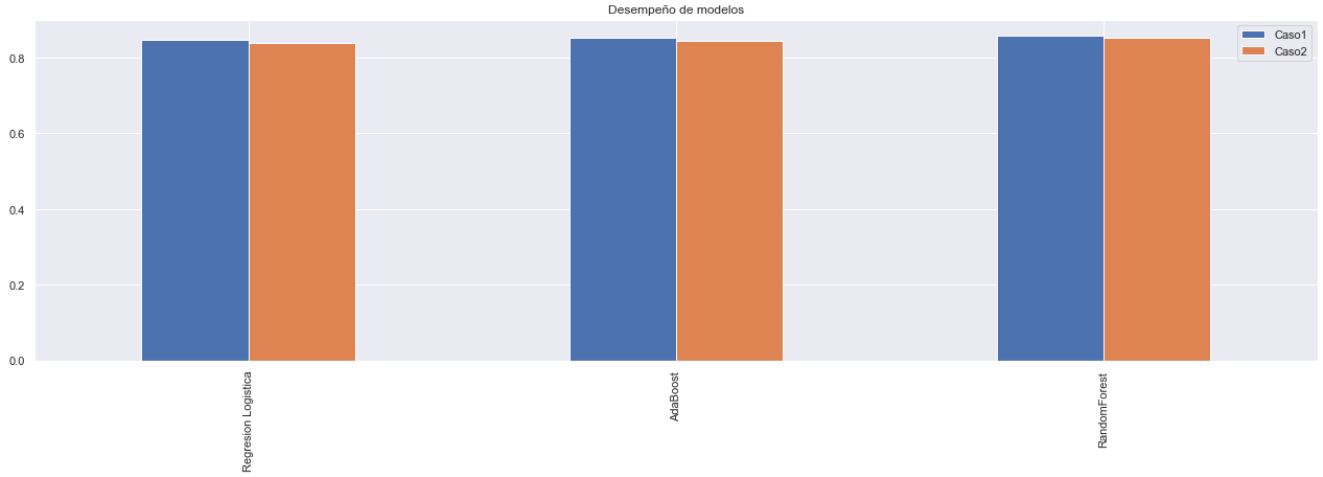
Comenzaremos con la feature "RainToday", esta es una feature categórica y binaria, donde alrededor del 1% de sus datos no son válidos. Considerando el bajo porcentaje se decidió borrar estas muestras con valores no válidos.
Como hemos visto en la sección anterior estas dos variables tienen gran correlación con MinTemp y MaxTemp respectivamente. Por ello, y porque tienen mayor cantidad de datos no válidos no utilizaremos estas features.
Para aquellos features con un porcentaje menor al 5% de Nan reemplazaremos estos valores por la mediana de su respectiva característica, ya que para distribuciones con oblicuidad es una mejor representación.
Las features restantes que poseen entre 5% y 10% de valores no válidos aplicaremos el método KNN para estimar los valores faltantes por semejanza a los más próximos
Se grafican por método de pearson y por información mutua las relaciones entre las features y la salida:



4 - Entrenamientos de modelos

Se aplicaron 3 modelos de aprendizaje automático, regresión lineal, clasificador adaboost y clasificador random forest, obteniendo los siguientes resultados:

	Caso1	Caso2
Regresion Logistica	0.848163	0.839406
AdaBoost	0.851944	0.844047
RandomForest	0.858029	0.853754



Como conclusión de los resultados obtenidos se encontró con que el modelo de Random Forest presenta una mayor precisión en la predicción de lluvia del día siguiente, al ser un modelo no lineal seguramente represente mucho mejor algunas relaciones algunos features y la salida que otros modelos como regresión lineal.

Por otro lado, se puede observar que el dataset al que se le eliminaron las muestras datos no válidos posee una mejor performance que el caso del modelo a la que se aplicó ingeniería de features. Esto se lo adjudico a que por ejemplo si vemos las gráficas de información mutua se encontro en el caso 1 que la salida depende mucho más de Temp9am (la cual fue eliminada en el caso 2) que de la TempMin.