TRABAJO FINAL
ANÁLISIS DE DATOS

NICOLÁS CECCHI

Introducción

Este documento presenta el Trabajo Integrador de la materia Análisis de Datos en el marco de la primera cohorte de la Especialización en Inteligencia Artificial de la Facultad de Ingeniería de la UBA.

El proyecto propuesto es el análisis y tratamiento de datos asociados al clima de Australia, donde se cuenta con columnas de mediciones de datos atmosféricos y una columna a predecir, que es un evento de tipo booleano indicando si al día siguiente lloverá o no, dados los datos del estado del tiempo el día anterior.

No forma parte del alcance del presente proyecto la obtención de los datos ni la construcción de la base de datos, esto viene de datos disponibles online propuestos por el cuerpo docente.

El trabajo si comprende:

- Análisis exploratorio de datos
- Tratamiento de datos
- Entrenamiento de modelos para predecir la variable objetivo

Análisis exploratorio inicial

En esta primera etapa se comienza a **conocer los datos**, se observan **tablas descriptivas** que incluyen: mínimo, máximo, promedio, Q25 y Q75, así como es desvío estándar. También se observan las cardinalidades de las variables categóricas.

De aquí se desprende que la información es bastante heterogénea. Por el lado de las **variables numéricas** hay muchas variaciones en las escalas y los niveles de concentración de los datos. Por el lado de las **variables categóricas**, se destaca "Location" que indica la ciudad o estación meteorológica de la medición.

La variable **Date,** se trata de una variable temporal, pero como los registros abarcan varios años, se decide transformarla en la codificación de las cuatro estaciones del año en Australia.

Se realiza un **primer análisis de nulos,** que luego será tenido en cuenta a la hora de transformar o descartar variables.

Respecto a la variable objetivo, se identifica que es una variable booleana indicando si al día siguiente llueve o no llueve, dados los datos recolectados. Esta variable tiene un 2.25% de los registros nulos, que se descartan teniendo en consideración que una "estrategia de imputación" no tiene sentido para esta variable ya que se trata de aquello que se quiere modelar posteriormente.

Esquema de validación de resultados

Utilizando la función `sklearn.preprocessing.train_test_split` se separa el conjunto de datos con el 80% en train y 20% en test.

Al momento de realizar la separación se tiene en cuenta el desbalance de los datos en la variable target, así como los efectos de las estaciones del año en las lluvias. Se define entonces el parámetro `Stratify` para que el muestreo aleatorio se realice manteniendo las proporciones de "evento/no-evento" y "estaciones" entre los dos data-sets generados, y así evitar que por aleatoriedad los dos conjuntos tengan proporciones que no representan el verdadero conjunto de datos.

Limpieza y preparación de datos

Creación de nuevas variables

Se crean variables artificiales a partir de las existentes. Teniendo en cuenta que hay mediciones en diferentes momentos del día (9am y 3pm), se construyen las variables de "variación intradiaria", como la diferencia entre la medición a las 3 pm y a las 9 am.

Existe también el registro de las temperaturas máximas y mínimas, cuya diferencia define la "amplitud térmica" de la jornada de medición.

Análisis gráfico de variables numéricas

Se realiza un análisis gráfico de las variables con **histogramas** y **box-plots**. Se adjunta un único gráfico de la variable "Humidity3pm" y algunas de las conclusiones que pueden sacarse de su análisis. El resto de los gráficos pueden consultarse en la notebook del proyecto.

Histogramas

Para construir el **histograma** de una variable numérica, Seaborn calcula los intervalos en que discretizar la variable y cuenta la cantidad de registros que caen dentro de cada bin, luego lo representa de manera gráfica.

Los histogramas se grafican diferenciando las clases "lluvia/no-lluvia" con colores diferentes.

Box-plot

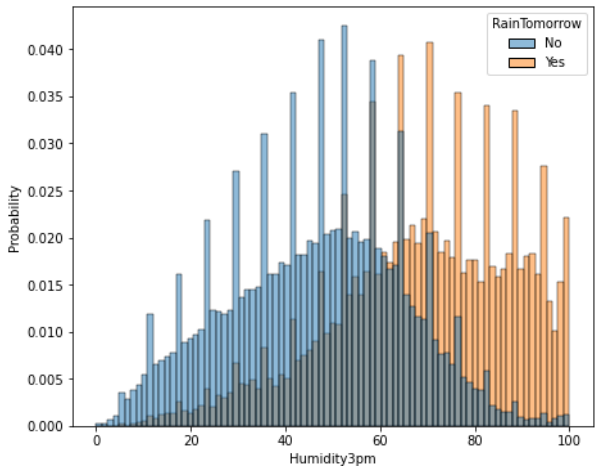
El **boxplot** presenta de manera gráfica algunos de los principales estadísticos que caracterizan una distribución: Q25, Mediana (Q50) y Q75, que construyen el cuerpo del gráfico. También están los "bigotes" o whiskers que se calculan como Q75(Q25) + (-) 1.5 IQR, donde IQR es el rango intercuartílico.

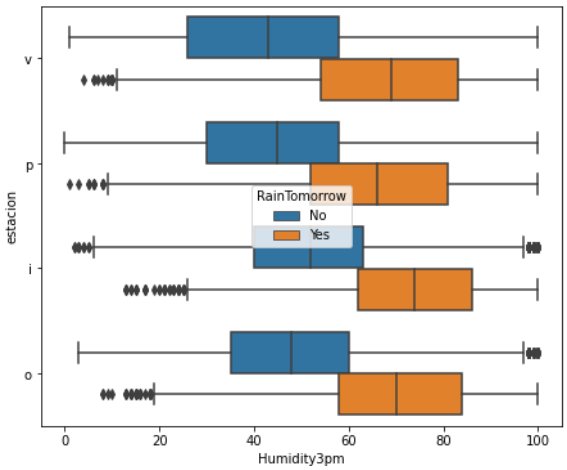
Teniendo en cuenta el split realizado en train-test, los gráficos también se generan diferenciando por "lluvia/no-lluvia" y por estaciones del año.

Ejemplo: Humidity3pm

En ambos gráficos se puede ver como los valores naranjas (Evento de lluvia al día siguiente) tienen una concentración hacia valores superiores. En el boxplot puede también observarse que el fenómeno ocurre durante las cuatro estaciones del año, donde el Q75 de "no llueve" tiene prácticamente el mismo valor que el Q25 de "Llueve al día siguiente". Es decir, el día anterior a un día lluvioso se caracteriza por tener una mayor humedad a las 3 pm.

Estos análisis son utilizados posteriormente en la notebook del proyecto para decidir sobre la transformación de una variable o si conviene directamente descartarla.





Correlación entre variables numéricas

Se realiza un análisis de la correlación entre las variables numéricas del dataset. Es deseable que el conjunto de variables predictoras no tengan altas correlaciones ya que se estaría usando información redundante. Tras el análisis de estas correlaciones, aquellos pares que tengan altos valores son analizados en mayor profundidad para decidir qué acciones tomar sobre cada variable. La matriz de correlaciones también es representada gráficamente con un heatmap.

Imputación de nulos

Se observa que muchas columnas poseen nulos, aunque en diferentes proporciones, por lo que se las estrategias adoptadas son las siguientes:

A la hora de imputar nulos a **variables numéricas** se utilizó el criterio siguiente:

- % nulos menor a 10% : Se imputa el promedio.
- % nulos mayor a 10%: Se realiza un análisis del caso particular (ver notebook).

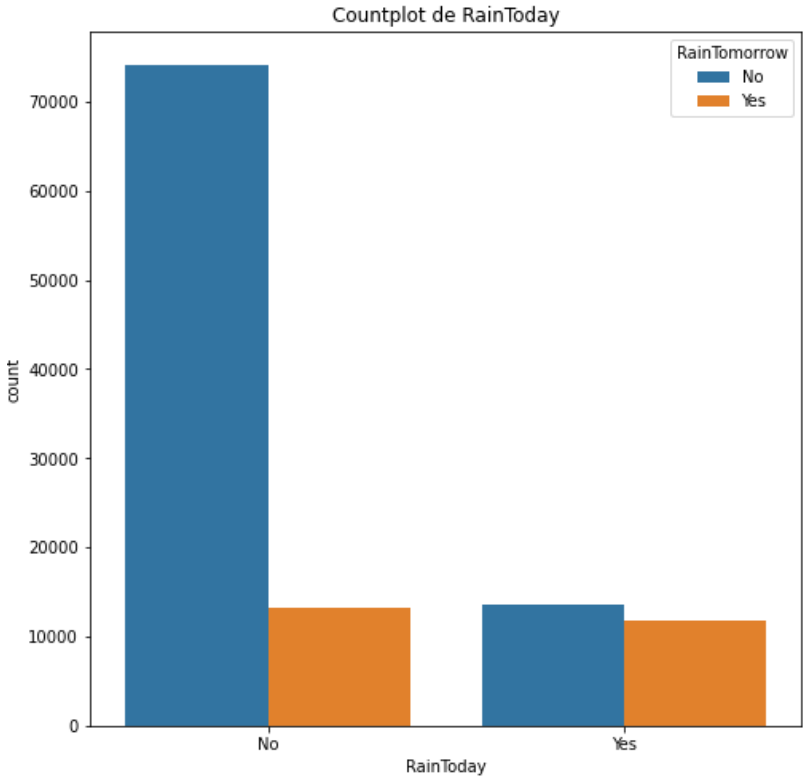
Por el lado de las **variables categóricas** se crea una clase que representa el dato faltante.

Transformación de variables numéricas

Las variables numéricas son transformadas utilizando el **RobustScaler** de sklearn. Esta técnica centra las variables por la mediana y estandariza por el IQR. Se prefiere esta metodología ya que los estadísticos utilizados para la transformación son **robustos frente a outliers**.

Análisis gráfico de variables categóricas

El análisis gráfico de variables categóricas se realiza con countplots que muestran la frecuencia de las diferentes clases. A modo de ejemplo, se muestra el gráfico de "RainToday", variable booleana que indica si llovió al día de las mediciones informadas.



Transformación de variables categóricas

Algunas de las variables categóricas presentan una **alta cardinalidad**, por lo que se decide transformarlas con alguna codificación que tenga mayor concentración, para facilitar el aprendizaje estadístico, dado que no se cuenta con una muestra suficientemente grande de datos.

Como ejemplo, las variables asociadas a la dirección del viento fueron transformadas tomando en cuenta la dirección principal. De esta manera se reduce de 16 a 4 categorías, más la categoría "missing".

Codificación de variables categóricas

One hot encoding

Previo al entrenamiento del modelo se utiliza la técnica de **OneHotEncoding** para las variables categóricas. La metodología utilizada genera N-1 columna (donde N es el

número de clases) ya que la clase N queda implícitamente representada cuando las N-1 columnas toman el valor 0.

Entrenamiento de modelos

Metodología

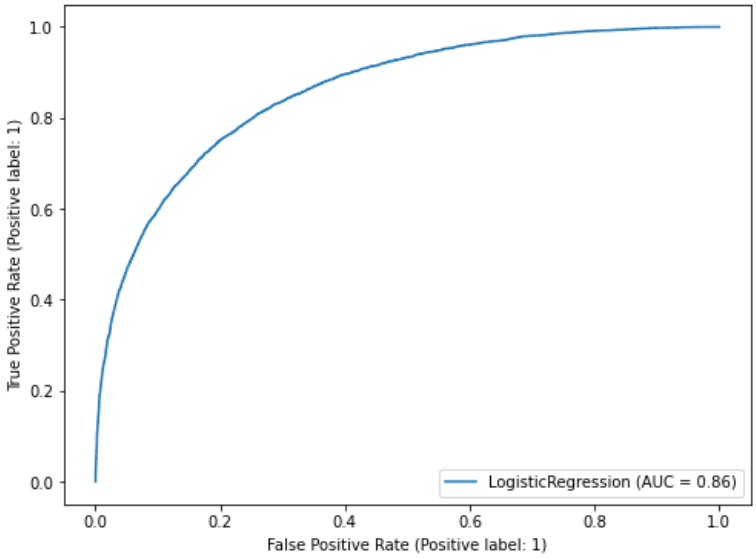
En una primera instancia se corren un conjunto de modelos diferentes con parámetros genéricos de la librería sklearn y midiendo la **balanced_accuracy** y **f1_score** como métricas de referencia. El modelo que mejor performance tuviera sobre los datos con este setup, se trabajará con mayor profundidad, realizando un procedimiento de **búsqueda de hiper parámetros**.

Resultados

El modelo que presenta mejor performance fue la **Regresión Logística** con las siguientes métricas:

Métrica	**Score**
balanced_accuracy_score	0.776
f1_score	0.612
roc_auc_score	0.776

Siguiendo la **metodología propuesta**, se realiza una búsqueda de hiper parámetros con la función RandomizedSearchCV y se obtiene la mejor regresión logística tras 50 iteraciones, la curva AUC asociada al mejor modelo se presenta a continuación:



Conclusiones

Tras la realización del trabajo puede destacarse la importancia de la ingeniería de features para la realización de un modelo de aprendizaje automático. Si bien muchas veces suele hacerse foco en los algoritmos de aprendizaje, gran cantidad de horas deben dedicarse a la obtención de datos de calidad y a su correcto tratamiento, ya que es la información a consumir por los algoritmos para encontrar los patrones relevantes que puedan existir en los datos para realizar la tarea objetivo.

Otro aspecto a resaltar es la necesidad de familiarizarse con los datos y con el problema. Conocimientos sobre las ciencias de la atmósfera y la geografía australiana, seguramente hubieran sido de gran ayuda para este proyecto.

Finalmente, la utilización de diferentes técnicas, numéricas y gráficas, aporta mucho valor al análisis de los mismos, tanto para tomar decisiones en el tratamiento de los mismos como para su presentación.