Análisis de Datos
Trabajo Integrador

Lautaro Delgado

18 de junio de 2021

1. Introducción

En el presente trabajo se realizó el análisis y desarrollo de un modelo de aprendizaje automático supervisado para un dataset de datos meteorológicos de Australia, donde el objetivo principal es predecir si lloverá al día siguiente o no. Por lo tanto, se trata de un problema de clasificación binaria.

Se realiza mayor enfoque en la etapa de preparación de los datos, análisis de las variables de entrada y de salida, transformación y codificación de las variables e ingeniería de nuevas features ya que es el core de la materia. No se ha realizado una elección detallada de los modelos ni del ajuste de sus hiper-parámetros por escapar al alcance principal.

2. Análisis Exploratorio Inicial

Se realizó una primera verificación del dataset, estudiando sus columnas, dimensiones, estadísticas principales de las variables numéricas y se determino el tipo de cada una de las variables. Luego se realizó un análisis separado para las variables de entrada, compuestas y de salida.

2.1. Variables de Entrada Numéricas

Para cada una de las variables numéricas se realizó un estudio de su distribución utilizando boxplots e histogramas (ver Figura 1). Asimismo, se estudió en la mayoría de los casos la relación entre las variables (ver Figura 2) .

2.2. Variables de Entrada Categóricas

En el caso de las variables categóricas se estudió principalmente su cardinalidad y la representatividad de cada clase (ver Figura 3). Para el caso de la ubicación, se obtuvo para cada ciudad, las coordenadas geográficas a modo de poder generar nuevas features y estudiar la relación entre la ubicación y el resto de las variables (ver Figura 4) .

2.3. Variables Compuestas

Se consideró como variable compuesta a la fecha y se analizó la composición de días, meses y años. Se estudió la relación temporal de algunas variables como la temperatura y las precipitaciones y se realizó por último un análisis conjunto de algunas variables y su evolución en el tiempo y en el espacio (ubicación), haciendo uso de heatmaps temporales.

Distribuciones de Sunshine, Cloud9am, Cloud3pm

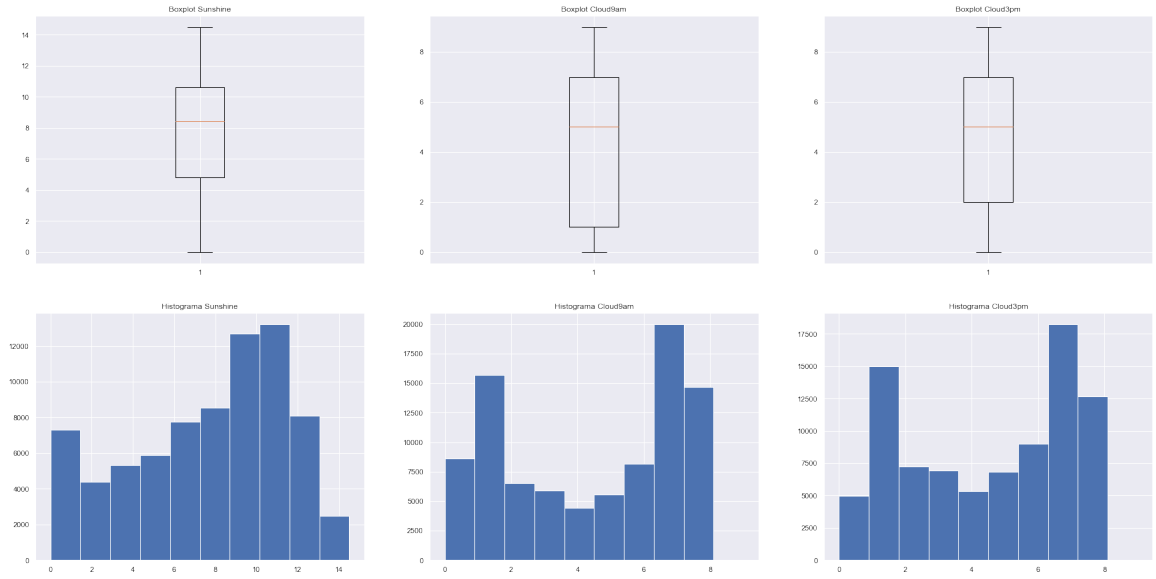


Figura 1: Gráficas de Distribución

2.4. Variable de Salida

Para la variables de salida se estudió principalmente su cardinalidad y se analizó si la mísma se encuentra o no balanceada, donde en éste caso hay un claro desbalance.

3. Esquema de Validación

Previo a realizar el split del dataset, se eliminaron las muestras que presentaban valores faltantes en la salida. Luego se separó el dataset en train y validation, teniendo en cuenta el balance de las clases de salida. Para ello, se utilizó stratified como parámetro en Scikit-Learn.

4. Ingeniería de Features

Se realizó en primer lugar un análisis de los valores faltantes de cada feature, buscando relaciones entre los mísmos y con la salida, para poder determinar si se trataba en cada caso de un proceso MAR, MCAR o MNAR. Se estudió también la cantidad total de muestras faltantes.

Una vez finalizado el análisis, se desarrollaron distintas estrategias de imputación de los valores faltantes, eliminando las muestras en el caso de la variable RainToday, comparando la imputación estadística con los métodos multivariados para el caso de las features numéricas y analizando crear una nueva clase o imputar por la moda en las features de dirección del viento.

Posterior a la imputación, se procedió a codificar las variables categóricas, donde para cada feature, se enssayaron distintas estrategias como one hot encoding, transformaciones numéricas, por grupos e incluso cíclicas. También se diseñaron nuevas features en el caso de la ubicación mediante distintos modelos de clusterización, y se agregó un indicador de ubicación insular y la distancia a la costa.

El siguiente pasó fue estudiar los outliers de los datasets resultantes, donde se empleó directamente

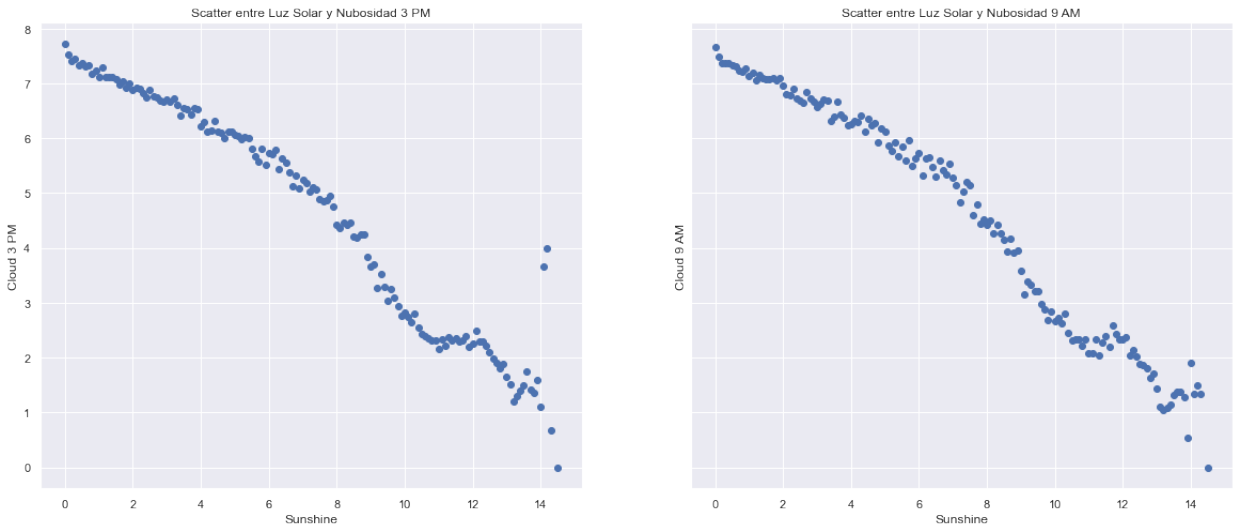


Figura 2: Relación entre variables

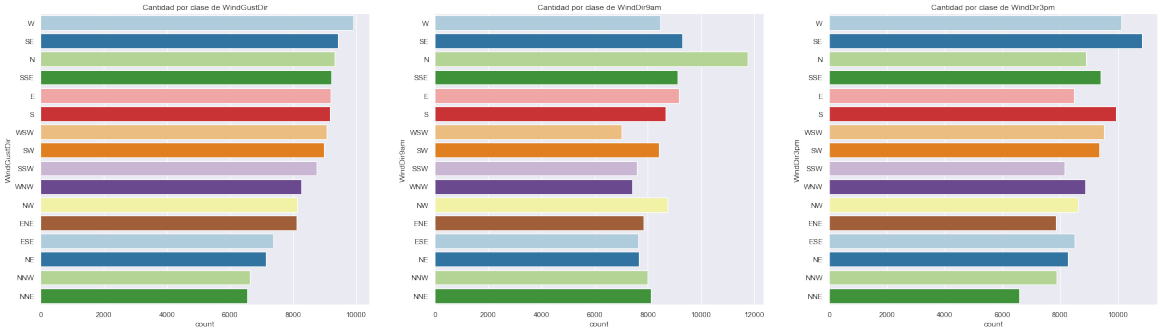


Figura 3: Cardinalidad y Representatividad

Local Outlier Factor por cuestiones de extensión. Dado que la cantidad de outliers no era elevada, se decidió eliminar directamente esas muestras.

El último paso de la sección consistió en el estudio de la relación entre las variables, haciendo uso de la correlación de Pearson, Kendall, el criterio de información mutua e incluso Cramer's V para las variables categóricas. A partir de ese análisis, se empezó a sacar conclusiones respecto de cuáles podían ser los mejores predictores para los modelos. Para complementar el análisis, se utilizaron técnicas de selección de features por modelo y de Recursive Feature Elimination, para dos tipos de modelos de familias distintas como es el caso de Regresión Logística y Random Forest.

Se concluyó la sección almacenando cada uno de los datasets y transformaciones realizadas. Se deja para trabajo futuro, el diseño de clases personalizadas que hereden de Column Transformer y BaseEstimator para poder utilizar las operaciones diseñadas, en un pipeline de Scikit-Learn que agilice la experimentación.

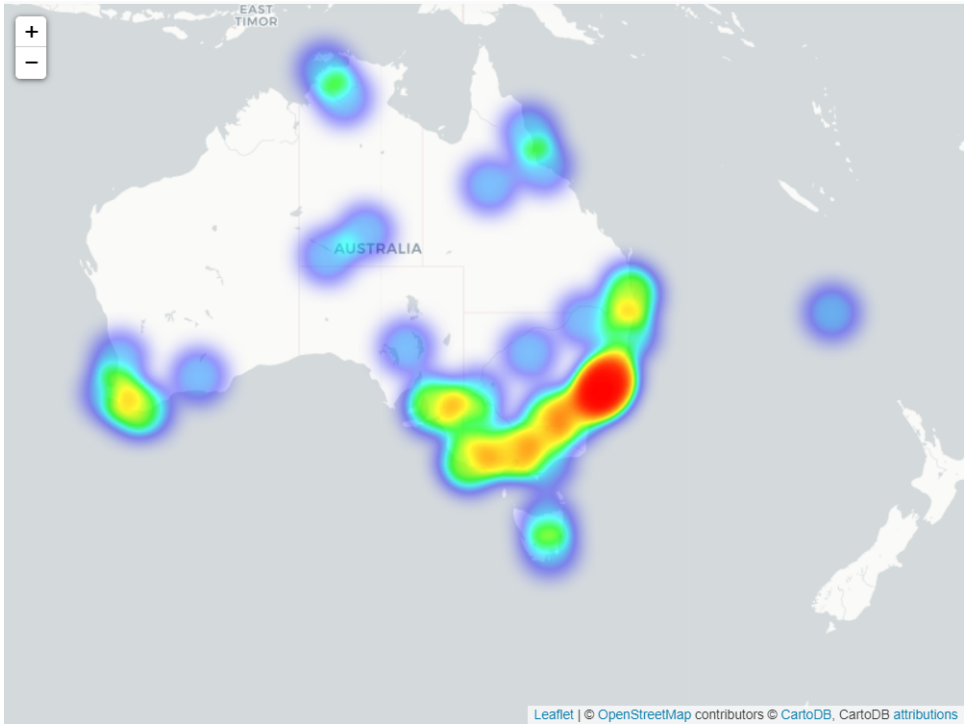


Figura 4: Mapa de precipitaciones

5. Entrenamiento de Modelos

Para el entrenamiento de los modelos se debe tener en cuenta que se está trabajando con un dataset desbalanceado. Existen varias técnicas que se pueden emplear como SMOTE (aplicada en éste caso), oversampling, undersampling e incluso se pueden utilizar modelos donde se balancea la función de costo (Regresión Logística por mencionar alguno) o modelos One Class. Por una cuestión de extensión y de que escapa al alcance del trabajo, se decidió emplear directamente SMOTE y comparar sus resultados.

Dado que se tiene una gran combinación de features, técnicas de imputación, entre otros, se explorarán únicamente algunas de ellas.

Para poder comparar los modelos, es necesario establecer la métrica a utilizar. Generalmente con datasets desbalanceados, no es deseable utilizar Accuracy. Existe literatura especializada sobre la elección de métricas en éstos casos. Para simplificar, se decidió utilizar F1 que es frecuente en la práctica.

Se comenzará con el dataset donde las direcciones del viento se imputaron por la moda, y se ensayarán, para los modelos de Regresión Logística, KNN, Random Forest, Árboles de Decisión y AdaBoost, distintas configuraciones temporales (ver Figura 5). Una vez elegida la mejor combinación, se analizarán los distintos clusters de ubicación. Para la mejor configuración encontrada, se ensayarán el resto de los datasets de imputación (dirección del viento con nueva clase y eliminación de muestras con Nans). Los tres casos se analizan luego para el dataset balanceado con SMOTE.

Dado que se obtienen buenos resultados con SMOTE y con el datasets donde se eliminaron los valores faltantes de dirección del viento, se utiliza ésta configuración para evaluar el dataset de validación. Previo a ello, se deben realizar todas las operaciones de transformación y codificación sobre el dataset

de test que se realizaron para el de train. En la Figura 6, se pueden observar los resultados obtenidos para Random Forest.

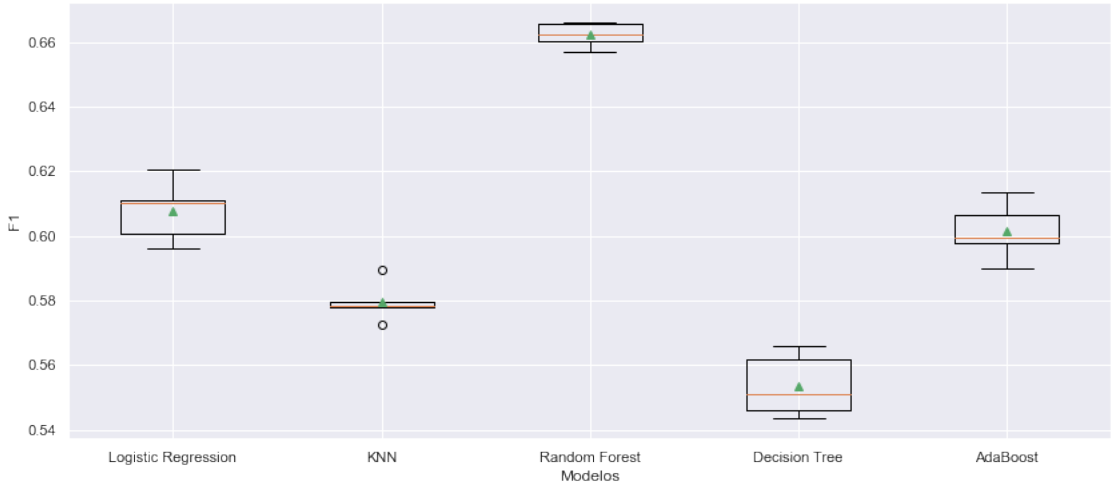
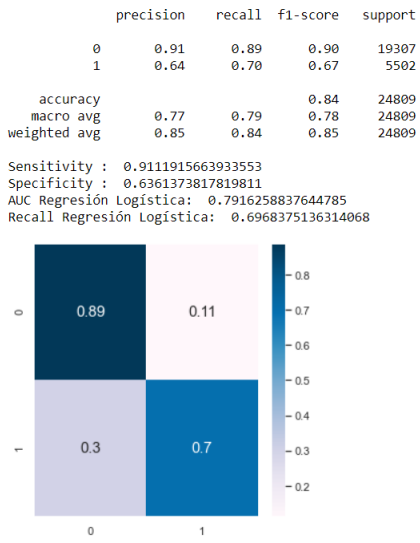


Figura 5: Comparación de Modelos



	precision	recall	f1-score	support
0	0.91	0.89	0.90	19307
1	0.64	0.70	0.67	5502
accuracy			0.84	24809
macro avg	0.77	0.79	0.78	24809
weighted avg	0.85	0.84	0.85	24809

Sensitivity : 0.9111915663933553
Specificity : 0.6361373817819811
AUC Regresión Logística: 0.7916258837644785
Recall Regresión Logística: 0.6968375136314068

Figura 6: Resultados para Random Forest