FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Informe de Avance del Trabajo Final de la Carrera de Especialización en Inteligencia Artificial

Análisis de datos

Trabajo final integrador

Autor

Ing. Hernán Contigiani

2021

Tabla de contenido

Revisión	Cambios realizados	Fecha
1.0	Creación del documento	16/06/2021

1. Breve descripción del desafío

Se propone realizar el ciclo completo del desarrollo de un modelo de aprendizaje automático supervisado basado en el dataset de pronóstico de lluvia tomado de diferentes estaciones meteorológicas de Australia.

El objetivo es predecir si lloverá o no al día siguiente (variable RainTomorrow), en función datos meteorológicos del día actual. Para ello se debe realizar el análisis de datos bajo la siguientes premisas:

1. Análisis exploratorio inicial: identificar los tipos de datos y realizar un análisis de la naturaleza de los features de entrada y la variable de salida.
2. Limpieza y preparación de datos: analizar la naturaleza de los datos faltantes y definir estrategias de imputación de datos.
3. Entrenamiento de modelos: armar cadena de procesamiento con diferentes estrategias de imputación y limpieza de datos antes definidas con diferentes modelos de clasificación
4. Evaluación de resultados: evaluar cuáles fueron las mejores estrategias y obtener conclusiones.

En el link a continuación podrá encontrar el notebook con el detalle de la resolución a este desafío:

Github notebook

2. Análisis exploratorio

A continuación se detallan alguna de las conclusiones obtenidas de la observación del dataset:

- El dataset se compone por más de 14000 filas y 23 columnas, por lo tanto hay una gran cantidad de datos para resolver el problema planteado.
- Se encontraron 3 tipos de datos entre los features de entrada:
    - compuestos → fecha
    - categóricos → locación y los relativos a la dirección del viento
    - numéricos → todas las demás columnas
- La variable de salida es una variable categórica binaria, por lo que el modelo de inteligencia artificial que se utilizará para resolver esta problemática es un clasificador binario (regresión logística, KNN, RandomForest, etc)
- **IMPORTANTE:** el dataset no está balanceado, el ~75% de los datos están relacionados a pronósticos sin lluvia.

Las variables categóricas mencionadas antes se encuentran bien balanceadas. El único detalle a tener en cuenta es que la locación tiene casi 50 clases, como se puede observar en la Figura 1.

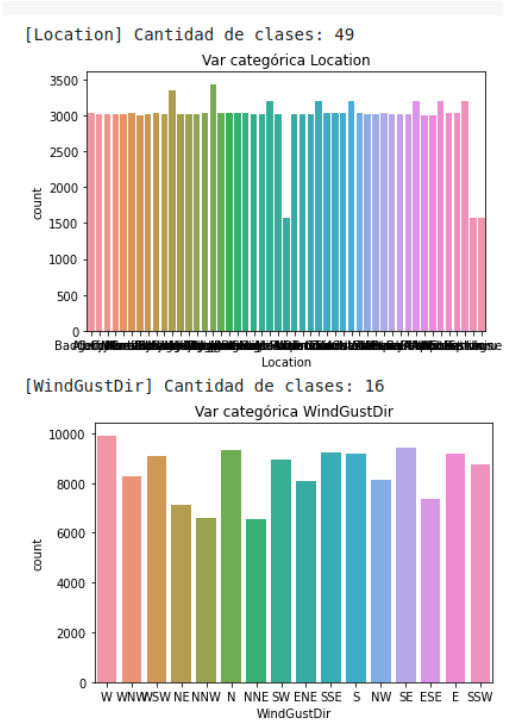


Figura 1

Se analizó cómo descomponer la columna “fecha”. En la Figura 2 se puede observar que el día o el año no aportan demasiado valor, por lo que solo se aprovecha el “mes” de la columna fecha:

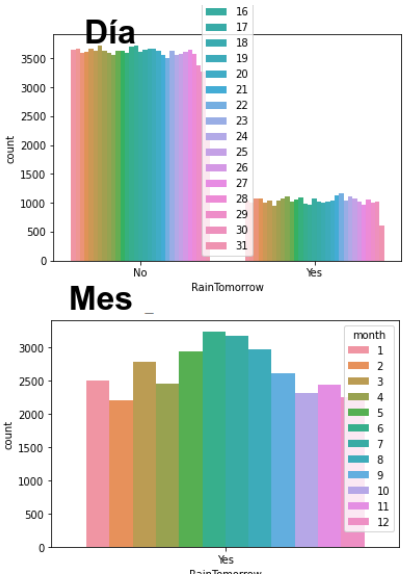


Figura 2

2. Limpieza y preparación de datos

Se analizó el porcentaje de datos faltantes en el dataset:

	column_name	sum_missing	percent_missing
Location	Location	0	0.000000
MinTemp	MinTemp	1025	1.006659
MaxTemp	MaxTemp	863	0.847558
Rainfall	Rainfall	2282	2.241166
Evaporation	Evaporation	43955	43.168470
Sunshine	Sunshine	48920	48.044627
WindGustDir	WindGustDir	7186	7.057414
WindGustSpeed	WindGustSpeed	7145	7.017148
WindDir9am	WindDir9am	7396	7.263656
WindDir3pm	WindDir3pm	2964	2.910962
WindSpeed9am	WindSpeed9am	1237	1.214865
WindSpeed3pm	WindSpeed3pm	2138	2.099743
Humidity9am	Humidity9am	1876	1.842431
Humidity3pm	Humidity3pm	3179	3.122115
Pressure9am	Pressure9am	10550	10.361219
Pressure3pm	Pressure3pm	10536	10.347469
Cloud9am	Cloud9am	39133	38.432755
Cloud3pm	Cloud3pm	41548	40.804541
Temp9am	Temp9am	1240	1.217811
Temp3pm	Temp3pm	2539	2.493567
RainToday	RainToday	2282	2.241166
RainTomorrow	RainTomorrow	2269	2.228399
month	month	0	0.000000

- Para todas aquellas columnas cuya cantidad de faltantes fuera inferior al 5% se eliminaron aquellas filas (por ejemplo WinDir3pm).
- Para todas las demás columnas se examinó por separado: las columnas con un porcentaje de cantidad de faltantes mayor al 20% (por ejemplo Cloud9am) y las demás (por ejemplo Pressure9am).

Para ambos casos de columnas faltantes se llegó a la conclusión que no existe un motivo aparente de porque faltan aquellos datos, ya que la cantidad de datos faltantes corresponde con la distribución de datos en el dataset (es decir que proporcionalmente faltan tantos datos para aquellos casos donde llueve para los que no llueve):

```
Analisis para Evaporation por RainToday
RainTomorrow 0: 30475 | 77.59%
RainTomorrow 1: 8801 | 22.41%
Analisis para Sunshine por RainToday
RainTomorrow 0: 28562 | 77.77%
RainTomorrow 1: 8164 | 22.23%
Analisis para Cloud9am por RainToday
RainTomorrow 0: 23352 | 80.59%
RainTomorrow 1: 5625 | 19.41%
Analisis para Cloud3pm por RainToday
RainTomorrow 0: 23209 | 80.45%
RainTomorrow 1: 5640 | 19.55%
```

Tampoco se encontró una relación en común entre los datos faltantes en sí:

Dicho lo anterior se concluyó que los datos faltantes corresponden a datos tipo **MCAR** (totalmente aleatorio) y por lo tanto se utilizarán estrategias de imputación alineadas con dicho comportamiento (imputación por la mediana o más frecuente, imputación por KNN, etc).

Por último se analizó si podría descartarse una columna/feature que estuviera poco relacionada con la salida o muy representada por otra columna. En definitiva no se halló la suficiente evidencia para eliminar columnas.

3. Entrenamiento de los modelos

Estrategias de imputación y transformación:

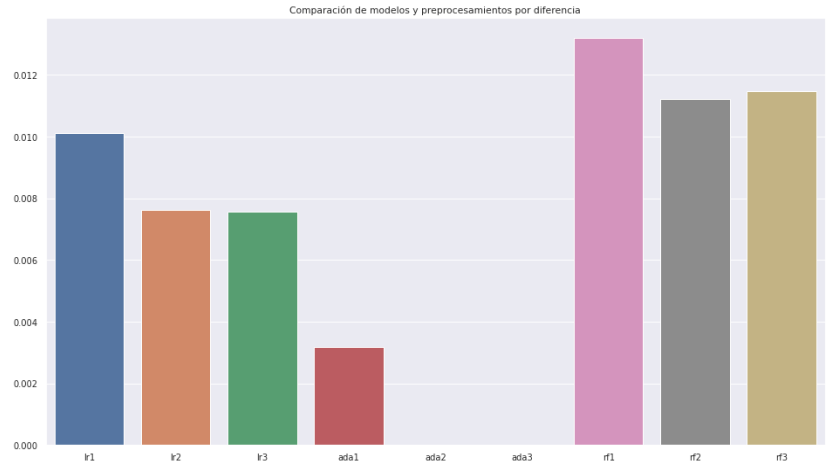
- Descartar todas las filas con datos faltantes en sus columnas
- Reemplazar por la mediana o el valor más frecuente (según si es una variable numérica o categórica)
- Aplicar una transformación a distribución normal

Modelos seleccionados:

- LogisticRegression
- RandomForestClassifier
- AdaBoostClassifier

IMPORTANTE: Se descartó el uso de KNN tanto para imputación como modelo de clasificación dado que los tiempos de cómputo pasan de pocos minutos a horas

4. Resultados



- En cuanto a los modelos seleccionados la mejor performance se obtuvo con RandomForest, era de esperarse ya que se trata de un modelo tipo ensamble muy potente.
- En cuanto a la estrategia de imputación y transformación la mejor performance se obtuvo eliminando todas aquellas filas con datos. Es coherente el resultado por los siguientes motivos:
    - Había 4 o 5 columnas que tenían casi la mitad de las filas con faltantes, por lo que realizar cualquier tipo de imputación representaría afectar en grandes rasgos la distribución de ese feature. La mitad de esas columnas estaban fuertemente relacionadas con la salida.
    - Los datos faltantes eran MCAR (aleatorios), por lo que era menor el riesgo de al eliminarlos producir algún tipo de desbalance en los datos.