Prediciendo la probabilidad de lluvias en Australia

Un trabajo de:

Ezequiel Scordamaglia, Santiago González Achaval y
Federico Glancszpigel

Aprendizaje Maquina I

.UBA fiuba
FACULTAD DE INGENIERÍA

Carrera de Especialización en Inteligencia Artificial

Objetivo del Trabajo

El objetivo de este trabajo es **seleccionar el mejor modelo** para **predecir si lloverá en Australia en el día de mañana**. Es decir, estamos frente a un problema de clasificación supervisado. La idea es únicamente estimar la probabilidad de lluvia de mañana, no la cantidad de precipitaciones en milímetros (mm) que caerá. Para esta tarea se utilizo el dataset ***Rain in Australia*** de Kaggle.

Descripción del Dataset

Utilizamos el dataset ***Rain in Australia*** de Kaggle:
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package?select=weatherAUS.csv

Este dataset cuenta con **23 variables**

- 22 variables independientes. Algunas de las mas relevantes son:
  - *Date*: La fecha de observación
  - *Location*: El nombre común de la ubicación de la estación meteorológica.
  - *MinTemp*: La temperatura mínima en grados centígrados
  - *MaxTemp*: La temperatura máxima en grados centígrados
  - *Rainfall*: La cantidad de lluvia registrada para el día en mm

- 1 variable dependiente o target llamada “RainTomorrow” que vale 1 si si la precipitación (en mm) del día siguiente excede 1 mm, de lo contrario 0.

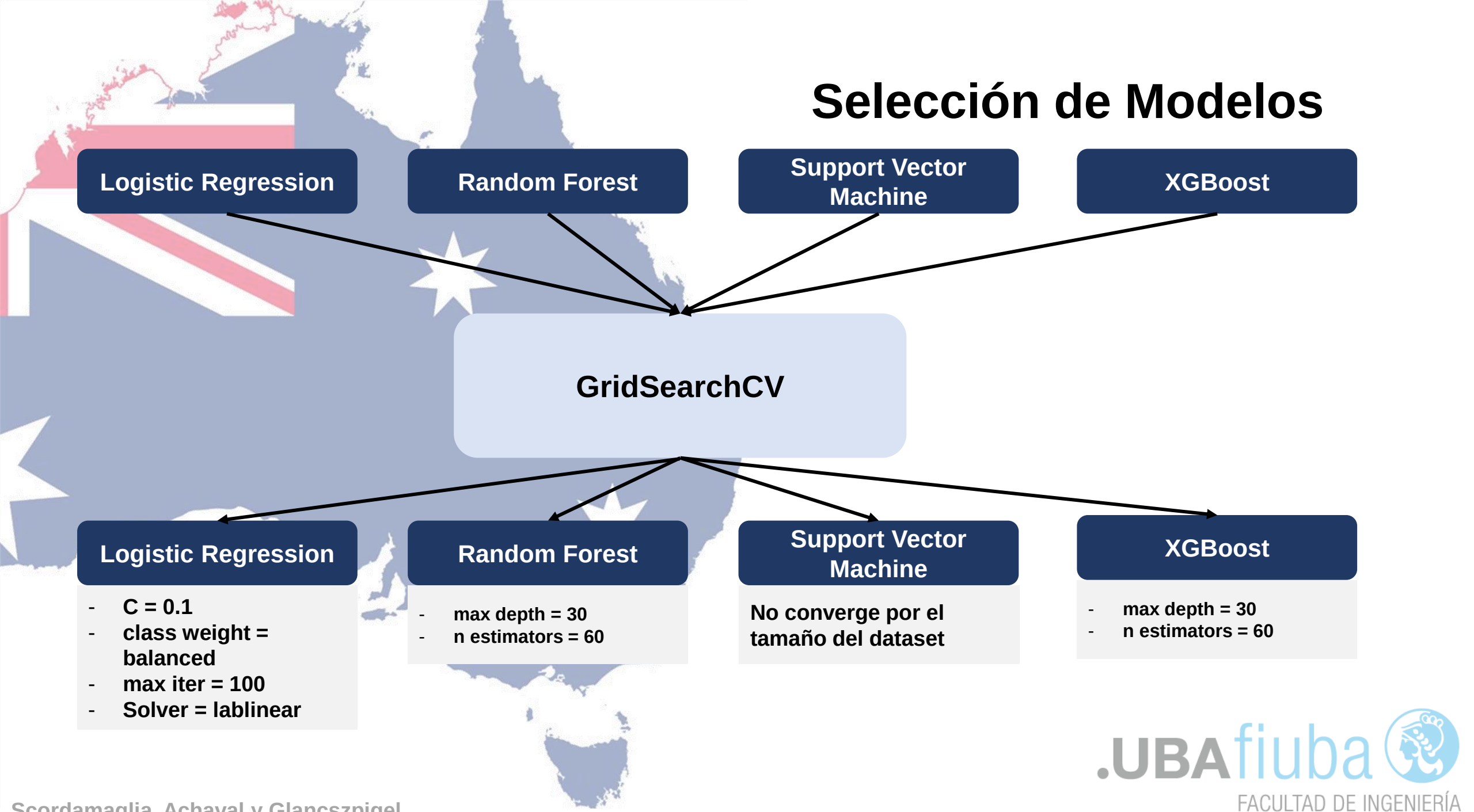
El dataset contiene **145,000 observaciones** ordenadas por fecha. Se utilizo el 75% del dataset para entrenamiento y el 25% restante para testeo.

Pre-procesamiento de datos

Para el Pre-procesamiento de los datos se realizaron las siguientes tareas:

- Eliminación de columnas con gran cantidad de valores nulos (Evaporation, Sunshine, Cloud9am, Cloud3pm)
- Eliminación de columnas que tenían gran correlación con otras (RainToday, Temp9am, Tem3pm)
- Eliminación de registros con mas de 4 valores faltantes.
- Eliminación de registros con variable objetivo nula.
- División de dataset en train (75%) y test (25%), usando Stratify = True para mantener la proporción de los datos
- Capping de outliers en columnas WindGustSpeed, WindSpeed9am, WindSpeed3pm, Rainfall.
- Imputación de valores faltantes (Usando media, mediana y MICE)
- Transformación de Yeo-Jhonson para normalizar las distribuciones
- Encoding:
    - RainTomorrow a booleana
    - Dirección del viento a numérico (de 0 a 360)
    - Location a dummies
    - Fecha a tres columnas numéricas (día, mes , año)
- Balanceo del dataset usando técnicas como SMOTE, KMeansSMOTE y RandomUnderSampler
- Escalado de los datos usando StandarScaler

Selección de Modelos



Resultados de Performance

Modelo	Tecnica de sampleo	Precision	Recall	F1 Score
Logistic Regression	Standardization	0.4	0.88	0.55
	SMOTE (1) + No undersampling + No standardization	0.57	0.66	0.61
	SMOTE (0.5) + No undersampling	0.68	0.52	0.59
	SMOTE (0.5) + Standardization	0.56	0.69	0.62
	KMeansSMOTE (0.5) + Random undersampling (0.8) + Standardization	0.43	0.86	0.57
Random Forest	Standardization	0.47	0.8	0.6
	SMOTE (1) + No undersampling + No standardization	0.69	0.58	0.63
	SMOTE (0.5) + No undersampling	0.71	0.54	0.62
	SMOTE (0.5) + Standardization	0.55	0.73	0.63
	KMeansSMOTE (0.5) + Random undersampling (0.8) + Standardization	0.47	0.8	0.6
XGBoost	Standardization	0.75	0.55	0.63
	SMOTE (1) + No undersampling + No standardization	0.76	0.55	**0.64**
	SMOTE (0.5) + No undersampling	0.5	0.79	0.61
	SMOTE (0.5) + Standardization	0.26	0.95	0.41
	KMeansSMOTE (0.5) + Random undersampling (0.8) + Standardization	0.38	0.91	0.54

La técnica de sampleo SMOTE (1) sin estandarización de los datos de entrada y sin undersampling es la que genera el máximo F1 score promedio

→ Modelo que maximiza el F1

→ Modelo que maximiza la precisión

→ Modelo que maximiza el recall

Aplicación de AutoML

Utilizando la técnica de sampleo SMOTE(1) + No undersampling + No standardization

Modelo	Accuracy	AUC	Recall	Precision	F1 score	Kappa	MCC	TT (Sec)
Light Gradient Boosting Machine	0.86	0.89	0.54	0.76	0.63	0.55	0.56	1.54
Random Forest Classifier	0.86	0.89	0.50	0.78	0.61	0.53	0.55	12.20
Extra Trees Classifier	0.86	0.89	0.49	0.79	0.60	0.52	0.55	13.57
Gradient Boosting Classifier	0.85	0.87	0.50	0.75	0.60	0.51	0.53	22.12
Linear Discriminant Analysis	0.85	0.86	0.52	0.72	0.60	0.51	0.52	0.82
Ridge Classifier	0.85	0.00	0.46	0.77	0.57	0.49	0.51	0.42
Ada Boost Classifier	0.84	0.86	0.49	0.72	0.58	0.49	0.51	5.09
Naive Bayes	0.84	0.85	0.50	0.68	0.58	0.48	0.49	0.24
Logistic Regression	0.83	0.83	0.44	0.70	0.54	0.44	0.46	2.60
Decision Tree Classifier	0.79	0.70	0.54	0.53	0.53	0.40	0.40	1.80
K Neighbors Classifier	0.78	0.70	0.27	0.52	0.35	0.24	0.26	5.69
Dummy Classifier	0.78	0.50	0.00	0.00	0.00	0.00	0.00	0.15
Quadratic Discriminant Analysis	0.61	0.73	0.74	0.33	0.46	0.22	0.26	0.49
SVM - Linear Kernel	0.61	0.00	0.49	0.51	0.32	0.15	0.22	8.10

Tanto Light GBM como Extra Trees Classifier son modelos de ensamble de arboles de decisión, ***de la misma familia que XGBoost***

Aplicación de AutoML

Por ultimo, hacemos un *tuning* de hiper-parametros del LightGBM (mejor modelo). Los siguientes resultados fueron obtenidos

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.85	0.88	0.57	0.71	0.63	0.54	0.55
1	0.84	0.86	0.56	0.67	0.61	0.51	0.51
2	0.85	0.87	0.58	0.69	0.63	0.54	0.54
3	0.84	0.87	0.57	0.67	0.62	0.52	0.52
4	0.85	0.87	0.58	0.69	0.63	0.53	0.53
5	0.85	0.88	0.56	0.70	0.62	0.53	0.54
6	0.85	0.88	0.59	0.71	0.64	0.55	0.56
7	0.85	0.87	0.58	0.69	0.63	0.53	0.54
8	0.85	0.88	0.60	0.70	0.64	0.55	0.56
9	0.85	0.88	0.58	0.70	0.63	0.54	0.54
Mean	0.85	0.87	0.58	0.69	0.63	0.53	0.54
Std	0.00	0.01	0.01	0.01	0.01	0.01	0.01

Comparativa final de Modelos

Modelo	Precision	Recall	F1 Score
Logistic Regression	0.57	0.66	0.61
Random Forest	0.69	0.58	0.63
XGBoost	0.76	0.55	**0.64**
Light GBM	0.69	0.58	0.63

Luego de un análisis detallado, llegamos a la conclusión que el mejor modelo para predecir la probabilidad de lluvia en Australia es XGBoost, para el dataset elegido.

¡Muchas Gracias!

Un trabajo de:
Ezequiel Scordamaglia, Santiago González Achaval y
Federico Glancszpigel

Aprendizaje Maquina I

.UBA fiuba
FACULTAD DE INGENIERÍA

Carrera de Especialización en Inteligencia Artificial