UNIVERSIDAD
PANAMERICANA

Proyecto Machine Learning
Clasificación

Esteban Viniegra Pérez Olagaray
Santiago Valdez Bocardo

Dataset: Weather

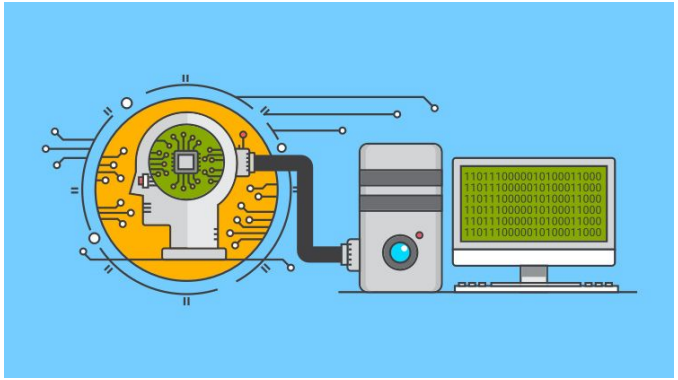
- Observaciones Diarias del Clima
- Datos y métricas de clima en Brasil
  - Lluvia hoy/mañana?
  - Horas de Sol
  - Temperatura mínima/máxima
- Generación de modelos de clasificación para predecir el clima en Brasil
- Crear un modelo de predicción para determinar el clima

Tipo de Modelo y Aprendizaje

- La naturaleza de la predicción climática implica asignar una categoría específica (clase) a una serie de variables meteorológicas en un momento dado. Por ejemplo, clasificar si lloverá o no, si habrá temperaturas extremas, etc.

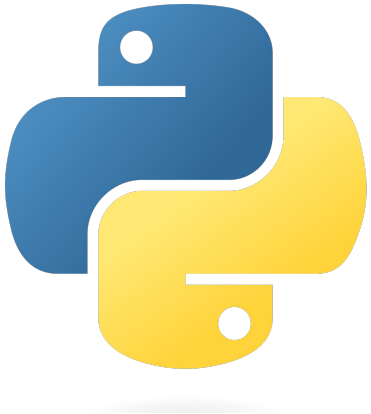
- Para entrenar modelos de clasificación, se necesitan datos históricos con etiquetas precisas que indiquen las condiciones climáticas reales.

- Aprendizaje supervisado: ya se cuenta con un conjunto de datos históricos (dataset “weather.csv”) que incluyen las condiciones climáticas observadas junto con las etiquetas correspondientes (por ejemplo, si llovió o no).



Transformación Dataset

- Imputación de valores nulos
- Conversión de fechas
- Eliminación de duplicados
- Codificación de variables categóricas
- Eliminación de columnas innecesarias
- Normalización de variables numéricas
- Eliminar espacios, acentos, ñ, etc.



Fecha	Ubicación	Temperatura_Mínima	Temperatura_Máxima	Precipitación	Evaporación	Horas_de_Sol	Dirección_ráfaga_viento	Velocidad_ráfaga_viento	Dirección_viento_9am	Dirección_viento_3pm	Velocidad_viento_9am
01/12/2008	ufepam	13.4	22.9	0.6			W	44	W	WNW	20
02/12/2008	ufepam	7.4	25.1	0			WNW	44	NNW	WSW	4
03/12/2008	ufepam	12.9	25.7	0			WSW	46	W	WSW	19
04/12/2008	ufepam	9.2	28	0			NE	24	SE	E	11
05/12/2008	ufepam	17.5	32.3	1			W	41	ENE	NW	7
06/12/2008	ufepam	14.6	29.7	0.2			WNW	56	W	W	19
07/12/2008	ufepam	14.3	25	0			W	50	SW	W	20
08/12/2008	ufepam	7.7	26.7	0			W	35	SSE	W	6
09/12/2008	ufepam	9.7	31.9	0			NNW	80	SE	NW	7
10/12/2008	ufepam	13.1	30.1	1.4			W	28	S	SSE	15
11/12/2008	ufepam	13.4	30.4	0			N	30	SSE	ESE	17
12/12/2008	ufepam	15.9	21.7	2.2			NNE	31	NE	ENE	15
13/12/2008	ufepam	15.9	18.6	15.6			W	61	NNW	NNW	28
14/12/2008	ufepam	12.6	21	3.6			SW	44	W	SSW	24
15/12/2008	ufepam	8.4	24.6	0					S	WNW	4
16/12/2008	ufepam	9.8	27.7				WNW	50		WNW	
17/12/2008	ufepam	14.1	20.9	0			ENE	22	SSW	E	11
18/12/2008	ufepam	13.5	22.9	16.8			W	63	N	WNW	6

Fecha	Temperatura_Minima	Temperatura_Maxima	Precipitacion	Horas_de_Sol	Velocidad_rafaga_viento	Velocidad_viento_9am	Velocidad_viento_3pm	Humedad_9am	Humedad_3pm	Presion_9am
01/12/2008	0.516509434	0.52362949	0.001617251	0.524908795	0.294573643	0.153846154	0.275862069	0.71	0.22	0.449586777
02/12/2008	0.375	0.565217391	0	0.524908795	0.294573643	0.030769231	0.252873563	0.44	0.25	0.497520661
03/12/2008	0.504716981	0.576559546	0	0.524908795	0.310077519	0.146153846	0.298850575	0.38	0.3	0.447933884
04/12/2008	0.41745283	0.620037807	0	0.524908795	0.139534884	0.084615385	0.103448276	0.45	0.16	0.61322314
05/12/2008	0.613207547	0.701323251	0.002695418	0.524908795	0.271317829	0.053846154	0.229885057	0.82	0.33	0.500826446
06/12/2008	0.544811321	0.652173913	0.000539084	0.524908795	0.387596899	0.146153846	0.275862069	0.55	0.23	0.474380165
07/12/2008	0.537735849	0.563327032	0	0.524908795	0.341085271	0.153846154	0.275862069	0.49	0.19	0.480991736
08/12/2008	0.382075472	0.595463138	0	0.524908795	0.224806202	0.046153846	0.195402299	0.48	0.19	0.543801653
09/12/2008	0.429245283	0.693761815	0	0.524908795	0.573643411	0.053846154	0.32183908	0.42	0.09	0.469421488
10/12/2008	0.509433962	0.65973535	0.003773585	0.524908795	0.170542636	0.115384615	0.126436782	0.58	0.27	0.438016529
11/12/2008	0.516509434	0.665406427	0	0.524908795	0.186046512	0.130769231	0.068965517	0.48	0.22	0.517355372
12/12/2008	0.575471698	0.50094518	0.005929919	0.524908795	0.19379845	0.115384615	0.149425287	0.89	0.91	0.495867769
13/12/2008	0.575471698	0.442344045	0.042048518	0.524908795	0.426356589	0.215384615	0.32183908	0.76	0.93	0.228099174
14/12/2008	0.497641509	0.487712665	0.009703504	0.524908795	0.294573643	0.184615385	0.229885057	0.65	0.43	0.34214876
15/12/2008	0.398584906	0.555765595	0	0.524908795	0.263838993	0.030769231	0.344827586	0.57	0.32	0.482644628
16/12/2008	0.431603774	0.61436673	0.006363661	0.524908795	0.341085271	0.108026353	0.252873563	0.5	0.28	0.543801653
17/12/2008	0.533018868	0.485822306	0	0.524908795	0.124031008	0.084615385	0.103448276	0.69	0.82	0.523966942
18/12/2008	0.518867925	0.52362949	0.045283019	0.524908795	0.441860465	0.046153846	0.229885057	0.8	0.65	0.418181818

EDA

- Análisis Exploratorio de Datos (EDA) para poder comprender mejor la naturaleza de los datos y las características relevantes.
- Spark SQL

Días con cambios bruscos en la velocidad del viento:

Fecha	wind_change
2011-08-05	0.896551724137931
2012-08-17	0.7471264367816092
2011-09-20	0.7471264367816092
2013-10-29	0.7087533156498674
2013-10-13	0.67789566755084

Días con mayor y menor precipitación:

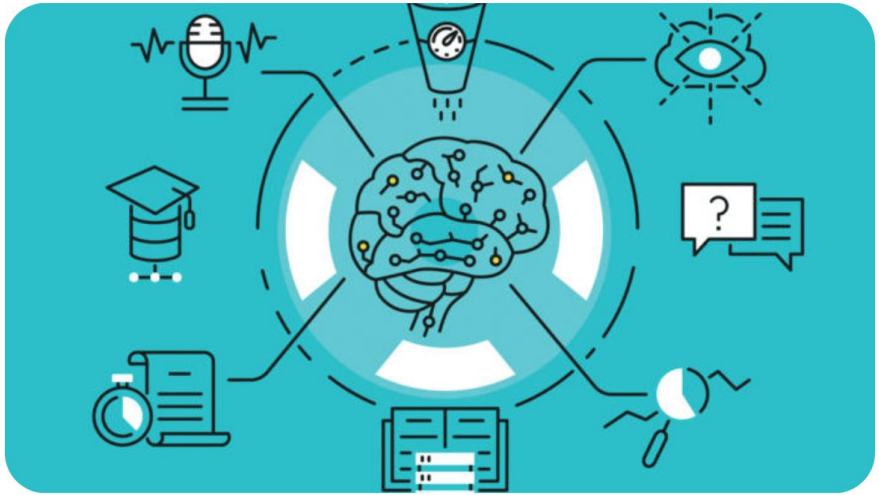
Fecha	Precipitacion
2009-11-07	1.0
2011-02-16	0.9908355795148248
2009-01-12	0.7504043126684636
2011-02-04	0.7239892183288411
2015-02-08	0.6663072776280323

Fecha	Precipitacion
2015-02-13	2.695417789757E-4
2013-08-01	2.695417789757E-4
2015-07-10	2.695417789757E-4
2013-09-11	2.695417789757E-4
2015-02-26	2.695417789757E-4

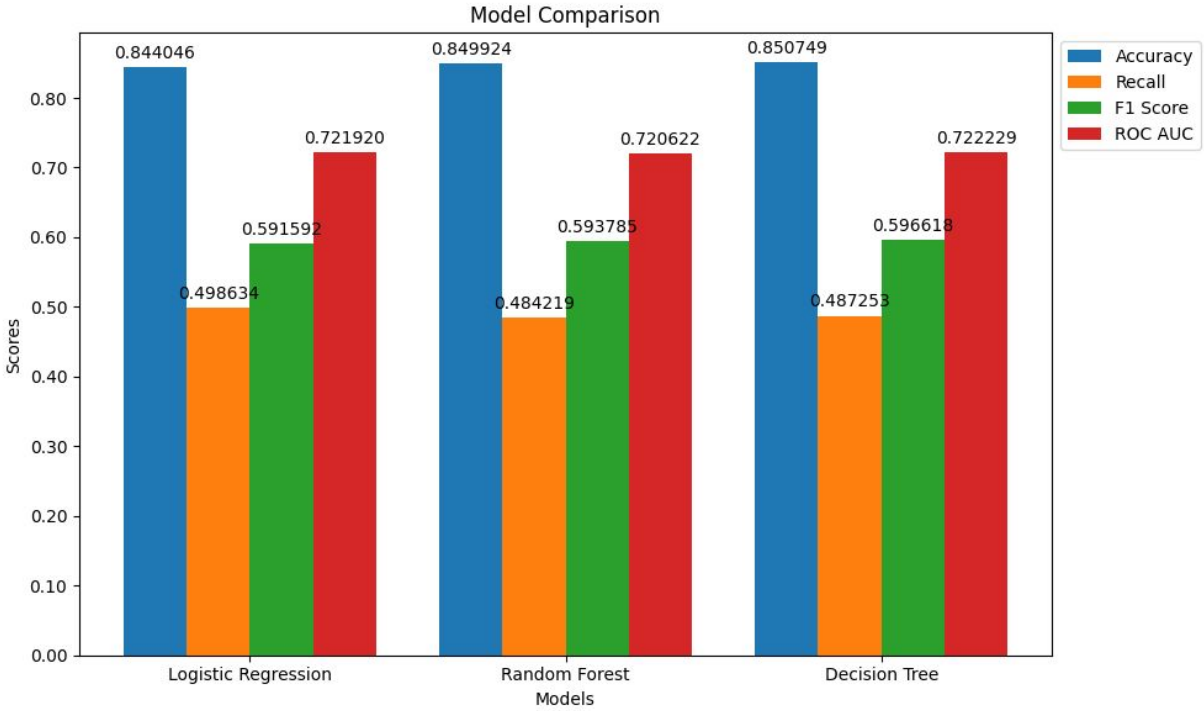
Modelos de clasificación

- Modelos elegidos:

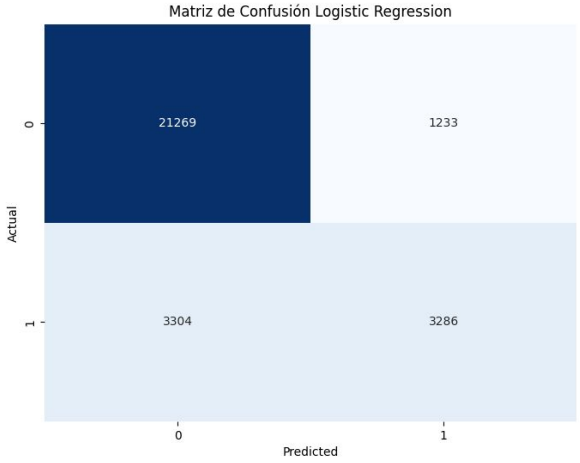
1. Linear Regression
2. Random Forest
3. Decision Tree

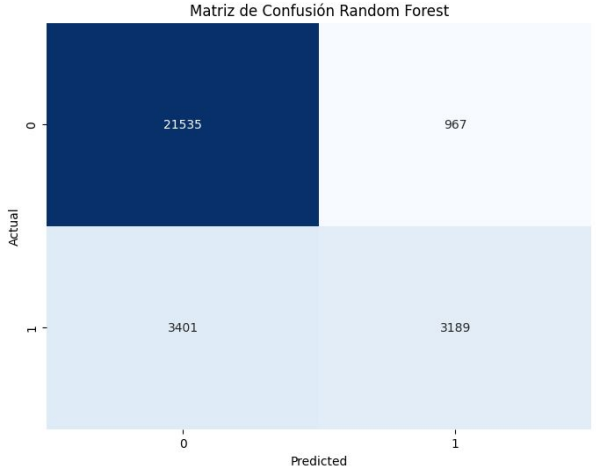


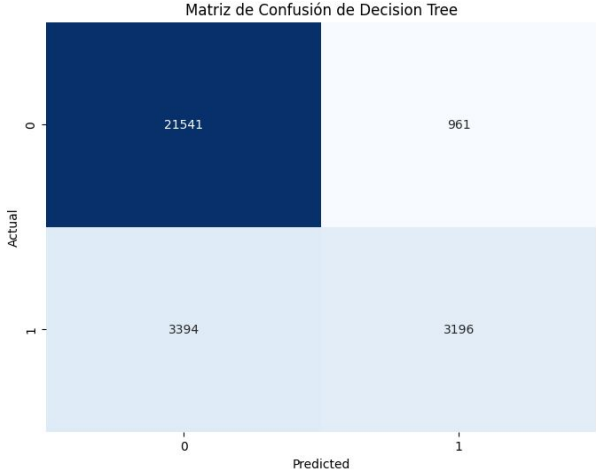
Comparación de modelos



Comparación de modelos





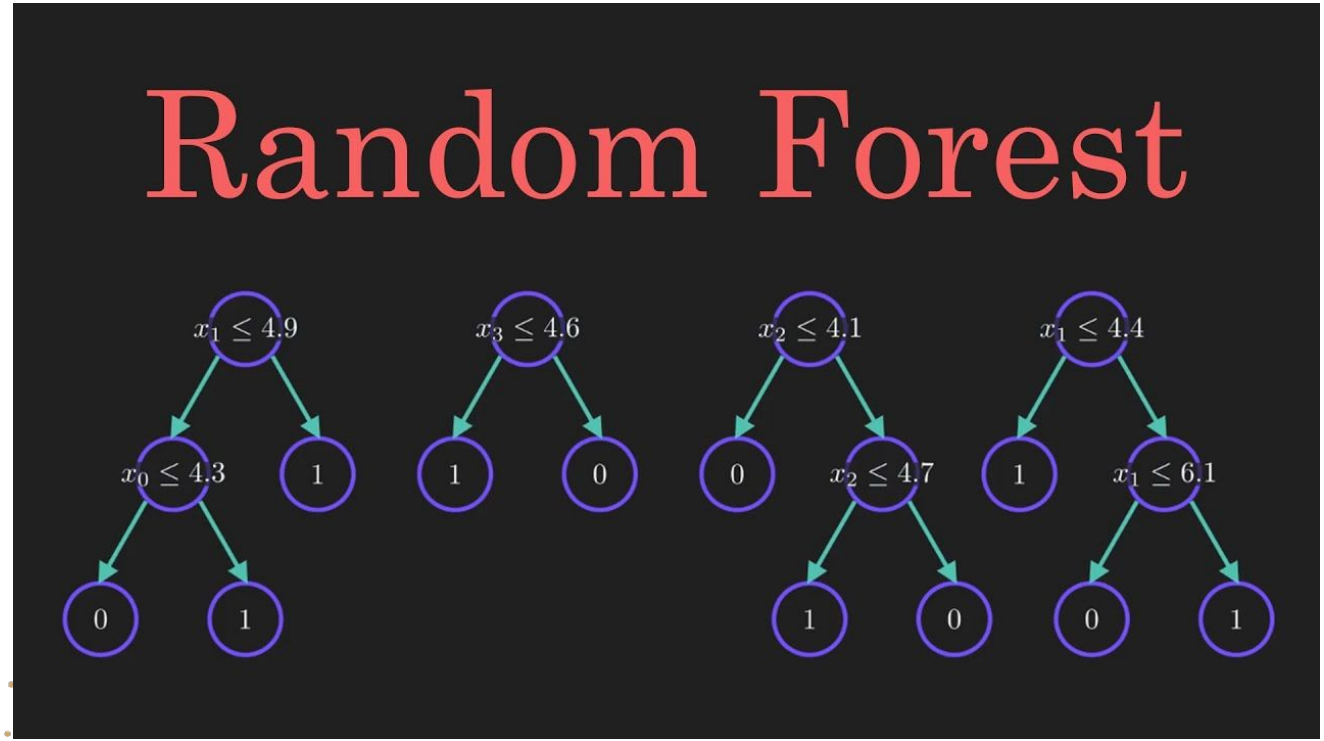


Análisis de Modelos

- Precisión (Accuracy): Es la proporción de predicciones correctas entre el total de casos.
- Recuperación (Recall): Es la proporción de casos positivos reales que fueron identificados correctamente.
- Puntuación F1 (F1 Score): Es el promedio armónico de la precisión y la recuperación. Es útil cuando quieres un balance entre precisión y recuperación, especialmente si las clases están desbalanceadas.
- ROC AUC: Mide la capacidad del modelo para distinguir entre clases. Un valor más alto indica una mejor discriminación.

- Regresión Logística: Cuenta el menor valor de **precisión**, pero el mayor valor de ROC AUC, indicando así una mejor capacidad para distinguir entre clases.
- Bosque Aleatorio y Árbol de Decisión: Extrañamente, estos dos modelos tienen exactamente las mismas métricas. La precisión de ambos es ligeramente superior a la regresión logística, pero su ROC AUC es marginalmente menor.

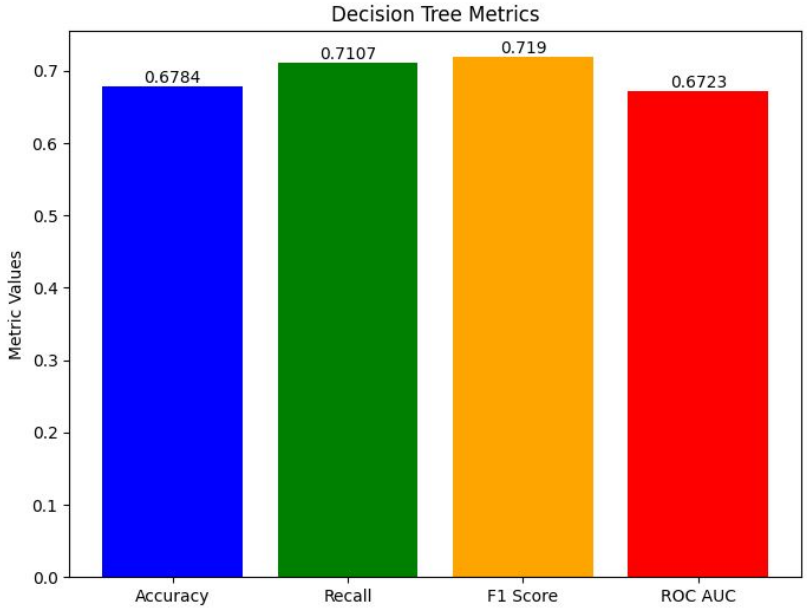
Modelo Final

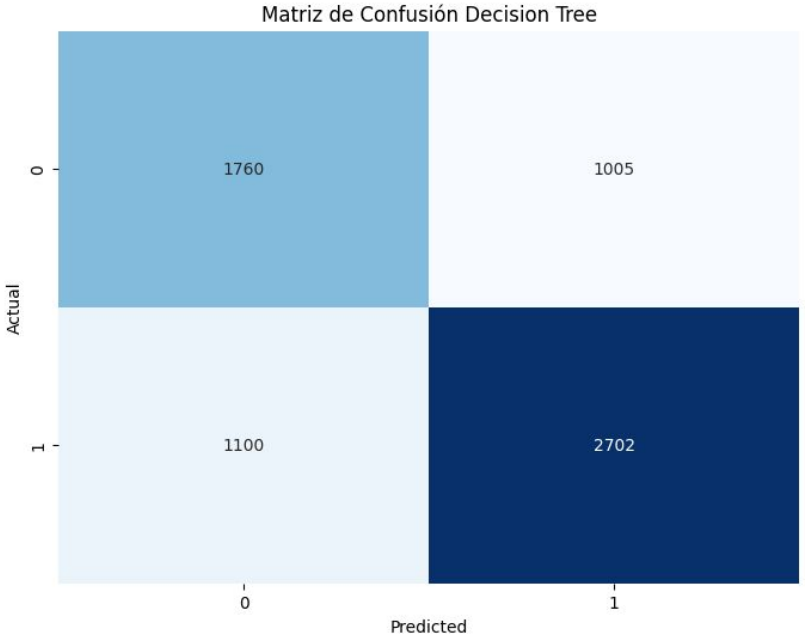


Nuevo dataset

- Sacamos un dataset de kaggle
- Más de 30,000 canciones

Resultados de Modelo





Gracias