Universidad Panamericana

1238-LM0047 Aprendizaje de Máquina 2152

Due Date: November 27, 2023

Teacher: Andrei Noguera Gil

Proyecto Final Modelo Clasificación

ID	Name	Career
0235320	Esteban Viniegra Pérez Olagaray	Data Intelligence and Cybersecurity Engineering
0231635	Santiago Valdez Bocardo	Data Intelligence and Cybersecurity Engineering

Introducción

Machine Learning es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la **capacidad de identificar patrones en datos masivos y elaborar predicciones** (análisis predictivo). Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, **sin necesidad de ser programados**.

Características Principales

- Aprendizaje a partir de Datos: El aprendizaje automático se basa en la capacidad de las máquinas para analizar y aprender de datos. Cuantos más datos relevantes se proporcionen, mejor será la capacidad del modelo para realizar predicciones o tomar decisiones.

- Generalización: Los modelos de machine learning están diseñados para generalizar patrones a partir de datos de entrenamiento y aplicar esos patrones a nuevas situaciones. El objetivo es que el modelo funcione bien incluso con datos que no ha visto antes.

- Adaptabilidad: Los modelos de machine learning son capaces de adaptarse y mejorar a medida que se les proporciona más información. Esto contrasta con los programas tradicionales que siguen reglas estrictas y no mejoran con el tiempo.

- Automatización: La automatización es fundamental en machine learning. Una vez que el modelo está entrenado, puede realizar tareas específicas sin intervención humana, lo que permite la automatización de procesos complejos.

Proceso de Machine Learning

1. Definición del Problema: Identificar claramente el problema que se quiere resolver y si el aprendizaje automático es la mejor aproximación para abordarlo.

2. Recopilación de Datos: Recolectar datos relevantes para entrenar y evaluar el modelo. La calidad y cantidad de los datos son críticos para el éxito del aprendizaje automático.

3. Preprocesamiento de Datos: Limpieza y transformación de datos para garantizar que estén en un formato adecuado para el entrenamiento del modelo.

4. Selección del Modelo: Elegir el algoritmo de machine learning más adecuado para el problema en cuestión.

5. Entrenamiento del Modelo: Utilizar datos de entrenamiento para ajustar los parámetros del modelo y permitirle aprender patrones.

6. Evaluación del Modelo: Evaluar el rendimiento del modelo utilizando datos de prueba independientes para asegurarse de que generalice bien a nuevas situaciones.

7. Ajuste y Optimización: Modificar parámetros o elegir un modelo diferente para mejorar el rendimiento.

8. Predicciones y Despliegue: Utilizar el modelo entrenado para realizar predicciones en situaciones del mundo real.

Ciencia de Datos

La ciencia de datos es un campo interdisciplinario que utiliza técnicas, procesos y sistemas científicos para extraer conocimiento y comprensión de datos en diversas formas. Incluye la exploración de datos, el análisis estadístico, el diseño y la construcción de modelos predictivos, y la interpretación de resultados para informar la toma de decisiones.

En el contexto del machine learning, la ciencia de datos es esencial para el éxito del proyecto, ya que abarca la recolección, limpieza y preparación de datos, así como la interpretación y comunicación de los resultados obtenidos a partir de modelos de machine learning. La ciencia de datos proporciona el marco necesario para transformar datos en información significativa y conocimiento accionable.

Modelos de clasificación

Los modelos de clasificación en machine learning son algoritmos diseñados para asignar categorías a datos según ciertos patrones identificados durante el entrenamiento del modelo. La tarea principal de un modelo de clasificación es predecir la pertenencia a una o más clases o categorías para nuevas instancias de datos. Aquí hay algunas características y ejemplos comunes de modelos de clasificación:

- Regresión Logística:
    - Características:
        - Utilizada para problemas de clasificación binaria.
        - Calcula la probabilidad de que una instancia pertenezca a una categoría específica.
        - Se basa en la función logística para realizar la clasificación.

- Máquinas de Soporte Vectorial (SVM):
    - Características:
    - Puede utilizarse para problemas de clasificación binaria y multiclase.
    - Busca el hiperplano que mejor separa las instancias de diferentes clases en un espacio dimensional más alto.
    - Es efectiva en espacios de alta dimensión.

- Árboles de Decisión:
    - Características:
        - Estructura jerárquica de decisiones basada en condiciones lógicas.
        - Puede manejar tanto problemas de clasificación como de regresión.

        - Intuitivos y fácilmente interpretables.

- Random Forest:
    - Características:
        - Conjunto de árboles de decisión que trabajan juntos para realizar la clasificación.
        - Reduce el sobreajuste y mejora la precisión promediando las predicciones de varios árboles.

- K Vecinos Más Cercanos (K-NN):
    - Características:
        - Asigna una instancia a la clase más común entre sus k vecinos más cercanos.
        - Simple pero puede ser efectivo en conjuntos de datos más pequeños.
        - Depende de la elección del parámetro k.

- Naive Bayes:
    - Características:
        - Basado en el teorema de Bayes.
        - Supone independencia condicional entre las características, lo que puede simplificar el cálculo de probabilidades.
        - A menudo utilizado en problemas de clasificación de texto.
- Gradient Boosting:
    - Características:
        - Combina múltiples modelos más débiles para crear un modelo fuerte.
        - Ajusta iterativamente los modelos para corregir los errores del modelo anterior.

Problema

El problema que se aborda en este proyecto se centra en la generación de modelos de clasificación para predecir el clima en Brasil. Esta tarea es fundamental, ya que la capacidad de prever con precisión el clima desempeña un papel crítico en diversas áreas de la vida. Desde agricultores que cultivan cosechas hasta familias que planifican un fin de semana de vacaciones o la toma de decisiones logísticas en aerolíneas, en particular, influye en gran medida en los planes. En algunos casos, el impacto del clima puede tener grandes consecuencias financieras. Por lo tanto, existe un fuerte interés por parte de numerosos interesados en la capacidad de prever con precisión el clima. El objetivo de este proyecto es utilizar los datos disponibles para crear un modelo de predicción para determinar el clima. Un modelo de este tipo podría ser utilizado en una aplicación meteorológica en beneficio del público en general.

Descripción de dataset

El conjunto de datos se compone de un único archivo, denominado "weather.csv", que a su vez se originó en las Observaciones Diarias del Clima. Se pueden encontrar métricas

meteorológicas adicionales para Brasil en la aplicación web de Datos Climáticos en Línea. A continuación, se detallará y ampliará la descripción de estos atributos para una comprensión más completa:

- Fecha : Fecha de la observación.
- Ubicación : Ubicación de la estación meteorológica.
- Temperatura Mínima : Temperatura mínima en las 24 horas hasta las 9 a. m. Grados Celsius
- Temperatura Máxima : Temperatura máxima en las 24 horas hasta las 9 a. m. Grados Celsius
- Precipitación : Precipitación (lluvia) en las 24 horas hasta las 9 a. m. Milímetros
- Evaporación : Evaporación de la clase A en las 24 horas hasta las 9 a. m. Milímetros
- Horas de Sol : Brillo del sol en las 24 horas hasta la medianoche. Horas
- Dirección de Ráfaga de Viento : Dirección de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche. 16 puntos cardinales
- Velocidad de Ráfaga de Viento : Velocidad de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche. Kilómetros por hora
- Dirección del Viento a las 9 a. m: Dirección del viento a las 9 a. m. 16 puntos cardinales
- Dirección del Viento a las 3 p. m: Dirección del viento a las 3 p. m. 16 puntos cardinales
- Velocidad del Viento a las 9 a. m: Velocidad del viento a las 9 a. m. Kilómetros por hora
- Velocidad del Viento a las 3 p. m: Velocidad del viento a las 3 p. m. Kilómetros por hora
- Humedad a las 9 a. m: Humedad relativa a las 9 a. m. Porcentaje
- Humedad a las 3 p. m: Humedad relativa a las 3 p. m. Porcentaje
- Presión a las 9 a. m: Presión atmosférica reducida al nivel medio del mar a las 9 a.m.
- Hectopascales
- Presión a las 3 p. m: Presión atmosférica reducida al nivel medio del mar a las 3 p.m. Hectopascales
- Nubosidad a las 9 a. m: Fracción del cielo oscurecido por nubes a las 9 a. m. Octavos
- Nubosidad a las 3 p. m: Fracción del cielo oscurecido por nubes a las 3 p. m. Octavos
- Temperatura a las 9 a. m: Temperatura a las 9 a. m. Grados Celsius
- Temperatura a las 3 p. m: Temperatura a las 3 a. m. Grados Celsius
- Lluvia Hoy : ¿Recibió el día actual precipitación superior a 1 mm en las 24 horas hasta las 9 a. m.? Binario (0 = No, 1 = Sí)
- Lluvia Mañana : ¿Recibirá el próximo día precipitación superior a 1 mm en las 24 horas hasta las 9 a. m.? Binario (0 = No, 1 = Sí)

Problemática y valor agregado

Problemática

La problemática central en este proyecto radica en la **necesidad de predecir con precisión el clima en Brasil**. La predicción del clima es esencial en diversas esferas de la vida cotidiana y la toma de decisiones en varias industrias. Las fluctuaciones climáticas pueden tener un impacto significativo en la **agricultura, la planificación de actividades al aire libre, la gestión de recursos hídricos, las operaciones logísticas y la industria del transporte**, entre otras áreas.

Valor Generado por el Modelo

El modelo de clasificación que se propondrá ofrecerá un valor sustancial en varios aspectos:

- Precisión en la Toma de Decisiones: Proporciona información precisa y oportuna sobre las condiciones climáticas futuras, permitiendo que agricultores, planificadores de eventos, empresas de logística y otras partes interesadas tomen decisiones informadas.

- Mitigación de Riesgos: Ayuda a mitigar los riesgos asociados con eventos climáticos adversos al permitir una preparación temprana y una respuesta proactiva ante condiciones climáticas desafiantes.

- Eficiencia Operativa: En el caso de industrias como la aviación y la agricultura, la capacidad de anticipar las condiciones climáticas permite una planificación más eficiente de las operaciones diarias, lo que puede traducirse en ahorros de costos y recursos.

- Impacto Socioeconómico: Contribuye al bienestar general al proporcionar a la sociedad información confiable sobre el clima, facilitando así una mejor planificación y gestión de recursos.

- Facilita la Innovación Tecnológica: Alimenta el desarrollo de tecnologías avanzadas en el campo de la meteorología y la inteligencia artificial, fomentando la innovación y la mejora continua en la precisión de las predicciones climáticas.

Tipo de Aprendizaje y Categoría de Modelo

Uso de modelos de clasificación

El problema de predecir el clima en Brasil se presta naturalmente para el enfoque de modelos de clasificación. 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.

Justificación

- Naturaleza de la Salida Deseada: Debido a que la salida deseada es binaria (por ejemplo, lluvia sí/no), se ajusta a un problema de clasificación binaria.

- Disponibilidad de Datos Etiquetados: Para entrenar modelos de clasificación, se necesitan datos históricos con etiquetas precisas que indiquen las condiciones climáticas reales, algo con lo que ya se cuenta para este problema.

Tipo de Aprendizaje

En este contexto, se utilizará el aprendizaje supervisado. Esto se debe a que 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 de dataset

Para mejorar el análisis de los datos utilizando modelos de machine learning, primero se limpió el dataset proporcionado. Mediante un código de python, se arreglaron ciertas características del dataset, como imputación de valores nulos, conversión de fechas, eliminación de duplicados, codificación de variables categóricas, eliminación de columnas innecesarias, y normalización de variables numéricas. Adicionalmente, se eliminaron de los nombres de columnas los acentos. Luego de correr el código de python, se generó un nuevo dataset llamado “weather_clean.csv”.

EDA

Se realizó un Análisis Exploratorio de Datos (EDA) para poder comprender mejor la naturaleza de los datos y las características relevantes de ellos para poder entender mejor el problema y saber qué datos son relevantes para el proyecto. Las consultas fueron realizadas usando una sesión de spark usando SQL. Algunas de las consultas realizadas fueron:

- Temperatura máxima y mínima registrada:

```
Unset
result_avg_temp = spark.sql("SELECT AVG(Temperatura_Minima) AS
avg_temp_min, AVG(Temperatura_Maxima) AS avg_temp_max FROM
weather_data")
result_avg_temp.show()

+------------------+------------------+
```

```
|      avg_temp_min|      avg_temp_max|
+------------------+------------------+
|0.4880668486077552|0.5297041261937112|
+------------------+------------------+
```

- Días más cálidos y más fríos:

Unset

```
result_hottest_days = spark.sql("SELECT Fecha, Temperatura_Maxima
FROM weather_data ORDER BY Temperatura_Maxima DESC LIMIT 5")
result_coldest_days = spark.sql("SELECT Fecha, Temperatura_Minima
FROM weather_data ORDER BY Temperatura_Minima ASC LIMIT 5")
result_hottest_days.show()
result_coldest_days.show()
```

```
+----------+------------------+
|     Fecha|Temperatura_Maxima|
+----------+------------------+
|2011-01-25|               1.0|
|2014-01-03|0.9848771266540642|
|2017-02-12|0.9848771266540642|
|2017-02-11|0.9792060491493384|
|2017-02-11|0.9773156899810964|
+----------+------------------+

+----------+------------------+
|     Fecha|Temperatura_Minima|
+----------+------------------+
|2009-06-11|               0.0|
|2011-07-29|0.0070754716981132|
|2015-08-04|0.0070754716981132|
|2010-06-29|0.0117924528301886|
|2011-07-29|0.0117924528301886|
+----------+------------------+
```

- Días con mayor cambio de temperatura

Unset

```
result_temp_change = spark.sql("SELECT Fecha, (Temperatura_Maxima
- Temperatura_Minima) AS temp_change FROM weather_data ORDER BY
temp_change DESC LIMIT 5")
result_temp_change.show()
```

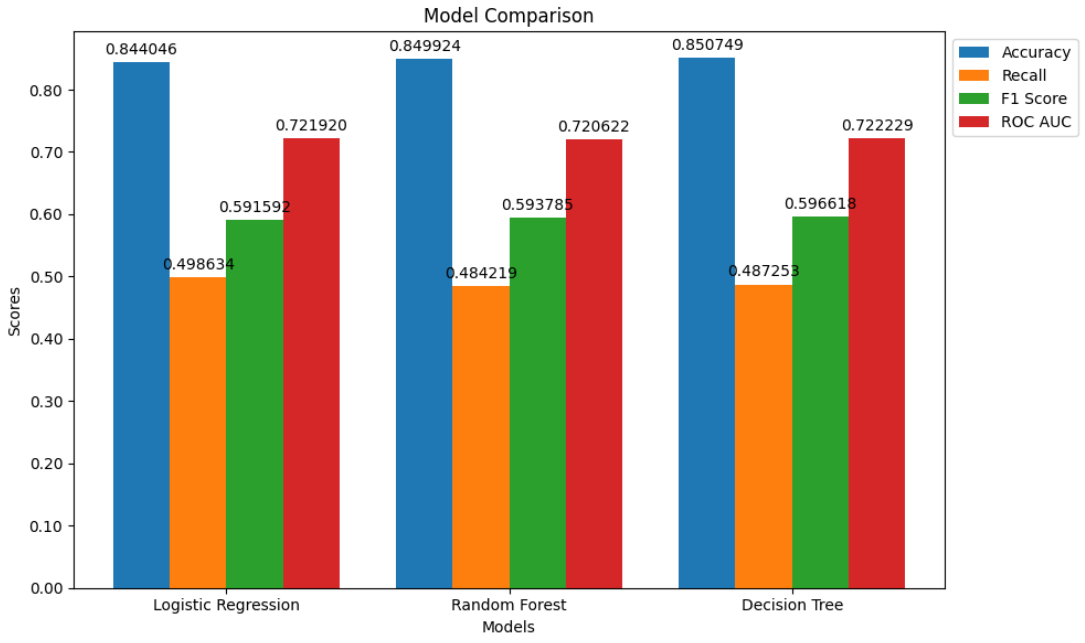
```
+----------+-------------------+
|     Fecha|        temp_change|
+----------+-------------------+
|2011-06-07| 0.4683833714767212|
|2009-10-01| 0.4448050790027464|
|2016-11-05|  0.443400684809359|
|2012-02-24| 0.4426070906302387|
|2009-08-31|0.43064752242011745|
+----------+-------------------+
```

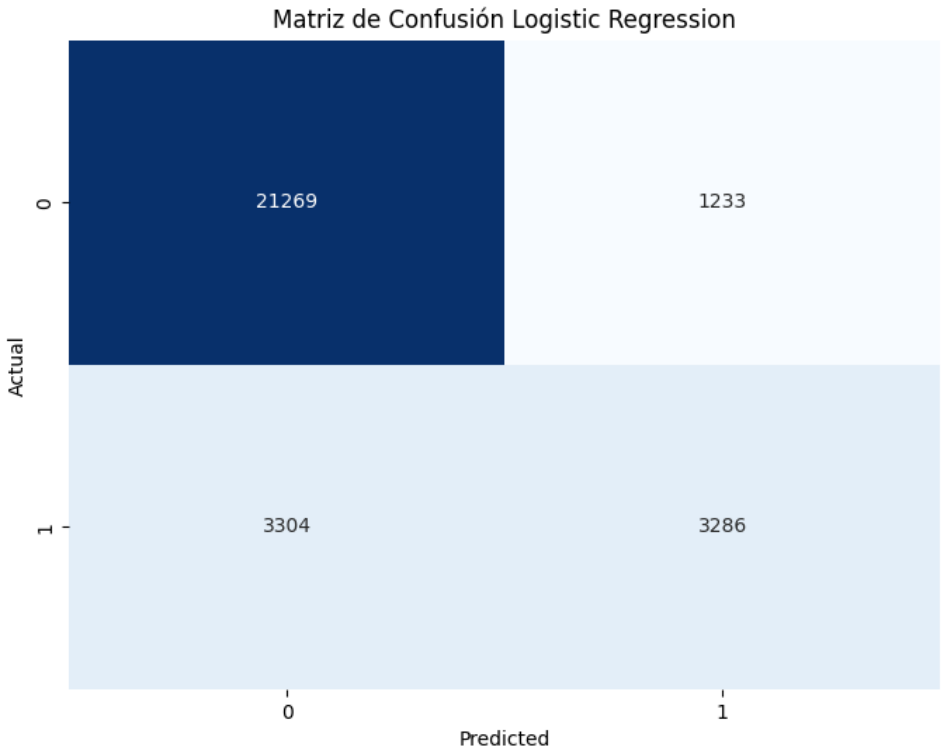
Elección de modelos y justificación

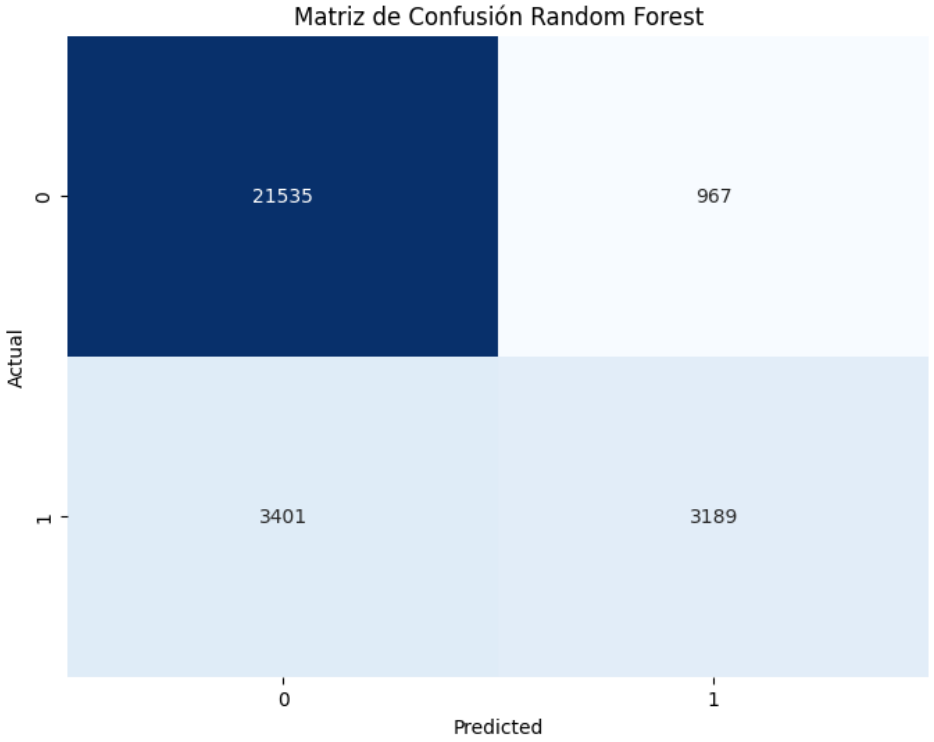
Los 3 modelos que se eligieron para probar métricas con el dataset fueron:

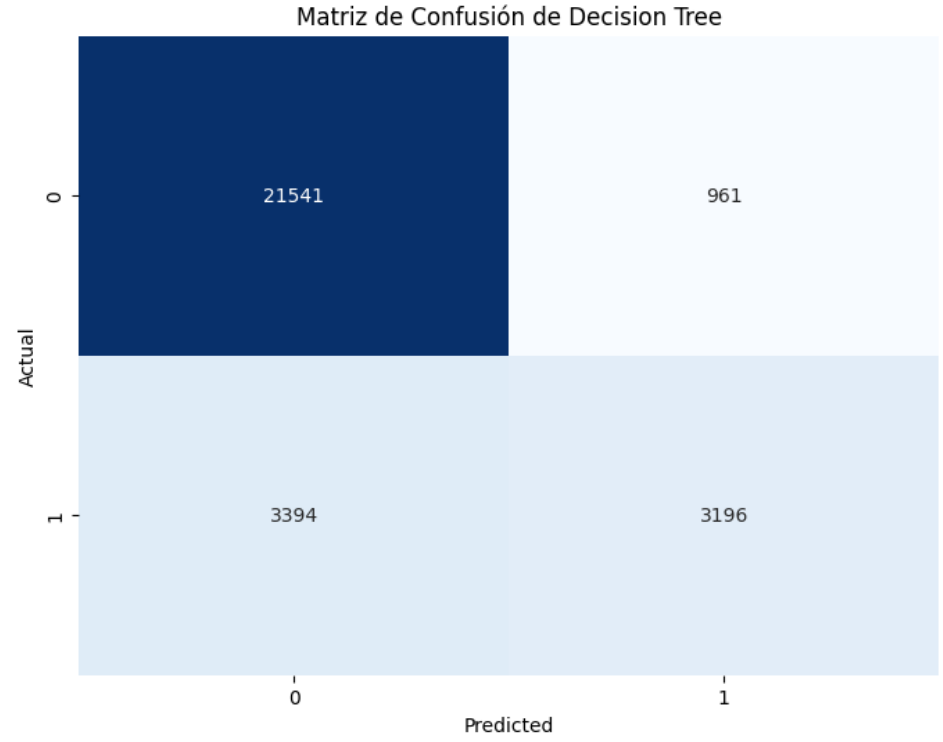
- Logistic Regression
- Random Forest
- Decision Tree

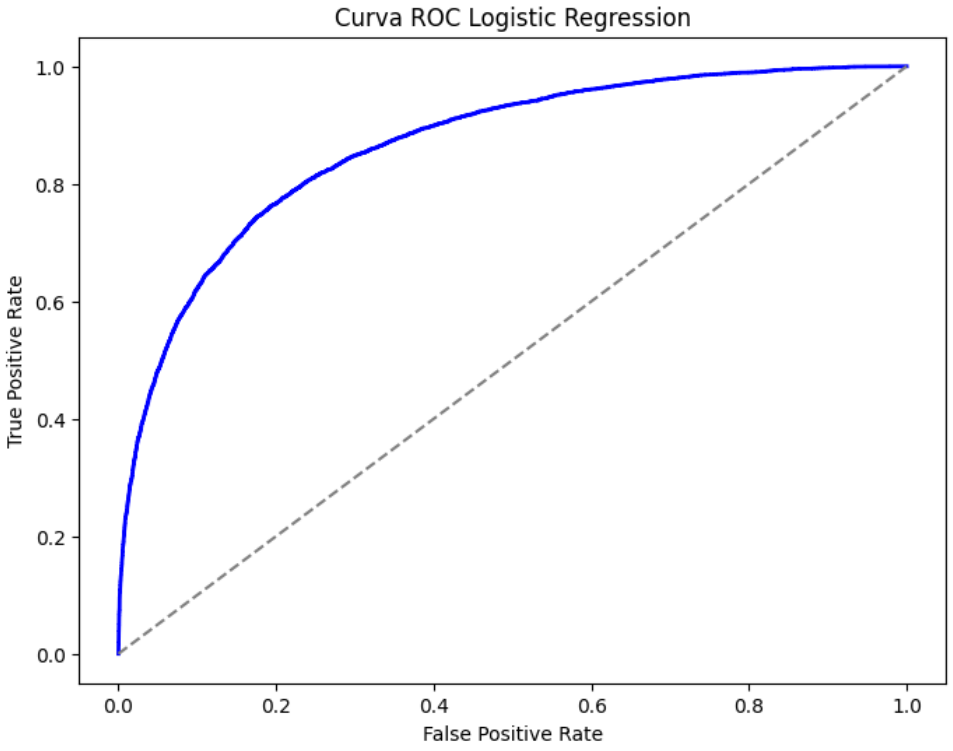
Luego de correr estos modelos, obtuvimos estos resultados de métricas utilizando gráficas:

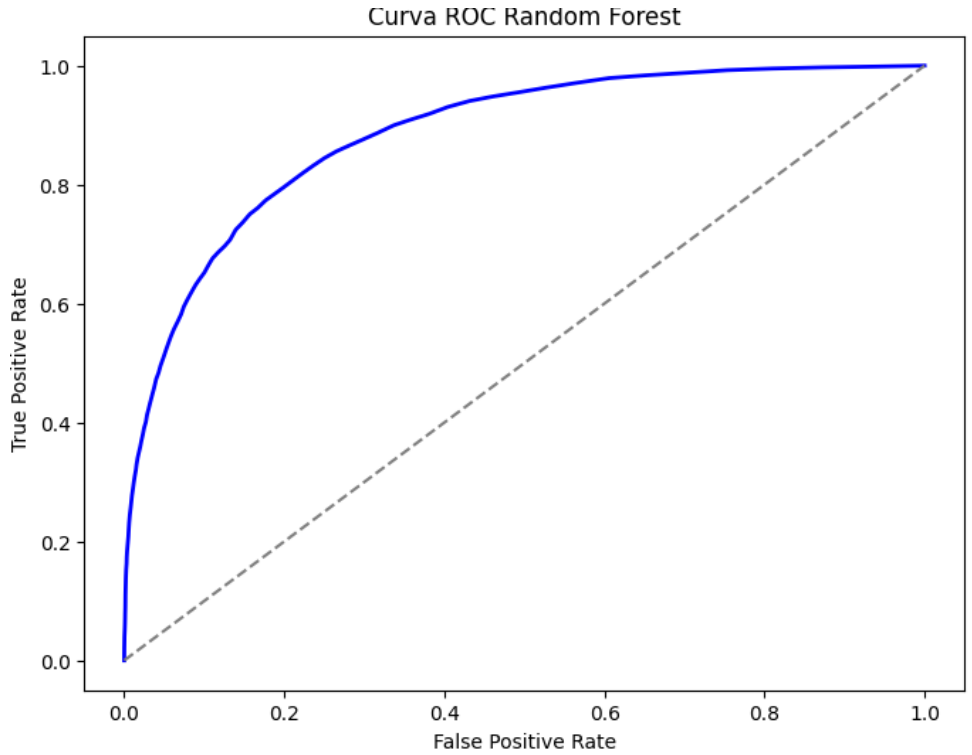


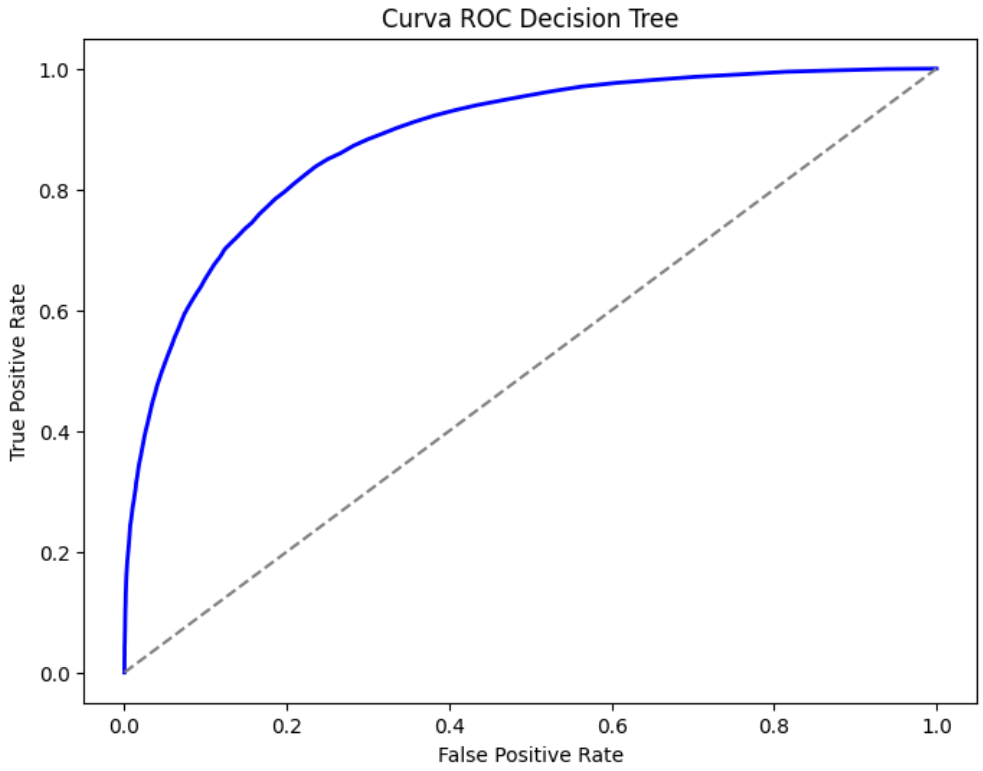

Matriz de Confusión Logistic Regression
21269
1233
3304
3286
Actual
Predicted
0
1


Matriz de Confusión Random Forest
21535
967
3401
3189
Actual
Predicted
0
1


Matriz de Confusión de Decision Tree
21541
961
3394
3196
Actual
0
1
0
1
Predicted


Curva ROC Logistic Regression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0





El modelo que elegimos, luego de analizar las métricas obtenidas, es Decision Tree. Esta decisión está fundamentada de acuerdo a los siguiente:

- Precisión (Accuracy): Es la proporción de predicciones correctas entre el total de casos. Es una buena medida general, pero puede ser engañosa si las clases están desbalanceadas.

- Recuperación (Recall): Es la proporción de casos positivos reales que fueron identificados correctamente. Es crucial si los falsos negativos son más problemáticos que los falsos positivos.

- 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.

Ahora, evaluando los modelos implementados:

- 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.

Por último, ya que el objetivo principal del proyecto es maximizar la capacidad general de predicción (precisión), el Bosque Aleatorio o el Árbol de Decisión podrían ser ligeramente mejores. Entre estos dos modelos, decidimos utilizar Árbol de Decisión

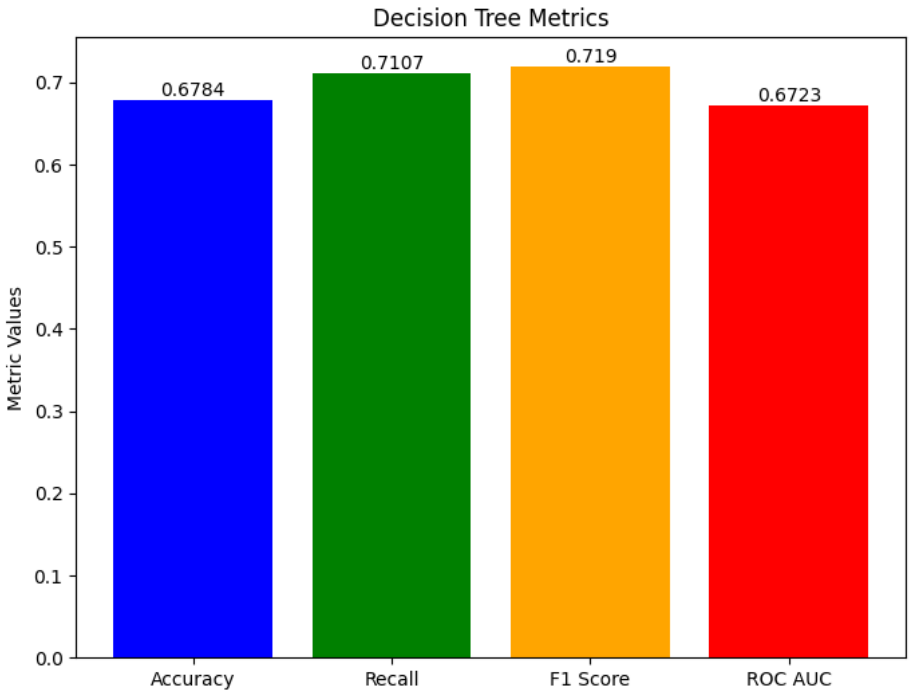
Nuevo Dataset Usando Árbol de Decisión

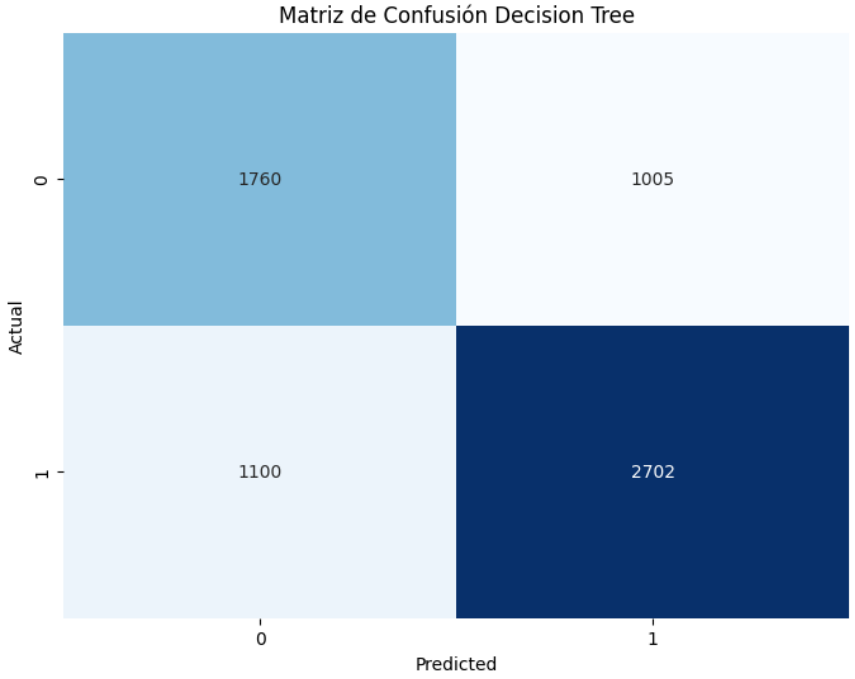
Tras haber encontrado el mejor modelo de machine learning, elegimos un nuevo dataset para utilizar el modelo. Elegimos un dataset de spotify con datos de 30,000 canciones en la plataform, dichos datos son:

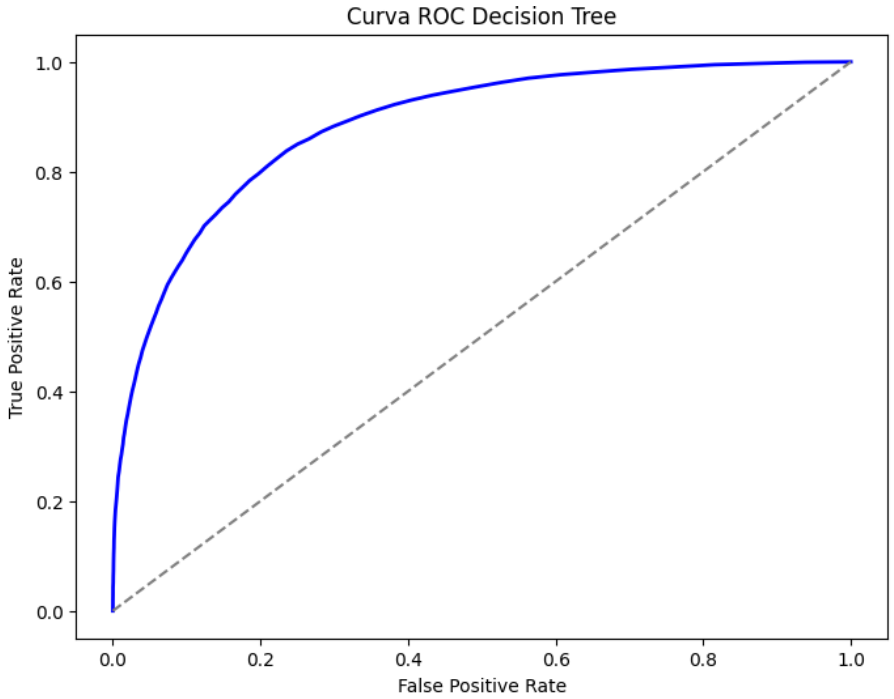
Variable	Class	Description
track_id	Character	Identificador único de la canción
track_name	Character	Name of the song
track_artist	Character	Artist of the song
track_popularity	Double	Song popularity (0-100), where higher is better
track_album_id	Character	Unique identifier of the album
track_album_name	Character	Name of the album of the song
track_album_release_date	Character	Release date of the album
playlist_name	Character	Name of the playlist
playlist_id	Character	Playlist identifier
playlist_genre	Character	Playlist genre
playlist_subgenre	Character	Playlist subgenre
danceability	Double	Danceability describes how suitable a track is for dancing. 0.0 is least danceable, 1.0 is most danceable
energy	Double	Energy is a measure from 0.0 to 1.0 representing a perceptual measure of intensity and activity
key	Double	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation
loudness	Double	The overall loudness of a track in decibels (dB). Values typically range between -60 and 0 dB
mode	Double	Mode indicates the modality (major or minor) of a track. Major is represented by 1 and minor is 0
speechiness	Double	La speechiness detecta la presencia de palabras habladas en una pista
acousticness	Double	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
instrumentalness	Double	Predicts whether a track contains no vocals
liveness	Double	Detects the presence of an audience in the recording
valence	Double	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track
tempo	Double	The overall estimated tempo of a track in beats per minute (BPM)
duration_ms	Double	Duration of song in milliseconds

Luego de haber aplicado el modelo, encontramos que no obtuvimos tan buenos resultados como en el dataset de weather, y deducimos que esto es debido a la naturaleza de los datos de este dataset, que son muy distintos a los que probamos anteriormente. Sin embargo, de igual manera analizamos los resultados de métricas obtenidas para poder llegar a esta conclusión.

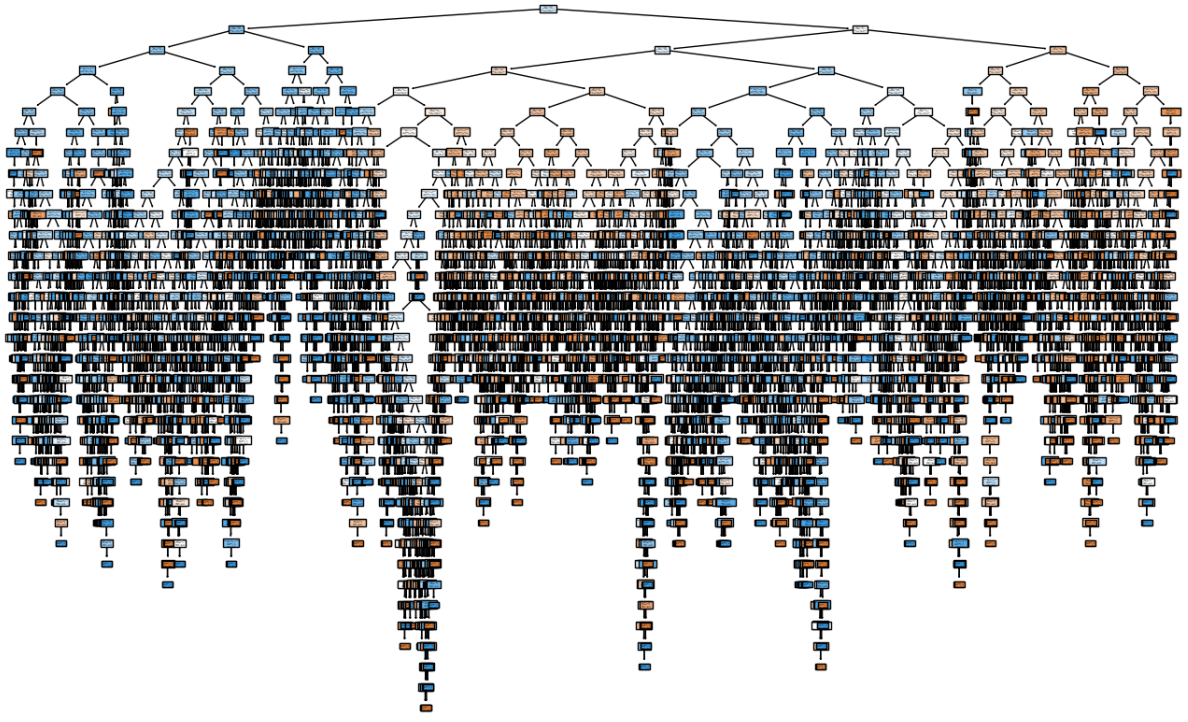
- Accuracy: 0.6794578955382975
- Recall: 0.7106785902156759
- F1 Score: 0.719669729657744
- ROC AUC: 0.6736033095743841


Decision Tree Metrics
Metric Values
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.6784
0.7107
0.719
0.6723
Accuracy
Recall
F1 Score
ROC AUC


Matriz de Confusión Decision Tree
Actual
0
1
1760
1005
1100
2702
0
1
Predicted



Visualización del Árbol



Conclusión

En este proyecto, pudimos aplicar el aprendizaje de máquina en el campo de predicción de clima en Brasil. Al momento de resolver la problemática de este proyecto de encontrar un modelo de ML adecuado para predecir diversas variables meteorológicas, pudimos apreciar la importancia de seleccionar el modelo de machine learning adecuado para resolver problemas específicos.

Adicionalmente, durante el proyecto y el curso aprendimos que la elección de un modelo de machine learning correcto para resolver un problema es una tarea altamente importante, ya que cada algoritmo tiene sus propias fortalezas y debilidades. Los diversos modelos y sus distintas características, desde Random Forest hasta regresión lineal, demuestra la variedad de herramientas disponibles para abordar distintos problemas de complejidad variable. La clave para elegir el modelo correcto para el problema consta en comprender la naturaleza del problema, analizar los datos que se tienen (mediante análisis como EDA) y seleccionar el modelo que se alinee de manera óptima con todas las características y a lo que se quiere llegar, proporcionando de esta forma soluciones más precisas.

Este proyecto nos fue de gran utilidad para comprender la relevancia del aprendizaje de máquina en la actualidad, donde la abundancia de datos, la complejidad de los problemas y el gran avance tecnológico con el que ya se cuenta permiten resolver problemas que requieren enfoques avanzados; desde la optimización de procesos hasta la toma de decisiones basada en datos, el aprendizaje de máquina es una herramienta importante para extraer información significativa a partir de conjuntos de datos masivos y variables y luego resolver problemas complejos.

Pasos siguientes

Para poder mejorar el proyecto, lo mejor sería tomar acciones para continuar mejorando nuestro modelo de ML, esto se puede lograr realizando optimización de hiperparámetros, validación cruzada. etc. transformación de datos adicionales, entre otros. Un modelo de machine learning entre más es entrenado un modelo con datos, más efectivo y eficiente se vuelve, por lo que lo principal para poder mejorar el proyecto es continuar entrenando al modelo elegido.

Bibliografía

- Corporativa, I. (n.d.). Descubre los principales beneficios del Machine Learning. Iberdrola. https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico
- Admin. (2021, November 3). ▷ ¿Qué es Data Science? | Universidad Complutense de Madrid. Máster Data Science. https://www.masterdatascienceucm.com/que-es-data-science/
- Parra, F. (n.d.). 6 Métodos de clasificación | Estadística y Machine Learning con R. https://bookdown.org/content/2274/metodos-de-clasificacion.html

- ¿Qué es el Aprendizaje supervisado en Machine Learning? (2022, October 13). Inesdi. https://www.inesdi.com/blog/aprendizaje-supervisado-machine-learning/
- https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs/