



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

Facultad de Ingeniería Asignatura Machine Learning

Unidad de aprendizaje 1: Supervisado – Clasificación

Introducción

Hola, estamos en la primera unidad de la asignatura denominada: **“Aprendizaje Supervisado - Clasificación”**. En esta unidad aprenderemos qué es machine learning, qué es aprendizaje supervisado y para qué sirven los algoritmos de clasificación. Asimismo, aprenderemos que es EDA, preprocesamiento de datos y realizaremos algunas pruebas con los algoritmos: Naive Bayes, árbol de decisión y Gradient Boosting Classifier.



MACHINE LEARNING



Fuente: <https://trycore.co/transformacion-digital/aplicaciones-machine-learning-en-organizaciones/>





CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.1. ¿Qué es Machine Learning?

El aprendizaje automático, conocido como "machine learning" en inglés, es un subcampo de la inteligencia artificial (IA) que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y tomar decisiones basadas en datos y experiencias pasadas sin necesidad de programación explícita. En otras palabras, el machine learning permite que las máquinas mejoren su rendimiento en tareas específicas a medida que se les proporciona más información y experiencia.

El proceso fundamental del machine learning implica lo siguiente:

Datos: Se recopilan y utilizan datos relevantes para la tarea en cuestión. Estos datos pueden ser de diferentes tipos, como texto, imágenes, sonido o números.

Entrenamiento: Se utiliza un algoritmo de machine learning para analizar y aprender patrones en los datos de entrenamiento. Durante esta fase, el modelo ajusta sus parámetros internos para realizar predicciones o tomar decisiones más precisas.

Evaluación y Prueba: Después del entrenamiento, se evalúa el rendimiento del modelo utilizando datos que no se usaron durante el entrenamiento. Esto permite verificar si el modelo es capaz de generalizar y hacer predicciones precisas en datos no vistos.

Despliegue: Si el modelo se considera adecuado, se puede implementar en aplicaciones o sistemas para realizar tareas automatizadas, como reconocimiento de voz, detección de fraudes, recomendación de productos, diagnóstico médico y en hidroinformática a la gestión del agua y la comprensión de los sistemas hidrológicos.

El machine learning se utiliza en una amplia variedad de campos y sus aplicaciones son diversas y están en constante expansión debido a su capacidad para abordar problemas complejos y encontrar patrones en grandes conjuntos de datos. También, es una tecnología en constante evolución con un gran potencial para transformar industrias y mejorar la eficiencia en muchas áreas.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

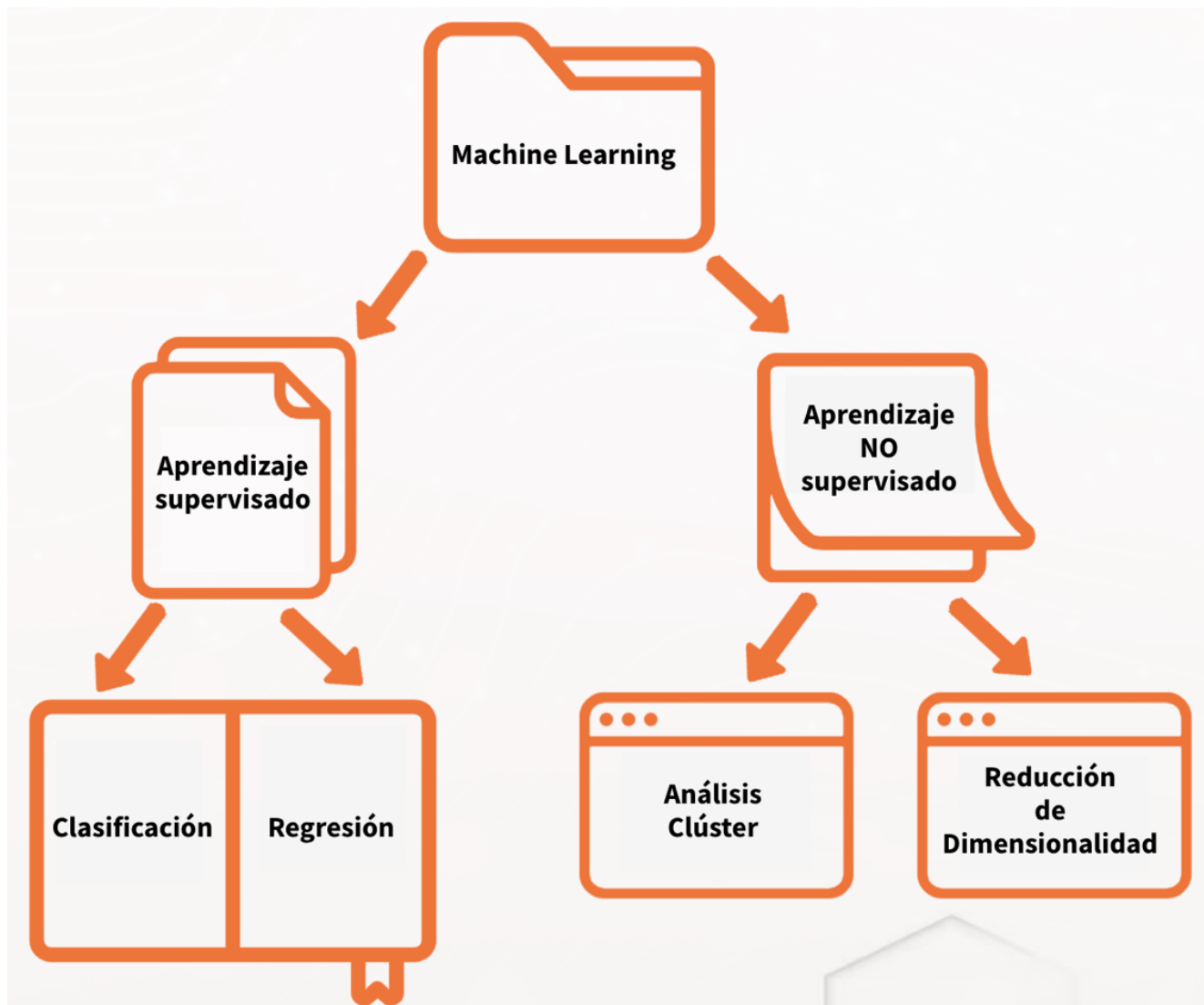
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828



Fuente: <https://www.diegocalvo.es/aprendizaje-supervisado-y-no-supervisado/>

📍 Sede Quirinal: Calle 21 No. 6 - 01
📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220
📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699
✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co
Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.2. ¿Qué es aprendizaje Supervisado?

El aprendizaje supervisado es un enfoque fundamental en el campo del aprendizaje automático (machine learning) en el que un modelo o algoritmo se entrena utilizando un conjunto de datos que contiene ejemplos de entrada (datos de entrada) junto con sus correspondientes salidas deseadas o etiquetas. El objetivo principal del aprendizaje supervisado es aprender una relación funcional entre las entradas y las salidas de manera que el modelo pueda hacer predicciones precisas sobre nuevas entradas basadas en el patrón aprendido de los datos de entrenamiento.

Las características clave del aprendizaje supervisado incluyen:

- **Datos de Entrenamiento Etiquetados:** En el aprendizaje supervisado, se proporcionan ejemplos de datos de entrenamiento que constan de pares de entrada-salida. Estas salidas deseadas se denominan etiquetas o respuestas correctas.
- **Aprendizaje de Patrones:** El modelo se entrena para identificar patrones y relaciones entre las entradas y las salidas. El objetivo es aprender una función matemática que mapee las entradas a las salidas de manera efectiva.
- **Predicciones y Generalización:** Una vez entrenado, el modelo puede realizar predicciones sobre nuevas entradas que no se vieron durante el entrenamiento. El objetivo es que el modelo generalice a partir de los datos de entrenamiento para hacer predicciones precisas en datos previamente desconocidos.
- **Tipos de Problemas:** El aprendizaje supervisado se utiliza en una variedad de problemas, como clasificación (donde se asigna una etiqueta a una entrada), regresión (donde se predice un valor numérico), y otros problemas de predicción y toma de decisiones.

Ejemplos comunes de aplicaciones de aprendizaje supervisado incluyen:

Clasificación de correo electrónico como spam o no spam.

Predicción de precios de bienes raíces en función de características como tamaño, ubicación y número de habitaciones.

Diagnóstico médico basado en síntomas y resultados de pruebas.

Reconocimiento de dígitos escritos a mano en aplicaciones de reconocimiento de caracteres.

El aprendizaje supervisado es una de las formas más utilizadas y poderosas de aprendizaje automático, ya que permite a las máquinas aprender de manera efectiva a partir de datos etiquetados y realizar tareas de toma de decisiones en una amplia variedad de dominios.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989

NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.3. ¿Qué son los algoritmos de clasificación?

Los algoritmos de clasificación son una categoría de técnicas de aprendizaje automático que se utilizan para asignar etiquetas o categorías a datos de entrada en función de ciertas características o atributos. Estos algoritmos son ampliamente utilizados en una variedad de aplicaciones para tomar decisiones automatizadas y realizar tareas de organización y categorización. Aquí hay algunas de las principales aplicaciones y propósitos de los algoritmos de clasificación:

1. Clasificación de Documentos: Los algoritmos de clasificación se utilizan en el procesamiento de texto y la minería de texto para categorizar documentos, correos electrónicos o mensajes en diferentes categorías, como spam o no spam, categorías de noticias, opiniones positivas o negativas, etc.

2. Detección de Fraudes: En el sector financiero, los algoritmos de clasificación se emplean para detectar actividades fraudulentas, como transacciones de tarjetas de crédito fraudulentas o solicitudes de préstamos sospechosas, clasificándolas como legítimas o fraudulentas.

3. Recomendación de Contenido: Los sistemas de recomendación, como los utilizados por plataformas de streaming de video o sitios de comercio electrónico, utilizan algoritmos de clasificación para sugerir productos, películas, música u otros contenidos basados en el historial y las preferencias del usuario.

4. Diagnóstico Médico: En aplicaciones médicas, los algoritmos de clasificación se utilizan para ayudar en el diagnóstico de enfermedades. Por ejemplo, pueden clasificar imágenes médicas, como radiografías o escaneos de resonancia magnética, para identificar patologías.

5. Clasificación de Imágenes y Video: Los algoritmos de clasificación se utilizan en aplicaciones de visión por computadora para identificar y clasificar objetos, rostros, emociones o actividades en imágenes o secuencias de video.

6. Filtrado de Spam: En el ámbito de la comunicación digital, los algoritmos de clasificación se utilizan para filtrar correos electrónicos no deseados (spam) y evitar que lleguen a la bandeja de entrada de los usuarios.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989

NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

7. Clasificación de Texto: Se utilizan para analizar y categorizar texto en redes sociales, comentarios de usuarios y reseñas de productos, lo que puede ser útil para el análisis de sentimientos, la detección de temas y la toma de decisiones comerciales.

8. Segmentación de Clientes: En marketing y análisis de negocios, los algoritmos de clasificación ayudan a segmentar clientes en grupos o perfiles basados en sus características y comportamiento, lo que permite una personalización más efectiva de las estrategias de marketing.

9. Calidad del agua (hidroinformática): Un ejemplo de clasificación en hidroinformática utilizando Machine Learning podría ser la clasificación de la calidad del agua. En esta aplicación, se utilizarían datos recopilados de cuerpos de agua, como ríos, lagos o embalses, junto con mediciones de diferentes parámetros de calidad del agua, para predecir si el agua es apta para el consumo humano o tiene presencia de algún contaminante.

En resumen, los algoritmos de clasificación son herramientas fundamentales para organizar, etiquetar y tomar decisiones automatizadas en una amplia gama de aplicaciones, lo que los convierte en una parte esencial de la inteligencia artificial y el aprendizaje automático. Permiten a las máquinas aprender a reconocer patrones y tomar decisiones basadas en datos, lo que aporta eficiencia y automatización a numerosos procesos y sistemas.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.4. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA, por sus siglas en inglés, Exploratory Data Analysis) es una etapa fundamental en el proceso de análisis de datos que tiene como objetivo entender y analizar un conjunto de datos antes de aplicar técnicas más avanzadas de modelado o inferencia. Su propósito principal es descubrir patrones, tendencias, relaciones y características clave en los datos, lo que proporciona una visión general completa y detallada de la información contenida en ellos.

Aquí hay algunos aspectos clave del análisis exploratorio de datos:

1. Resumen de Datos: Se realizan resúmenes estadísticos básicos para cada variable en el conjunto de datos. Esto incluye estadísticas descriptivas como la media, la mediana, la desviación estándar, los valores mínimos y máximos. Estos resúmenes proporcionan una idea inicial de la distribución de los datos.

2. Visualización de Datos: Se utilizan gráficos y visualizaciones para representar los datos. Los gráficos, como histogramas, diagramas de dispersión, diagramas de caja y gráficos de barras, ayudan a identificar patrones visuales en los datos, como la presencia de outliers (valores atípicos), distribuciones, correlaciones y más.

3. Limpieza de Datos: Se identifican y manejan outliers, valores faltantes y datos duplicados. Esto es importante para garantizar que los datos estén limpios y listos para su análisis.

4. Análisis de Correlación: Se examinan las relaciones entre las diferentes variables del conjunto de datos. Esto puede ayudar a identificar posibles variables predictoras y comprender cómo interactúan entre sí.

5. Análisis de Distribución: Se analiza la distribución de los datos para comprender cómo están dispersos. La normalidad de la distribución es un factor importante en muchos enfoques estadísticos.

6. Segmentación y Agrupamiento: Se pueden utilizar técnicas de agrupamiento para identificar grupos o patrones naturales en los datos, lo que puede ayudar a segmentar la información y aportar insights adicionales.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



Fuente: <https://www.excelr.com/exploratory-data-analysis-in-data-science#>

El análisis exploratorio de datos es una etapa crucial en el proceso de toma de decisiones basadas en datos, ya que proporciona información valiosa sobre los datos subyacentes y ayuda a definir la dirección del análisis posterior. Además, puede ayudar a identificar problemas en los datos que requieren corrección antes de realizar análisis más avanzados o construir modelos predictivos. En última instancia, el EDA es una herramienta esencial para comprender y obtener insights de los datos antes de realizar cualquier análisis o toma de decisiones significativos.

Ejemplo en Python:

EDA - Análisis Exploratorio de Datos

El conjunto de datos seleccionado presenta registros del crecimiento de la población en los años (1952 - 2007)

Diccionario de Datos

Variable	Tipo de dato	Definición
country	Cadena	País donde se originaron los registros
year	Entero	Año en que se tomo el número de población
population	Entero	Número de población

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

Paso 1: Importar las librerías

```
1 # Librería para operaciones Matemáticas o Estadísticas
2 import numpy as np
3 # Librería para el manejo de datos
4 import pandas as pd
5 # Librería para gráficas
6 import matplotlib.pyplot as plt
```

Paso 2: Cargar los datos en un DataFrame

```
1 # Se lee el archivo plano y se carga en un DataFrame
2 df = pd.read_csv("data/1.4-EDA.csv")
3 # Se imprime los primeros 5 registros
4 print(df.head(5))
```

	country	year	population
0	Afghanistan	1952.0	8425333.0
1	Afghanistan	1957.0	9240934.0
2	Afghanistan	1962.0	10267083.0
3	Afghanistan	1967.0	11537966.0
4	Afghanistan	1972.0	13079460.0

Paso 3: Exploramos los datos

```
1 # Se imprime el número de Filas y Columnas
2 print("Filas, Columnas")
3 print(df.shape)
```

Filas, Columnas
(1704, 3)

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
1 # Se identifican los valores NaN del DataFrame
2 print("Columna      Cantidad NaN")
3 print(df.isnull().sum(axis = 0))
```

Columna	Cantidad NaN
country	0
year	72
population	72

dtype: int64

```
1 # Se eliminan los valores NaN del DataFrame porque generan ruido
2 data = df.dropna()
3 # Se imprime el número de Filas y Columnas
4 print("Filas, Columnas")
5 print(data.shape)
```

Filas, Columnas
(1632, 3)

```
1 # Se observan las estadísticas de los datos (mínimo, máximo, media, SD, mediana)
2 data.describe()
```

	year	population
count	1632.000000	1.632000e+03
mean	1979.500000	3.014837e+07
std	17.265553	1.083943e+08
min	1952.000000	6.001100e+04
25%	1965.750000	2.748356e+06
50%	1979.500000	6.962964e+06
75%	1993.250000	1.859411e+07
max	2007.000000	1.318683e+09

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

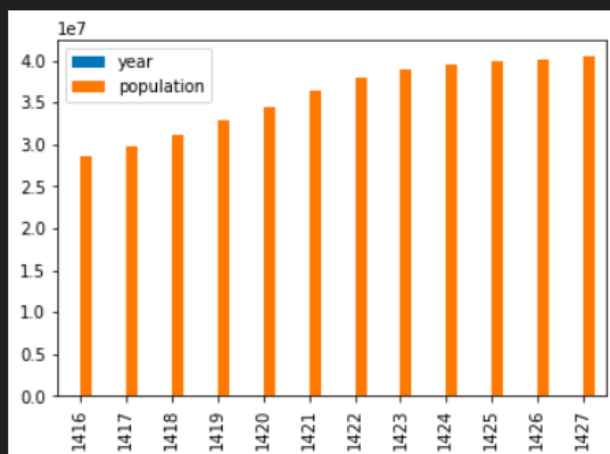
INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
1 # Se imprime los datos para el país de España
2 data_españa = data[data['country'] == 'Spain']
3 print(data_españa.head())
```

	country	year	population
1416	Spain	1952.0	28549870.0
1417	Spain	1957.0	29841614.0
1418	Spain	1962.0	31158061.0
1419	Spain	1967.0	32850275.0
1420	Spain	1972.0	34513161.0

```
1 # Se genera una gráfica de barras con los datos del país España
2 data_españa.plot(kind='bar')
```

<AxesSubplot:>



📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



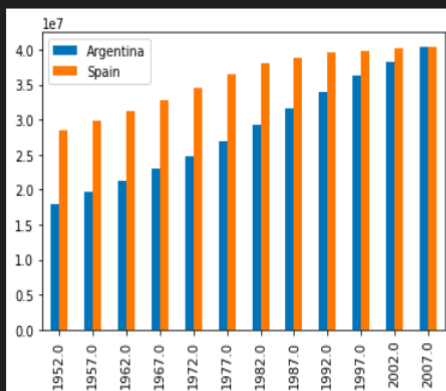
CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
1 # Ahora se compara el crecimiento de la población entre España y Argentina
2 data_argentina = data[(data['country'] == 'Argentina')]
3
4 # Se ajusta el eje x con los años correspondientes
5 anios = data_espana['year'].unique()
6 # Se consultan los valores de la población
7 poblacion_espana = data_espana['population'].values
8 poblacion_argentina = data_argentina['population'].values
9
10 # Se genera la gráfica de barras para la población de argentina y españa
11 data_grafica = pd.DataFrame({'Argentina': poblacion_argentina, 'Spain': poblacion_espana}, index=anios)
12 data_grafica.plot(kind='bar')
```

<AxesSubplot:>



Puede ver el código fuente de este ejemplo en: https://github.com/jose-llanos/ML_hidroinformatica/blob/main/1.4-EDA.ipynb

Video: Ejemplo EDA

https://www.youtube.com/watch?v=-KW4gT_oGU

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.5. Preprocesamiento de Datos

El preprocesamiento de datos es una fase crítica en el proceso de análisis de datos y aprendizaje automático que implica la preparación y limpieza de los datos brutos antes de que sean utilizados en análisis más avanzados o en la construcción de modelos predictivos.

Esta etapa es esencial para garantizar que los datos estén en condiciones óptimas para su análisis y que los resultados sean confiables y significativos. El preprocesamiento de datos aborda una serie de desafíos comunes en los datos, que incluyen:

1. Limpieza de Datos: Esto implica identificar y tratar valores atípicos (outliers) o valores que faltan (missing values). Los valores atípicos pueden distorsionar el análisis y la modelización, mientras que los valores faltantes pueden causar problemas en los algoritmos de aprendizaje automático.

2. Transformación de Datos: A menudo, es necesario transformar los datos para que cumplan con los supuestos de ciertos algoritmos o para mejorar su interpretación. Esto puede incluir la normalización (escalar los datos a una escala común), la estandarización (centrar y escalar los datos para que tengan media cero y desviación estándar uno), o la codificación de variables categóricas en formatos numéricos.

3. Selección de Características: En ocasiones, los conjuntos de datos contienen muchas características o variables que pueden no ser todas relevantes para la tarea en cuestión. La selección de características implica identificar y mantener solo las características más importantes o relevantes para reducir la dimensionalidad de los datos y mejorar el rendimiento del modelo.

4. Manejo de Datos Desbalanceados: En problemas de clasificación, puede haber desequilibrios en las clases, lo que significa que una clase tiene muchas más muestras que la otra. Esto puede afectar negativamente la capacidad del modelo para aprender y predecir correctamente. El preprocesamiento puede incluir técnicas para abordar este desequilibrio.

5. Tratamiento de Datos Categóricos: Los algoritmos de aprendizaje automático suelen requerir que las variables categóricas sean convertidas en representaciones numéricas adecuadas, como la codificación one-hot o la codificación ordinal.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

6. Reducción de Ruido: A veces, los datos pueden contener ruido o información irrelevante que puede dificultar la precisión del análisis. El preprocesamiento puede incluir la eliminación de ruido o datos irrelevantes.

7. Manejo de Valores Extremos: Dependiendo de la aplicación, los valores extremos pueden ser tratados de diversas maneras, desde su eliminación hasta su transformación o inclusión en análisis específicos.

El objetivo general del preprocesamiento de datos es asegurar que los datos sean de alta calidad, coherentes y adecuados para su uso en análisis o modelado. Esto contribuye a la robustez y la confiabilidad de los resultados finales y facilita la creación de modelos más precisos y útiles en el aprendizaje automático y el análisis de datos.

Ejemplo en Python:

Preprocesamiento de Datos

En este ejemplo se utiliza python para realizar el preprocesamiento de los siguientes datos:

Variable	Tipo	Descripción
clima	Cadena	Es el clima actual de la ciudad (soleado, nublado, lluvioso)
temperatura	Cadena	Temperatura actual de la ciudad (caliente, templado, frio)
Sequia	Cadena	Describe la existencia de sequía para la ciudad (si, no)

El conjunto de datos tiene 60 registros y los pasos que seguiremos son los siguientes:

1. **Revisión de los datos:** Verificamos la estructura y contenido de los datos.
2. **Limpieza de datos:** Eliminación de valores nulos y corrección de errores tipográficos.
3. **Conversión de tipos de datos:** Aseguramos que cada columna tenga el tipo de dato adecuado.
4. **Codificación de variables categóricas:** Convertimos las variables categóricas en variables numéricas si es necesario.
5. **Normalización/Estandarización:** Ajuste de los valores numéricos a una escala común.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 1: Revisión de los datos

Revisamos la estructura y el contenido del archivo.

```
[1] # Importamos la librería para el manejo de los datos
import pandas as pd

# Cargamos el archivo CSV a un Dataframe
data = pd.read_csv('1.5-clima.csv')

# Mostrar los datos
data
```



10	soleado	templado	si
11	nublado	templado	si
12	nublado	caliente	si
13	luvioso	templado	no
14	soleado	caliente	no
15	soleado	caliente	si
16	nublado	caliente	si
17	lluvioso	templado	si
18	luvioso	frio	no
19	lluvioso	frio.	no
20	nublado	frio	si
21	solead	templado	no
22	soleado	frio	si
23	lluvioso	templado	si
24	soleado	templado	si

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 2: Limpieza de datos

Eliminación de valores nulos y corrección de errores tipográficos.

```
[8] # Verificar valores nulos
    print(data.isnull().sum())

    # Se eliminan valores nulos
    data = data.dropna()

    # Se muestran los registros sin los valores nulos
    print("-----")
    data
```

```
clima      0
temperatura 0
sequia      0
dtype: int64
```

	clima	temperatura	sequia
0	soleado	caliente	si
1	soleado	caliente	si
2	nublado	caliente	si
3	lluvioso	templado	si
4	lluvioso	frio	no
5	lluvioso	frio	no
6	nublado	frio	si
7	soleado	templado	no
8	soleado	frio	si
9	lluvioso	templado	si

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
# Corregir errores tipográficos
data['clima'] = data['clima'].replace('luvioso', 'lluvioso')
data['clima'] = data['clima'].replace('solead', 'soleado')
data['temperatura'] = data['temperatura'].replace('frio.', 'frio')

# Mostrar los datos corregidos (10 primeras filas)
data.head(10)
```

	clima	temperatura	sequia
0	soleado	caliente	si
1	soleado	caliente	si
2	nublado	caliente	si
3	lluvioso	templado	si
4	lluvioso	frio	no
5	lluvioso	frio	no
6	nublado	frio	si
7	soleado	templado	no
8	soleado	frio	si
9	lluvioso	templado	si

▼ Paso 3: Conversión de tipos de datos

Aseguramos que cada columna tenga el tipo de dato adecuado, como los datos son de tipo cadena no es necesario hacer conversión.

```
[10] # Verificar y convertir tipos de datos si es necesario
print(data.dtypes)
```

clima	object
temperatura	object
sequia	object
dtype:	object

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



✓ Paso 4: Codificación de variables categóricas

Convertimos las variables categóricas a numéricas usando One-Hot Encoding.

```
# Codificar variables categóricas
data_encoded = pd.get_dummies(data, columns=['clima', 'temperatura'])

# Mostrar los datos codificados (5 primeras filas)
data_encoded.head(5)
```

	sequia	clima_lluvioso	clima_nublado	clima_soleado	temperatura_caliente	temperatura_frio	temperaturatemplado
0	si	False	False	True	True	False	False
1	si	False	False	True	True	False	False
2	si	False	True	False	True	False	False
3	si	True	False	False	False	False	True
4	no	True	False	False	False	True	False

Paso 5: Normalización/Estandarización

En este caso, no hay variables numéricas para normalizar.

Conclusión

Este preprocesamiento nos permite preparar los datos para su uso en análisis y modelos predictivos, mejorando la calidad y la estructura de los datos.

Puede ver el código fuente de este ejemplo en: https://github.com/jose-llanos/ML_hidroinformatica/blob/main/1.5-Preprocesamiento.ipynb

Videos: Ejemplos Preprocesamiento

Variables Dummy: <https://www.youtube.com/watch?v=DAQhQFp-sCg>

Detección de Outliers: <https://www.youtube.com/watch?v=P8ls1Wjmpkw>

Reemplazando los Valores Perdidos:
<https://www.youtube.com/watch?v=bY7OIJvTMrE>

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.6. Naive Bayes

Naive Bayes es un algoritmo de aprendizaje automático supervisado basado en el teorema de Bayes, que se utiliza comúnmente para tareas de clasificación y modelado probabilístico. El enfoque "naive" (ingenuo) en Naive Bayes se refiere a una suposición simplificada pero efectiva: asume que las características (o atributos) que se utilizan para la clasificación son independientes entre sí, aunque en la realidad esto puede no ser cierto. A pesar de esta suposición simplificada, el algoritmo de Naive Bayes suele funcionar sorprendentemente bien en una amplia variedad de aplicaciones prácticas.

El algoritmo de Naive Bayes se basa en el teorema de Bayes, que es una fórmula de probabilidad condicional que permite calcular la probabilidad de una causa dadas ciertas evidencias. En el contexto de la clasificación, se utiliza para calcular la probabilidad de que una instancia de datos pertenezca a una clase específica dado un conjunto de características observadas.

Aquí hay algunos puntos clave sobre Naive Bayes:

- **Suposición de Independencia Condicional:** Naive Bayes asume que las características son independientes entre sí dadas las clases. Esto significa que la probabilidad de que una instancia pertenezca a una clase se calcula como el producto de las probabilidades condicionales de cada característica dada esa clase.
- **Tres Variantes Principales:** Hay tres variantes comunes de Naive Bayes: Naive Bayes Gaussiano (para características numéricas continuas), Naive Bayes Multinomial (para características categóricas o conteos) y Naive Bayes Bernoulli (para características binarias).
- **Fácil de Implementar y Eficiente:** Naive Bayes es fácil de implementar y computacionalmente eficiente, lo que lo hace adecuado para conjuntos de datos grandes.
- **Aplicaciones:** Se utiliza en una amplia variedad de aplicaciones, incluyendo la clasificación de spam de correo electrónico, la categorización de documentos, el análisis de sentimientos, la detección de enfermedades médicas, la clasificación de documentos, entre otros.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

- **Ventajas:** Es rápido, eficiente y puede funcionar bien en conjuntos de datos con muchas características. Es especialmente útil cuando se dispone de pocos datos de entrenamiento.
- **Desventajas:** La suposición de independencia condicional puede ser problemática en conjuntos de datos donde las características están correlacionadas. Puede no ser el mejor enfoque cuando la relación entre las características y las clases es compleja.

A pesar de su simplicidad y suposiciones simplificadas, Naive Bayes es una herramienta poderosa y ampliamente utilizada en el aprendizaje automático, y a menudo sirve como un punto de referencia inicial para tareas de clasificación.

Ejemplo en Python:

Algoritmo: Naive Bayes

En este ejemplo se utiliza el algoritmo de clasificación: **Naive Bayes** para predecir si existe sequía o no en una ciudad a partir del clima y la temperatura, los datos son:

Variable	Tipo	Descripción
clima	Cadena	Es el clima actual de la ciudad (soleado, nublado, lluvioso)
temperatura	Cadena	Temperatura actual de la ciudad (caliente, templado, frío)
Sequía	Cadena	Describe la existencia de sequía para la ciudad (sí, no)

Nota: Los datos de este ejemplo ya tienen preprocesamiento.

Paso 1: Importar las librerías

```
[1] # Librería para operaciones matemáticas o estadísticas
import numpy as np
# Librería para manejo de datos
import pandas as pd
# Librerías para gráficas
import matplotlib.pyplot as plt
import seaborn as sb
# Librería para transformar datos
from sklearn.preprocessing import LabelEncoder
# Librería para separar el conjunto de datos (entrenamiento y pruebas)
from sklearn.model_selection import train_test_split
# Librería para Naive Bayes
from sklearn.naive_bayes import GaussianNB
# Librería para generar la métrica: matriz de confusión
from sklearn.metrics import confusion_matrix
# Librería para generar el reporte de clasificación
from sklearn.metrics import classification_report
```

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 2: Cargar el archivo plano en un DataFrame

```
[2] # Se cargan los datos del archivo plano: '1.6-clima.csv' a un DataFrame  
data = pd.read_csv("1.6-clima.csv")
```

✓ Paso 3: Explorar los datos

```
[3] # Se muestran los primeros 10 registros del DataFrame  
data.head(10)
```



	clima	temperatura	sequia
0	soleado	caliente	si
1	soleado	caliente	si
2	nublado	caliente	si
3	lluvioso	templado	si
4	lluvioso	frio	no
5	lluvioso	frio	no
6	nublado	frio	si
7	soleado	templado	no
8	soleado	frio	si
9	lluvioso	templado	si



📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación
y extensión de programas de pregrado, aplicando todos los
requisitos de las normas ISO implementadas en sus sedes
Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

▼ Paso 4: Procesar los datos

```
✓ 0 s [4] # Se reemplazan los valores de tipo cadena a numéricos para generar la predicción  
le = LabelEncoder()
```

```
data['clima'] = le.fit_transform(data['clima'])  
data['temperatura'] = le.fit_transform(data['temperatura'])  
data['sequia'] = le.fit_transform(data['sequia'])
```

data



clima temperatura sequia

0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
4	0	1	0



```
✓ 0 s [5] # Se cuentan los registros con valor 1 y 0 (es decir si existe o no sequia)  
sequia_si = np.sum(data['sequia'] == 1)  
sequia_no = np.sum(data['sequia'] == 0)  
  
print("Sequía si =", sequia_si)  
print("Sequía no =", sequia_no)
```



Sequía si = 32
Sequía no = 18

```
✓ 0 s [6] # También se pueden agrupar y traer su tamaño (0- No sequia, 1- Si sequia)  
print(data.groupby('sequia').size())
```



sequia
0 18
1 32
dtype: int64

NOTA: Los datos están desbalanceados, se debería aplicar la técnica de balanceo que incluye **Oversampling** o **undersampling**. Sin embargo, este proceso lo realizaremos con el algoritmo: Árbol de Decisión (siguiente ejemplo).

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



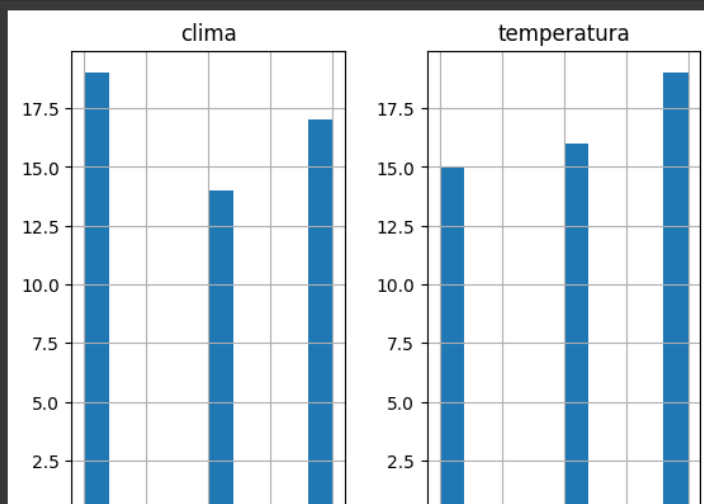
CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

▼ Paso 5: Generar las visualizaciones

```
[7] # Se genera el histograma de los datos, pero se excluye la clase (sequía) porque es lo que queremos predecir.  
data.drop(['sequia'], axis=1).hist() # La columna se elimina de manera temporal, es decir solo para la gráfica  
plt.rcParams['figure.figsize'] = (10, 10)  
plt.show()
```



```
[8] # Se verifica que todas las columnas están en el DataFrame data  
data.head()
```



	clima	temperatura	sequia
0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
4	0	1	0



📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[9] # Se generan las estadísticas de los datos
data.describe()
```

	clima	temperatura	sequia
count	50.00000	50.00000	50.00000
mean	0.96000	1.08000	0.64000
std	0.85619	0.82906	0.48487
min	0.00000	0.00000	0.00000
25%	0.00000	0.00000	0.00000
50%	1.00000	1.00000	1.00000
75%	2.00000	2.00000	1.00000
max	2.00000	2.00000	1.00000

NOTA: Las estadísticas permiten identificar: menor valor, mayor valor, media, desviación estándar (SD), mediana y cuantiles para las columnas numéricas del DataFrame.

✓ Paso 6: Generar la predicción con el algoritmo Gaussian Naive Bayes de la librería SKLearn

```
[10] # Se deja en X todas las características para el modelo
features = ['clima', 'temperatura']
X = data[features]
# Se deja en y la clase (sequia) porque es lo que queremos predecir
y = data['sequia'].values

# Dividimos el conjunto de datos en entrenamiento (80%) y pruebas (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, test_size=0.2, random_state=6)

# Creamos el modelo, lo ponemos a aprender con fit() y obtenemos predicciones sobre nuestro conjunto de test
# Se instancia el clasificador
nb = GaussianNB()
# Se entrena el clasificador
nb.fit(X_train, y_train)
# Se genera la predicción
prediccion = nb.predict(X_test)

# Se imprime la matriz de confusión
print(confusion_matrix(y_test, prediccion))
# Se imprime la Accuracy del modelo
print(classification_report(y_test, prediccion))
```

```
[[0 2]
 [2 6]]
```

precision recall f1-score support

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
# Se imprime la matriz de confusión
print(confusion_matrix(y_test, prediccion))
# Se imprime la Accuracy del modelo
print(classification_report(y_test, prediccion))
```

		precision	recall	f1-score	support
	0	0.00	0.00	0.00	2
	1	0.75	0.75	0.75	8
accuracy				0.60	10
macro avg		0.38	0.38	0.38	10
weighted avg		0.60	0.60	0.60	10

NOTA: Se puede observar que el accuracy del modelo es del 0.60%. Sin embargo, esta métrica no es tan precisa como: F1 score, que requiere de la **precisión** y el **recall**. En el próximo ejercicio utilizaremos estas métricas (ejemplo árbol de decisión).

```
[11] # Se genera una predicción con datos nuevos
# clima = 2 (soleado)
# temperatura = 1 (caliente)
print( nb.predict([[2, 1]]) )
#Resultado esperado 0-No sequía, 1-Sequia
# Para el ejemplo se predice que hay sequía [1]
```

Puede ver el código fuente de este ejemplo en:

https://github.com/jose-llanos/ML_hidroinformatica/blob/main/1.6-NaiveBayes.ipynb

Videos: Ejemplo Naive Bayes Python

https://www.youtube.com/watch?v=S2_SZ7WRy-Q

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.7. Árbol de Decisión

Un árbol de decisión es un modelo de aprendizaje automático que se utiliza en problemas de clasificación y regresión. Representa una estructura de tipo árbol en la que cada nodo interno representa una pregunta o prueba sobre una característica específica del conjunto de datos, y cada rama representa una posible respuesta a esa pregunta. Las hojas del árbol representan las decisiones o predicciones finales.

El árbol de decisión es una técnica popular en aprendizaje automático debido a su capacidad para manejar conjuntos de datos complejos y tomar decisiones lógicas basadas en reglas simples. Algunas de sus características clave incluyen:

- 1. Explicabilidad:** Los árboles de decisión son modelos fácilmente interpretables y explicables. Puedes seguir las ramas del árbol para entender cómo se toma una decisión.
- 2. Versatilidad:** Se pueden utilizar en una variedad de problemas, incluyendo clasificación y regresión. Además, pueden manejar tanto características numéricas como categóricas.
- 3. Robustez ante Valores Atípicos:** Son resistentes a valores atípicos y no requieren suposiciones sobre la distribución de los datos.
- 4. Selección de Características:** Los árboles de decisión pueden ayudar a identificar las características más importantes en un conjunto de datos al ubicarlas en niveles superiores del árbol.

El proceso de construcción de un árbol de decisión implica dividir repetidamente el conjunto de datos en subconjuntos más pequeños en función de las características y sus valores, de manera que se maximice la homogeneidad (similaridad) dentro de cada subconjunto y se minimice la heterogeneidad entre los subconjuntos. Esto se hace de manera recursiva hasta que se cumple un criterio de parada, como la profundidad máxima del árbol o un límite en la cantidad mínima de ejemplos en una hoja.

Algunos ejemplos de aplicaciones de árboles de decisión incluyen:

- **Clasificación de spam de correo electrónico:** Decidir si un correo electrónico es spam o no en función de sus características.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989

NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

- **Predicción de enfermedades médicas:** Determinar si un paciente tiene cierta enfermedad o no en función de sus síntomas y resultados de pruebas.
- **Predicción de precios de bienes raíces:** Predecir el precio de una propiedad en función de características como ubicación, tamaño y antigüedad.
- **Recomendación de productos:** Recomendar productos a los clientes en función de su historial de compras y preferencias.

En resumen, un árbol de decisión es una herramienta versátil y fácil de interpretar que se utiliza en una amplia variedad de problemas de aprendizaje automático y toma de decisiones. Su estructura de tipo árbol permite dividir un problema en decisiones lógicas y puede ser una excelente opción tanto para problemas simples como para aplicaciones más complejas.

Ejemplo en Python:

✓ Algoritmo: Árbol de Decisión (Clasificación)

En este ejemplo se utiliza el algoritmo de clasificación: **Árbol de Decisión** para predecir si existe sequía o no en una ciudad a partir del clima y la temperatura, los datos son:

Variable	Tipo	Descripción
clima	Cadena	Es el clima actual de la ciudad (soleado, nublado, lluvioso)
temperatura	Cadena	Temperatura actual de la ciudad (caliente, templado, frío)
Sequía	Cadena	Describe la existencia de sequía para la ciudad (sí, no)

Nota: Los datos de este ejemplo ya tienen preprocesamiento.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 1: Importar las librerías

```
✓ 3 s [1] # Librería para operaciones matemáticas o estadísticas
import numpy as np
# Librería para el manejo de datos
import pandas as pd
# Librerías para gráficas
import matplotlib.pyplot as plt
import seaborn as sb
# Librería para transformar datos
from sklearn.preprocessing import LabelEncoder
# Librería para el balanceo de los datos
from sklearn.utils import resample
# Librería para separar el conjunto de datos en entrenamiento y pruebas
from sklearn.model_selection import train_test_split
# Librería para Árbol de Decisión (clasificación)
from sklearn.tree import DecisionTreeClassifier
# Librerías para métricas del modelo
from sklearn.metrics import precision_score, recall_score, f1_score
```

✓ Paso 2: Cargar el archivo plano en un DataFrame

```
✓ 0 s [2] # Se cargan los datos del archivo plano: 1.6-clima.csv a un DataFrame
df = pd.read_csv("1.6-clima.csv")
```

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 3: Precesamiento de los datos

```
[3] # Se muestran los primeros 10 registros del DataFrame  
df.head(10)
```

	clima	temperatura	sequia
0	soleado	caliente	si
1	soleado	caliente	si
2	nublado	caliente	si
3	lluvioso	templado	si
4	lluvioso	frio	no
5	lluvioso	frio	no
6	nublado	frio	si
7	soleado	templado	no
8	soleado	frio	si
9	lluvioso	templado	si

```
[4] # Se identifican los valores NaN del DataFrame  
print("Columna Cantidad NaN")  
print(df.isnull().sum(axis = 0))  
  
#df.dropna()
```

Columna	Cantidad NaN
clima	0
temperatura	0
sequia	0
dtype: int64	

✓ NOTA: Como no hay valores NaN en las columnas, no es necesario aplicar 'dropna()' para eliminarlos

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[5] # Se reemplazan los valores de tipo cadena a numéricos para generar la predicción
le = LabelEncoder()

df['clima'] = le.fit_transform(df['clima'])
df['temperatura'] = le.fit_transform(df['temperatura'])
df['sequia'] = le.fit_transform(df['sequia'])

df
```

	clima	temperatura	sequia
0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
4	0	1	0
5	0	1	0
6	1	1	1
7	2	2	0

```
[6] # Se cuentan los registros con valor 1 y 0 (es decir si existe o no sequía)
sequia_si = np.sum(df['sequia'] == 1)
sequia_no = np.sum(df['sequia'] == 0)

print("Sequía si =", sequia_si)
print("Sequía no =", sequia_no)
```

```
Sequía si = 32
Sequía no = 18
```

📍 Sede Quirinal: Calle 21 No. 6 - 01
📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220
📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699
✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co
Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[7] # Se genera el balanceo de los datos
df_alto = df[df['sequia'] == 1] # Porque hay 32 registros se define alto
df_bajo = df[df['sequia'] == 0] # Porque hay 18 registros se define bajo

# Aplicamos undersample para dejar los 18 registros en 32
data_resample_bajo = resample(df_bajo,
                              replace = True,
                              n_samples = 32,
                              random_state = 1)

# Se concatenan (unen) los datos df_alto y los del resample
data = pd.concat([df_alto, data_resample_bajo])
# Se observan la cantidad de registros para 1- Sequia Si y 0- Sequia No
data['sequia'].value_counts()
```

sequia	count
1	32
0	32

dtype: int64

▼ Paso 4: Generar las visualizaciones

```
[8] # Se genera el histograma de los datos, pero se excluye la clase (sequia)
data.drop(['sequia'], axis=1).hist() # La columna se elimina de manera temporal, es decir solo la gráfica
plt.rcParams['figure.figsize'] = (10, 10)
plt.show()
```

clima	count
0	29
1	14
2	21

temperatura	count
10	18
20	24
30	22

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[9] # Se verifica que todas las columnas están en el DataFrame  
data.head()
```

	clima	temperatura	sequia
0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
6	1	1	1

```
[10] # Se generan las estadísticas del DataFrame  
data.describe()
```

	clima	temperatura	sequia
count	64.000000	64.000000	64.000000
mean	0.875000	1.062500	0.500000
std	0.881917	0.794325	0.503953
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	1.000000	1.000000	0.500000
75%	2.000000	2.000000	1.000000
max	2.000000	2.000000	1.000000

Nota: En count, se observan que los datos están balanceados (64 registros, 32 (1- sequia si) y 32 (0- sequia no)

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

▼ Paso 5: Predicción con el algoritmo Árbol de Decisión

```
[11] # Se deja en X todas las características para el modelo
features = ['clima', 'temperatura']
X = data[features]
# Se deja en y la clase (sequia) porque es lo que queremos predecir
y = data['sequia'].values

# Dividimos el conjunto de datos en entrenamiento (80%) y pruebas (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=6)

# Creamos el modelo, lo ponemos a aprender con fit() y obtenemos predicciones sobre nuestro conjunto de test
# Se instancia el clasificador
dtc = DecisionTreeClassifier()
# Se entrena el modelo
dtc.fit(X_train, y_train)
# Se genera la predicción
prediccion = dtc.predict(X_test)
# Métricas clasificación: Precisión, Recall, F1-Score
print("Precisión: ", round(precision_score(y_test, prediccion, average='weighted'), 2))
print("Recall: ", round(recall_score(y_test, prediccion, average='weighted'), 2))
print("F1-Score: ", round(f1_score(y_test, prediccion, average='weighted'), 2))
```

Precisión: 0.76
Recall: 0.77
F1-Score: 0.76

Puede ver el código fuente de este ejemplo en:

https://github.com/jose-llanos/ML_hidroinformatica/blob/main/1.7-ArbolDecision.ipynb

Videos: Ejemplo Árbol de Decisión Python

<https://www.youtube.com/watch?v=1ckCpOFuQV0>

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

1.8. Gradient Boosting Classifier

El Gradient Boosting Classifier (Clasificador de Impulso de Gradiente) es un algoritmo de aprendizaje automático que se utiliza comúnmente en tareas de clasificación. Pertenecce a la categoría de modelos de ensamble, lo que significa que combina múltiples modelos de aprendizaje débil (generalmente árboles de decisión poco profundos) para formar un modelo más fuerte y preciso.

La idea principal detrás del Gradient Boosting Classifier es corregir los errores cometidos por modelos anteriores, de manera que cada nuevo modelo se enfoque en las instancias que fueron mal clasificadas por los modelos anteriores. A continuación, se explican algunos conceptos clave relacionados con el Gradient Boosting Classifier:

1. Aprendizaje Secuencial: A diferencia de otros métodos de ensamble como el Bagging (Bootstrap Aggregating), donde los modelos base se entrenan de forma paralela, el Gradient Boosting entrena los modelos base de forma secuencial.

2. Gradiente Descendente: El nombre "Gradient Boosting" proviene de la técnica de optimización utilizada para minimizar una función de pérdida que mide los errores del modelo actual. Se utiliza el gradiente descendente para ajustar los parámetros del nuevo modelo base de manera que minimicen la pérdida.

3. Modelos Base Débiles: Los modelos base en Gradient Boosting suelen ser árboles de decisión poco profundos (también conocidos como árboles débiles o "stumps"). Estos árboles tienen una profundidad limitada, lo que los hace relativamente simples y evita el sobreajuste.

4. Peso de Aprendizaje (Learning Rate): Es un hiperparámetro que controla la contribución de cada modelo base al modelo final. Un valor más bajo del learning rate hace que el aprendizaje sea más lento y puede mejorar la generalización.

5. Número de Estimadores: Indica cuántos modelos base se van a entrenar en secuencia. Un número mayor de estimadores puede conducir a un modelo más complejo y, a menudo, más preciso, pero también puede aumentar el riesgo de sobreajuste.

6. Regularización: Algunas implementaciones de Gradient Boosting, como XGBoost y LightGBM, incluyen técnicas de regularización para controlar el sobreajuste y mejorar el rendimiento.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

Gradient Boosting Classifier es conocido por su alto rendimiento en una amplia gama de problemas de clasificación, incluyendo problemas desafiantes en los que otras técnicas pueden tener dificultades. Sin embargo, debido a su capacidad para ajustarse a los datos de entrenamiento, es importante afinar adecuadamente los hiperparámetros y aplicar técnicas de validación cruzada para evitar el sobreajuste.

Ejemplo en Python:

✓ Algoritmo: Gradient Boosting Classifier (GBC)

En este ejemplo se utiliza el algoritmo de clasificación: **GBC** para predecir si existe sequía o no en una ciudad a partir del clima y la temperatura, los datos son:

Variable	Tipo	Descripción
clima	Cadena	Es el clima actual de la ciudad (soleado, nublado, lluvioso)
temperatura	Cadena	Temperatura actual de la ciudad (caliente, templado, frío)
Sequía	Cadena	Describe la existencia de sequía para la ciudad (si, no)

Nota: Los datos de este ejemplo ya tienen preprocesamiento.

✓ Paso 1: Importar las librerías

```
[1] # Librería para operaciones matemáticas o estadísticas
import numpy as np
# Librería para el manejo de datos
import pandas as pd
# Librerías para gráficas
import matplotlib.pyplot as plt
import seaborn as sb
# Librería para transformar datos
from sklearn.preprocessing import LabelEncoder
# Librería para el balanceo de los datos
from sklearn.utils import resample
# Librería para separar el conjunto de datos en entrenamiento y pruebas
from sklearn.model_selection import train_test_split
# Librería para Gradient Boosting Classifier
from sklearn.ensemble import GradientBoostingClassifier
# Librerías para métricas del modelo
from sklearn.metrics import precision_score, recall_score, f1_score
# Librería para calcular la media y la desviación estándar utilizadas en las características
from sklearn.preprocessing import StandardScaler
# Librería para manejo de hiperparámetros
from sklearn.model_selection import GridSearchCV
```

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 2: Cargar el archivo plano en un DataFrame

```
[2] # Se cargan los datos del archivo plano: '1.6-clima.csv' a un DataFrame
df = pd.read_csv("1.6-clima.csv")
```

✓ Paso 3: Procesamiento de los datos

```
[3] # Se muestran los primeros 10 registros del DataFrame
df.head(10)
```



	clima	temperatura	sequia
0	soleado	caliente	si
1	soleado	caliente	si
2	nublado	caliente	si
3	lluvioso	templado	si
4	lluvioso	frio	no
5	lluvioso	frio	no
6	nublado	frio	si
7	soleado	templado	no
8	soleado	frio	si
9	lluvioso	templado	si



📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[4] # Se identifican los valores NaN del DataFrame
print("Columna Cantidad NaN")
print(df.isnull().sum(axis = 0))

#df.dropna()
```

```
Columna  Cantidad NaN
clima      0
temperatura  0
sequia     0
dtype: int64
```

✓ NOTA: Como no hay valores NaN en las columnas, no es necesario aplicar 'dropna()' para eliminarlos

```
[5] # Se reemplazan los valores de tipo cadena a numéricos para generar la predicción
le = LabelEncoder()

df['clima'] = le.fit_transform(df['clima'])
df['temperatura'] = le.fit_transform(df['temperatura'])
df['sequia'] = le.fit_transform(df['sequia'])

df
```

	clima	temperatura	sequia
0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
4	0	1	0
5	0	1	0
6	1	1	1
7	2	2	0

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[6] # Se cuentan los registros con valor 1 y 0 (es decir si existe o no sequía)
sequia_si = np.sum(df['sequia'] == 1)
sequia_no = np.sum(df['sequia'] == 0)

print("Sequía si =", sequia_si)
print("Sequía no =", sequia_no)
```

```
Sequía si = 32
Sequía no = 18
```

```
[7] # Se genera el balanceo de los datos
df_alto = df[df['sequia'] == 1] # Porque hay 32 registros se define alto
df_bajo = df[df['sequia'] == 0] # Porque hay 18 registros se define bajo

# Aplicamos undersample para dejar los 18 registros en 32
data_resample_bajo = resample(df_bajo,
                              replace = True,
                              n_samples = 32,
                              random_state = 1)

# Se concatenan (unen) los datos df_alto y los del resample
data = pd.concat([df_alto, data_resample_bajo])
# Se observan la cantidad de registros para 1- Sequia Si y 0- Sequia No
data['sequia'].value_counts()
```

```
count
sequia
1      32
0      32
```

dtype: int64

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



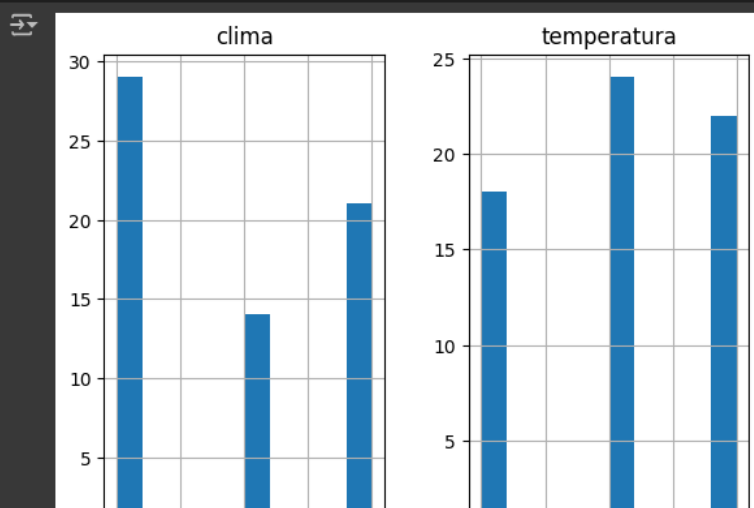
CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación


INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 4: Generar las visualizaciones


```
[8] # Se genera el histograma de los datos, pero se excluye la clase (sequia)  
data.drop(['sequia'], axis=1).hist() # La columna se elimina de manera temporal, es decir solo la gráfica  
plt.rcParams['figure.figsize'] = (10, 10)  
plt.show()
```



```
[9] # Se verifica que todas las columnas están en el DataFrame  
data.head()
```



	clima	temperatura	sequia
0	2	0	1
1	2	0	1
2	1	0	1
3	0	2	1
6	1	1	1



📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA
"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

```
[10] # Se generan las estadísticas del DataFrame
data.describe()
```

	clima	temperatura	sequia
count	64.000000	64.000000	64.000000
mean	0.875000	1.062500	0.500000
std	0.881917	0.794325	0.503953
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	1.000000	1.000000	0.500000
75%	2.000000	2.000000	1.000000
max	2.000000	2.000000	1.000000

▼ Paso 5: Predicción con GBC

```
[11] # Se deja en X todas las características para el modelo
features = ['clima', 'temperatura']
X = data[features]
# Se deja en y la clase (no comprar o comprar) porque es lo que queremos predecir
y = data['sequia'].values

# Dividimos el conjunto de datos en entrenamiento (80%) y pruebas (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=6)

# Creamos el modelo, lo ponemos a aprender con fit() y obtenemos predicciones sobre nuestro conjunto de test
# Se instancia el clasificador de ensemble
gbc = GradientBoostingClassifier()
# Se entrena el modelo
gbc.fit(X_train, y_train)
# Se genera la predicción
prediccion = gbc.predict(X_test)
# Métricas clasificación: Precisión, Recall, F1-Score
print("Precisión: ", round(precision_score(y_test, prediccion, average='weighted'), 2))
print("Recall: ", round(recall_score(y_test, prediccion, average='weighted'), 2))
print("F1-Score: ", round(f1_score(y_test, prediccion, average='weighted'), 2))
```

Precisión: 0.76
Recall: 0.77
F1-Score: 0.76

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

✓ Paso 6: Aplicar hiperparámetros

Aquí se aplica la técnica de hiperparámetros para mejorar la precisión del modelo

```
[12] scaler = StandardScaler()
      X_train = scaler.fit_transform(X_train)
      X_test = scaler.transform(X_test)

      # Algoritmo GBC
      gbc = GradientBoostingClassifier()

      # Hiperparámetros
      learning_rate = [0.01, 0.05, 0.1, 0.15, 0.2]
      criterion = ['friedman_mse', 'squared_error']
      max_depth = [3,5,8]
      max_features = ['log2','sqrt']

      grid = dict(learning_rate = learning_rate,
                  criterion = criterion,
                  max_depth = max_depth,
                  max_features = max_features)

      # Técnica de cuadrícula para hiperparámetros
      grid_search = GridSearchCV(estimator = gbc,
                                param_grid = grid,
                                cv= 10,
                                verbose=1,
                                n_jobs=-1.
```

Puede ver el código fuente de este ejemplo en:

https://github.com/jose-llanos/ML_hidroinformatica/blob/main/1.8-GBC.ipynb

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"



CORHUILA

CORPORACIÓN UNIVERSITARIA DEL HUILA
Vigilada Mineducación

INSTITUCIÓN DE EDUCACIÓN SUPERIOR SUJETA A INSPECCIÓN
Y VIGILANCIA POR EL MINISTERIO DE EDUCACIÓN NACIONAL - SNIES 2828

Bibliografía

- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media.
- Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Springer.
- Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.
- Müller, A. C., & Frank, E. (2016). Getting started with WEKA. University of Waikato.
- Müller, A. C., & Guido, S. (2017). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media.
- Raschka, S., & Mirjalili, V. (2019). Python machine learning (3rd ed.). Packt Publishing.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann.

📍 Sede Quirinal: Calle 21 No. 6 - 01

📍 Sede Prado Alto: Calle 8 No. 32 - 49 PBX: (608) 8754220

📍 Sede Pitalito: Carrera 2 No. 1 - 27 - PBX: (608) 8360699

✉ Email: contacto@corhuila.edu.co - www.corhuila.edu.co

Personería Jurídica Res. Ministerio de Educación No. 21000 de Diciembre 22 de 1989
NIT. 800.107584-2



CORPORACIÓN UNIVERSITARIA DEL HUILA - CORHUILA

"Diseño y prestación de servicios de docencia, investigación y extensión de programas de pregrado, aplicando todos los requisitos de las normas ISO implementadas en sus sedes Neiva y Pitalito"