Pontificia Universidad Católica del Ecuador
Seréis mis testigos

Preprocesamiento de datos

Mgtr. Sebastián Tamayo

> **Un viaje de mil millas comienza con el primer paso**
>
> *Lao Tse*

CONTENIDOS

1. Introducción
2. Técnicas de preprocesamiento
3. Detección de outliers
4. Análisis de Componentes Principales
5. Instalación de herramientas
6. Actividad Práctica

Pontificia Universidad Católica del Ecuador
Seréis mis testigos

01

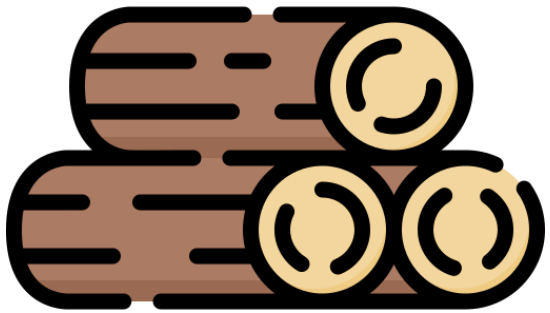
Introducción

¿En qué consiste esta fase?

- El preprocesamiento de datos es una etapa crucial en el análisis de datos y la minería de datos.

- Se preparan datos brutos para su posterior análisis.





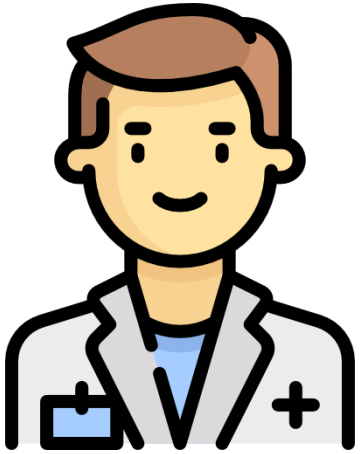

¿Qué implica?

- Esta fase implica una serie de técnicas y transformaciones que aseguran que los datos sean precisos, consistentes y adecuados para los modelos analíticos que se aplicarán.

- Sin un adecuado preprocesamiento, **los resultados de los análisis pueden ser engañosos o inexactos**.

¿Qué ocurre si se falla?

- ¿Qué ocurriría si en vez de leñador para talar arboles asignamos a un doctor?

Pontificia Universidad Católica del Ecuador
Seréis mis testigos

02

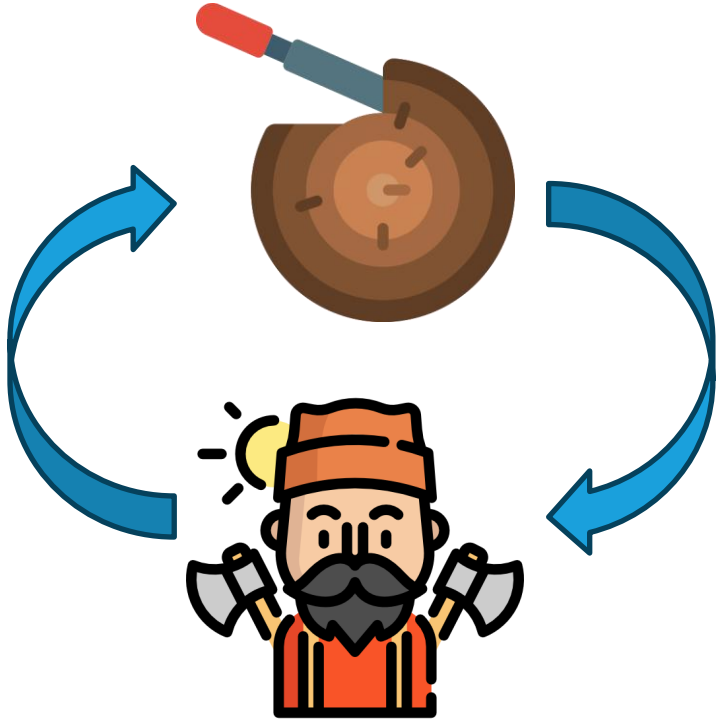
Técnicas de preprocesamiento

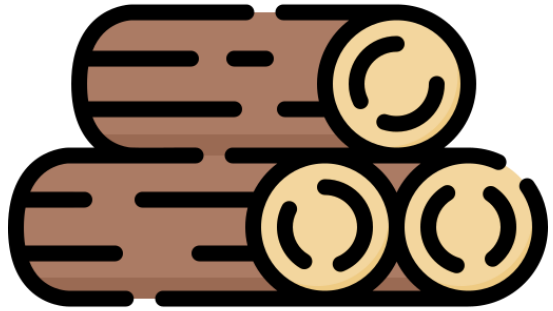
Limpieza de Datos

- Es uno de los primeros pasos en el preprocesamiento de datos.
- Consiste en la identificación y corrección de errores o inconsistencias en los datos.
- Puede incluir la eliminación de datos duplicados, el manejo de valores perdidos y la corrección de errores tipográficos.

Limpieza de Datos

- Entonces, el leñador no solo debe talar los árboles, debe hacerlo de tal forma que **estén listos para su procesamiento**.



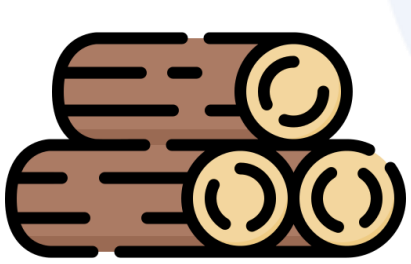


Reducción de Datos

- La reducción de datos busca **simplificar el conjunto de datos original**.
- Se realiza mediante la eliminación de variables irrelevantes o redundantes, lo que puede mejorar la eficiencia del análisis.
- Es crucial cuando se trabaja con conjuntos de datos grandes que tienen muchas variables.

Reducción de Datos

- Entonces, el leñador no solo debe talar los árboles, debe hacerlo de tal forma que **estén listos para su procesamiento**.



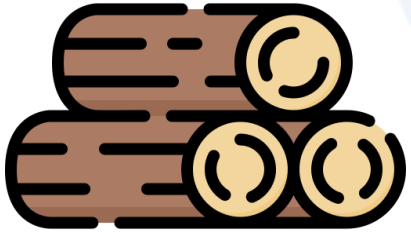
Enriquecimiento de Datos

- El enriquecimiento de datos implica agregar información adicional que puede mejorar la calidad y el valor de los datos existentes.

- Esta etapa es particularmente útil para crear un contexto más completo que permita un análisis más profundo.

Enriquecimiento de Datos

¿Cómo se modificaría el ejemplo del leñador?

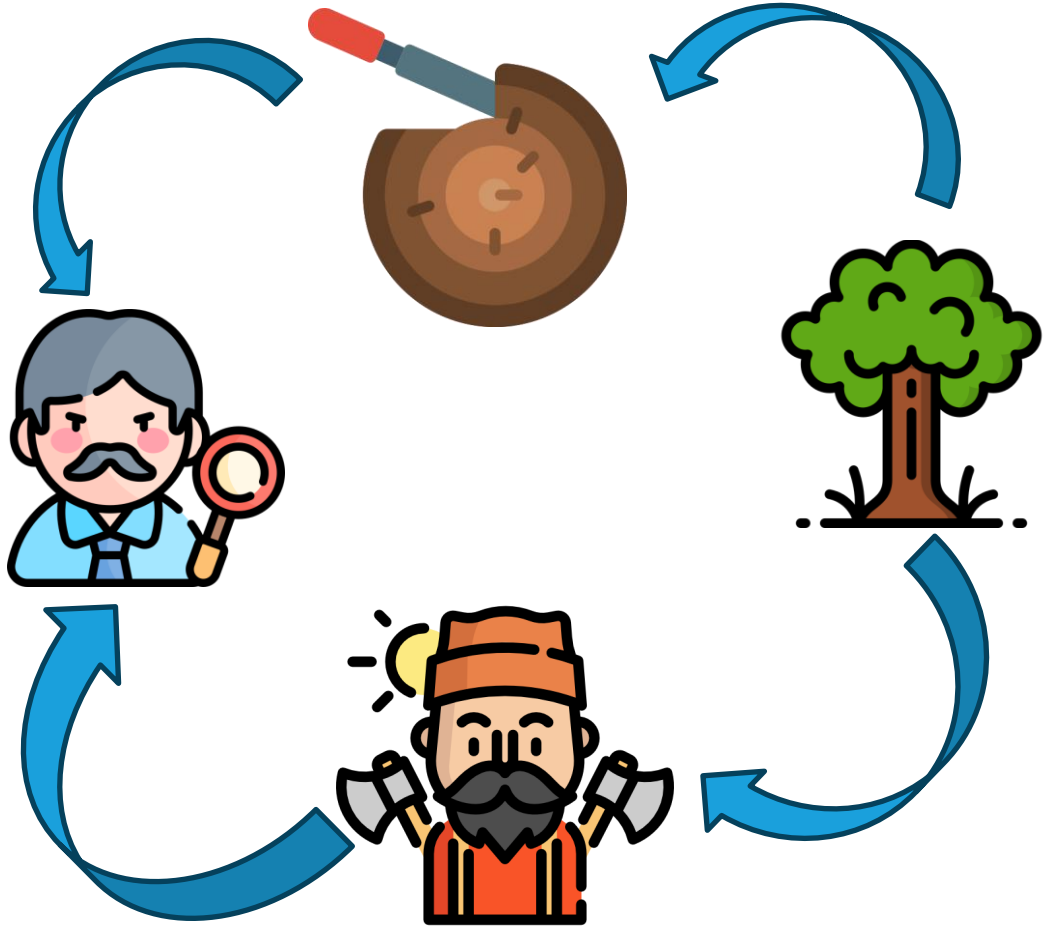


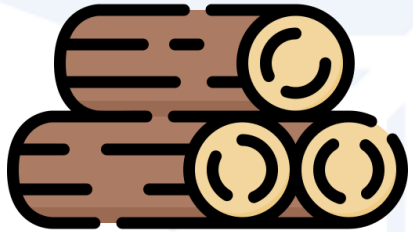
Validación de Datos

- La validación de datos es la etapa final del preprocesamiento y se refiere a la verificación de que los datos cumplen con criterios específicos de calidad y precisión.

- Esta etapa es esencial para garantizar que los datos sean confiables antes de ser utilizados en análisis posteriores.

Validación de Datos





Importancia del preprocesamiento de datos

Mejora la Calidad de los Datos

- Se logra mediante la limpieza de datos.
- Implica identificar y corregir errores.

Aumento de la Eficiencia Analítica

- Reduce el volumen de datos mediante la eliminación de atributos irrelevantes o redundantes.

Facilita el Machine Learning

- Muchos algoritmos requieren que los datos estén en un formato específico y que sean normalizados o estandarizados.

Mejora de la Interpretabilidad

- Al transformar los datos en un formato más comprensible, los analistas pueden comunicar sus hallazgos de manera efectiva.

Validación y Confiabilidad

- Asegura que los datos sean válidos y confiables.

Pontificia Universidad Católica del Ecuador
Seréis mis testigos

Detección de outliers

03

¿Por qué es importante?

- La detección de outliers es un componente crítico en el análisis de datos.
- Los valores atípicos pueden **influir de manera significativa en la validez y precisión de los resultados** obtenidos a partir de un conjunto de datos.
- Pueden distorsionar las medidas estadísticas, como la media y la desviación estándar, lo que puede llevar a conclusiones erróneas.

¿Qué son?

- Un outlier se define como un dato que **se encuentra considerablemente alejado de otros valores** en un conjunto.

- Esta anomalía puede deberse a variaciones naturales en el fenómeno que se está estudiando, errores en la recolección de datos o condiciones extraordinarias.

Métodos de Detección

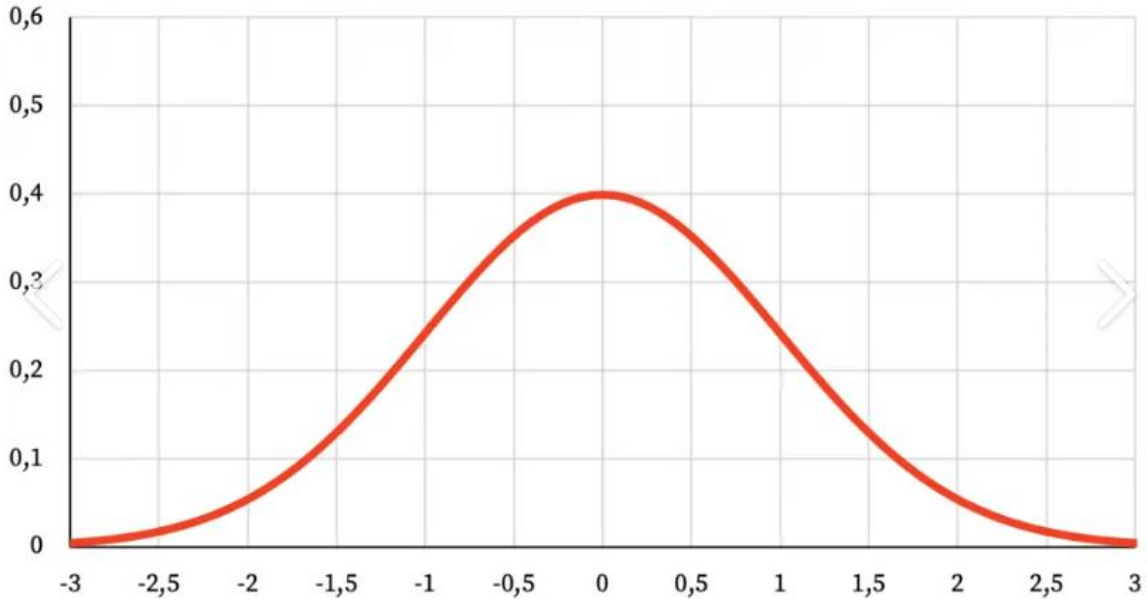
Métodos Estadísticos

- Un método común es el uso de la desviación estándar.
- Un valor se considera un outlier si se encuentra a más de tres desviaciones estándar de la media.
- Este método es efectivo para conjuntos de datos que siguen una distribución normal, pero puede no ser adecuado para distribuciones asimétricas.

Métodos por Machine Learning

- Los algoritmos de clustering, como DBSCAN, son útiles para identificar outliers al agrupar datos y detectar puntos que no pertenecen a ningún grupo.
- Esta técnica es particularmente valiosa en conjuntos de datos de alta dimensión donde la visualización es complicada.

Métodos de Detección



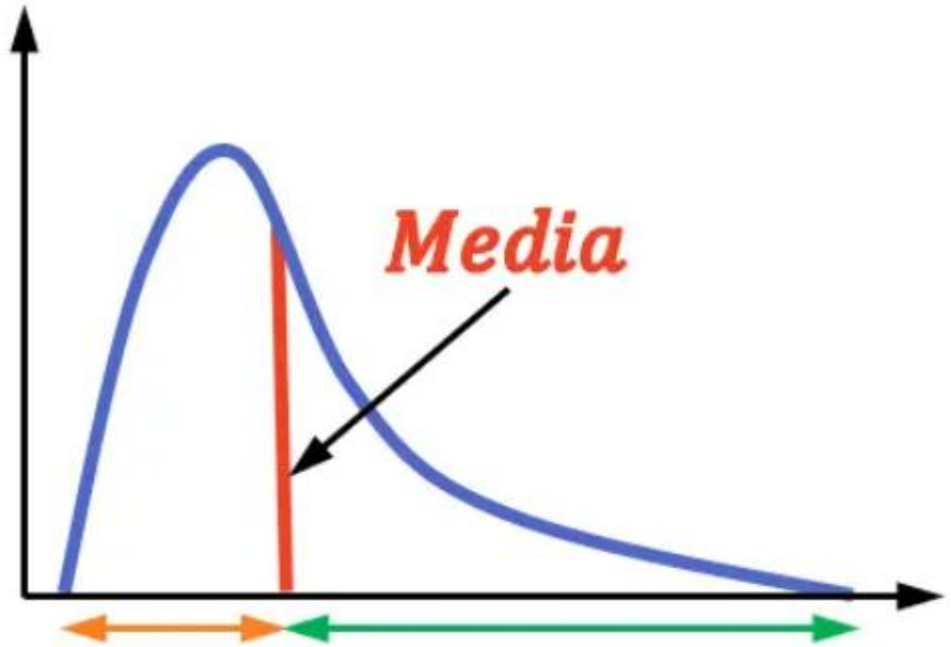
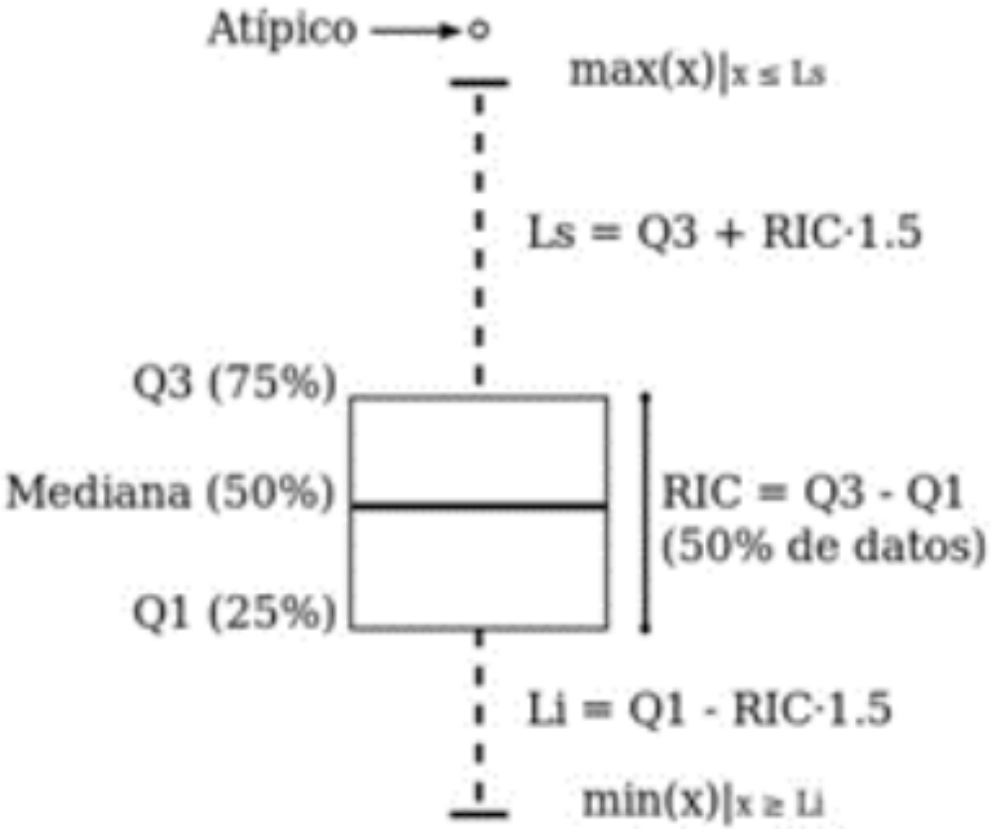


Diagrama de Caja y Bigotes

- El boxplot es una herramienta gráfica fundamental en la estadística que se utiliza para visualizar la distribución de un conjunto de datos.

- Su diseño permite representar la mediana, cuartiles y los valores atípicos, facilitando así la comparación de diferentes conjuntos de datos.

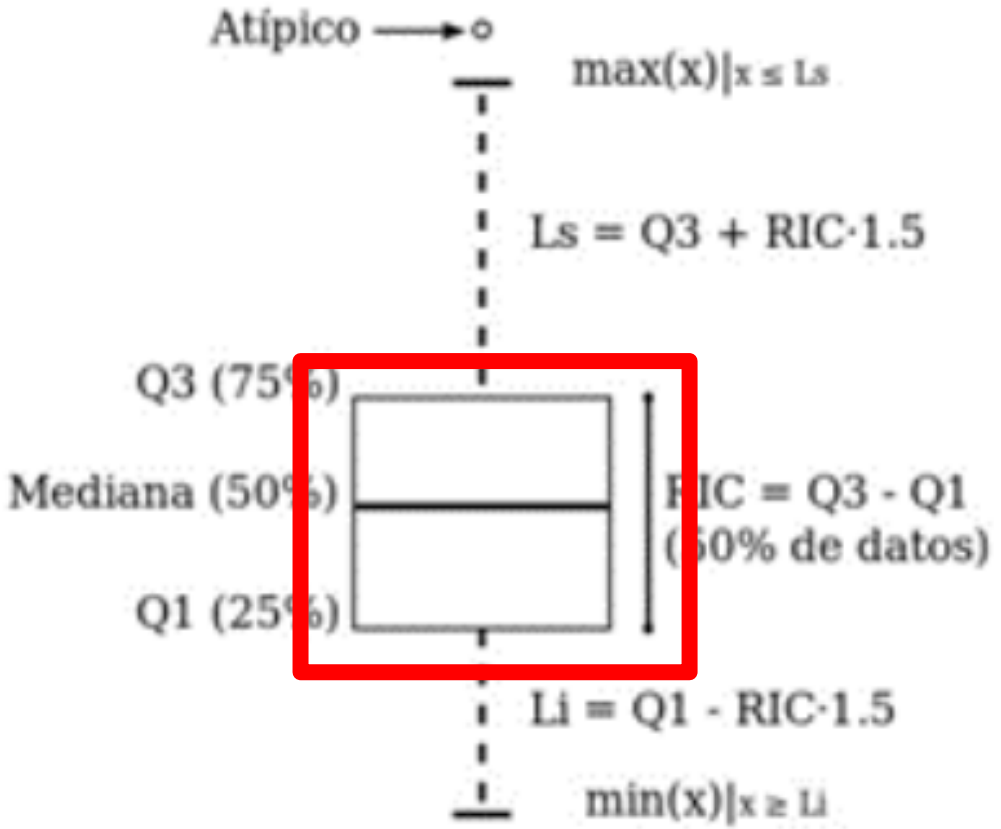
Diagrama de Caja y Bigotes



Componentes del Boxplot

La **caja** representa el rango intercuartílico, que es la distancia entre el primer cuartil y el tercer cuartil.

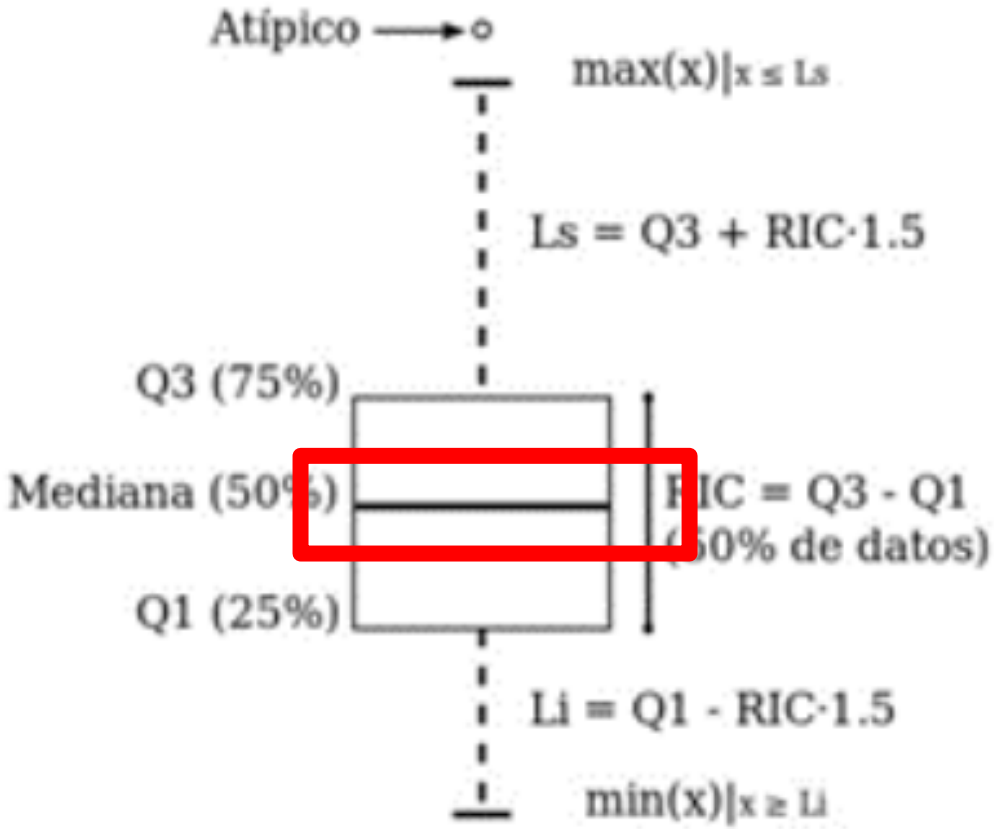
Su longitud indica la variabilidad de la mitad central del conjunto de datos.



Componentes del Boxplot

La **línea mediana** proporciona una medida robusta de tendencia central que no se ve afectada por outliers.

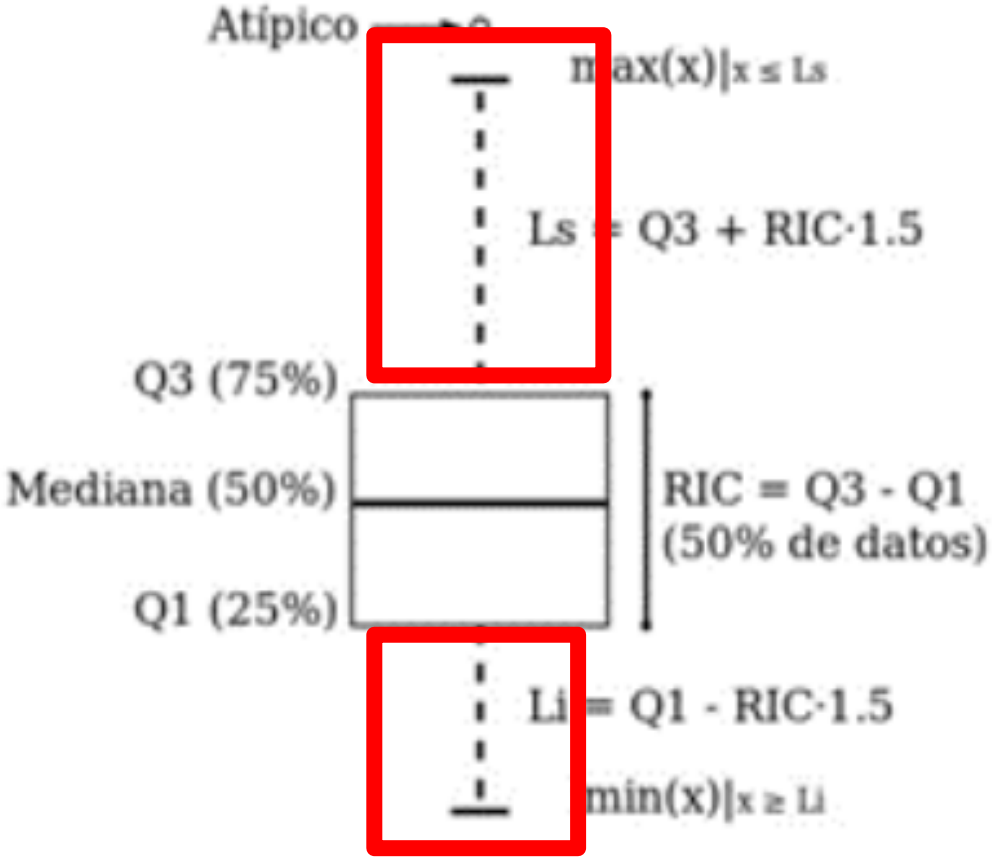
Se ubica dentro de la caja.



Componentes del Boxplot

Los **“bigotes”** son las líneas que se extienden desde la caja.

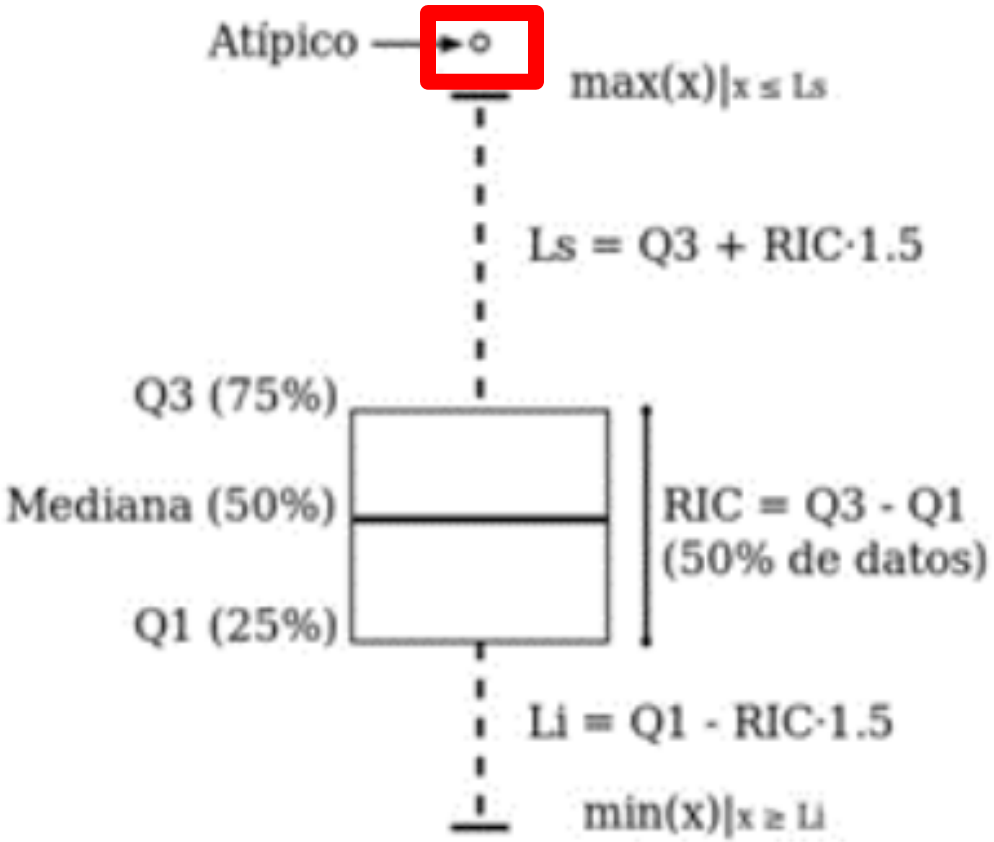
Muestran el rango de los datos y suelen extenderse hasta 1.5 veces el IQR desde los cuartiles, lo que ayuda a identificar valores atípicos.



Componentes del Boxplot

Los **valores atípicos** son los puntos que caen fuera del rango de los bigotes.

Se representan como puntos individuales, lo que permite identificar rápidamente datos que requieren atención especial.



Ventajas y Limitaciones del Boxplot

Ventajas

- Tiene gran capacidad de resumir un gran volumen de datos en una representación gráfica simple.
- No solo muestran la forma de la distribución, sino que también destacan valores atípicos y la variabilidad dentro de los datos.
- Su diseño permite comparaciones rápidas entre múltiples conjuntos de datos.

Limitaciones

- No muestran la distribución completa de los datos, como la forma exacta de la distribución o la presencia de agrupamientos.
- Pueden ser menos informativos en conjuntos de datos muy pequeños o en distribuciones con sesgos extremos.

Pontificia Universidad Católica del Ecuador
Seréis mis testigos

Análisis de Componentes Principales

04

¿Qué es?

- El Análisis de Componentes Principales (PCA), es una técnica estadística multivariada ampliamente utilizada en el campo de la exploración y el análisis de datos.

- Su principal objetivo es **reducir la dimensionalidad** de un conjunto de datos mientras se conserva la mayor cantidad de variabilidad posible.

¿Cuál es su objetivo?

- El PCA busca **transformar** un conjunto de variables **correlacionadas** en un nuevo conjunto de variables **no correlacionadas**.

- Estas nuevas variables se llaman **componentes principales**.

Proceso del PCA

Normalización de los Datos
- Se deben normalizar los datos.
- Se puede optar por la estandarización.

Cálculo de la Matriz de Covarianza
- Se calcula para identificar cómo varían conjuntamente las variables.
- Permite entender la relación entre las variables.

Cálculo de los Valores y Vectores Propios
- Se obtienen valores propios de la matriz de covarianza.
- Estos valores indican la cantidad de varianza que se captura.
- Los vectores representan la dirección de los componentes.

Selección de Componentes Principales
- Se seleccionan aquellos con los valores más altos.
- Estos son los que retienen la mayor parte de la variabilidad.

Transformación de los Datos
- Se proyectan los datos originales en el espacio de los componentes
- Resulta un nuevo conjunto de datos con menos dimensiones.

Ventajas y Limitaciones del PCA

Ventajas

- Simplifica el análisis de datos.
- Reduce el ruido.
- Mejora en la visualización de datos de alta dimensión.

Limitaciones

- Es difícil interpretar los componentes principales, ya que son combinaciones lineales de las variables originales.
- El PCA assume que las relaciones entre las variables son lineales, lo que puede no ser cierto en todos los casos.

Puntos clave

- Las componentes principales son variables que no están relacionadas entre sí y que explican el comportamiento de los datos acumulando cierto porcentaje de variabilidad.

- Los loadings (o cargas) de estas componentes ayudan a la interpretación de si mismas.

Puntos clave

- Si la componente principal tiene todas sus cargas positivas, esto indica que es una componente de tamaño.

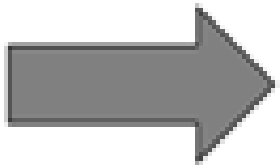
  Mientras más alta sea esta componente, más altos serán los valores de la variable.

- Si las cargas son positivas y negativas, esto convierte a la componente en una componente de contraste.

Mientras más alta sea esta componente, mayores serán los contrastes entre las variables con cargas positivas y negativas.

PCA

id	chol	presa	ldlc	presb	insu	gluc	hdlc	trig
1	100	120	100	48	20	142	50	150
2	220	125	103	45	23	140	51	154
3	130	112	103	42	25	139	55	142
4	150	122	108	40	21	138	52	148



	CP1	CP2	CP3
chol	0,42	-0,11	0,15
presa	0,4	0,12	0,08
ldlc	0,4	-0,16	0,04
presb	0,36	0,07	0,35
insu	0,35	0,03	-0,46
gluc	0,32	0,12	-0,67
hdlc	0,29	-0,64	0,22
trig	0,23	0,7	0,32

Pontificia Universidad Católica del Ecuador
Seréis mis testigos

Instalación de herramientas

05

VSCode

- La herramienta que aprenderán en esta ocasión se llama Visual Studio Code, es gratuita y les permitirá interactuar con los laboratorios y secciones prácticas de la asignatura.

- Pueden descargar la herramienta en el siguiente enlace:

https://code.visualstudio.com/download

Python

- Es muy posible que ustedes ya se hayan instalado Python, sin embargo, para quienes aun no lo han hecho aun pueden hacerlo de dos formas:

- Windows, a través del Windows Store:

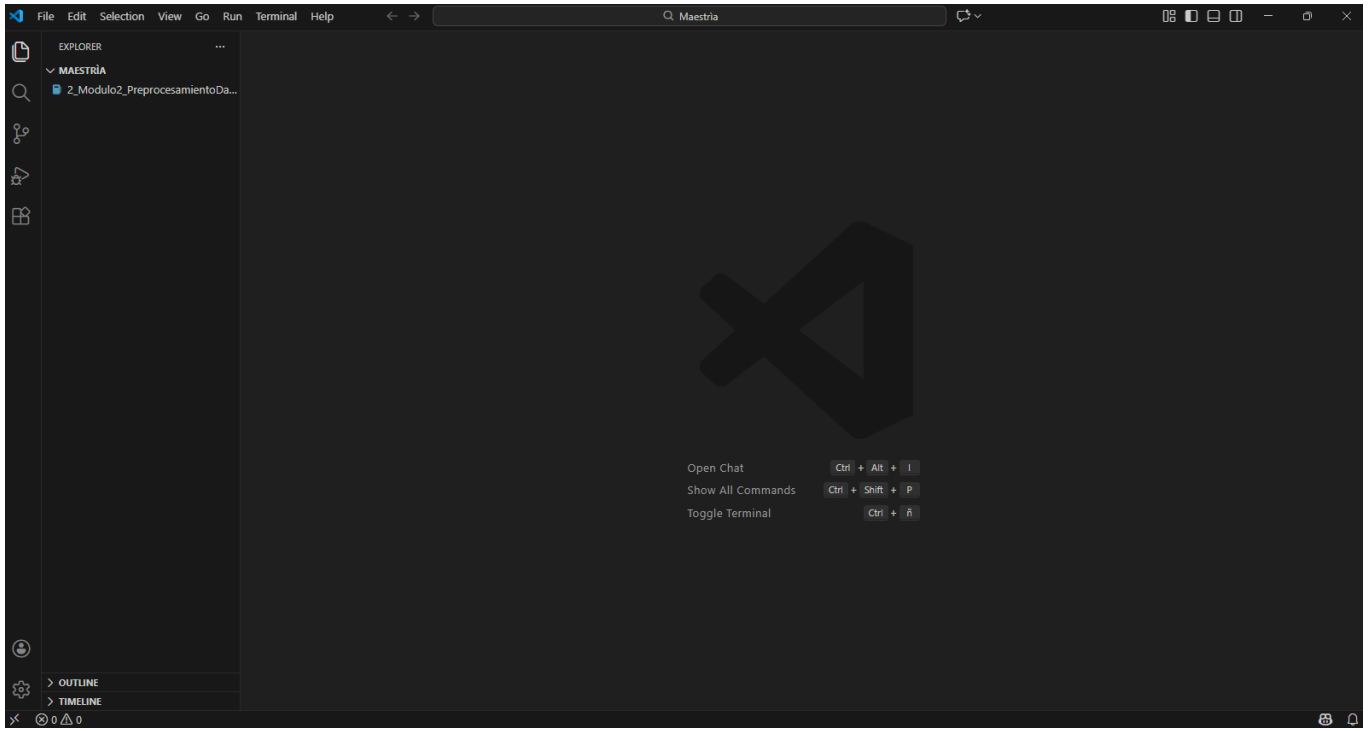
https://apps.microsoft.com/detail/9PNRBTZXMB4Z?hl=en-us&gl=EC&ocid=pdpshare

- Sistema operativo a elección:

https://www.python.org/downloads/

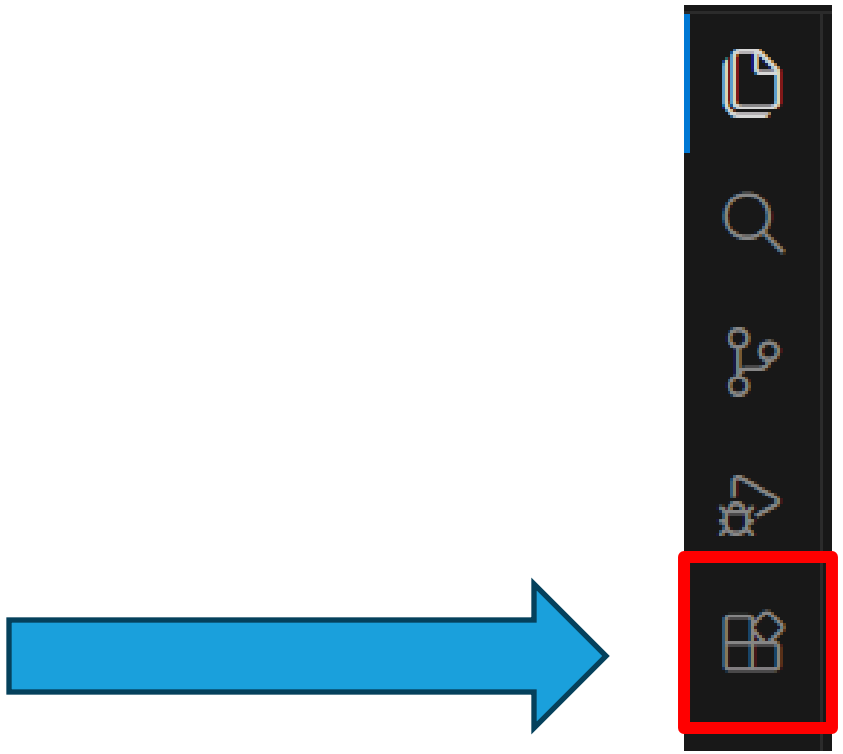
En VSCode

- Habiendo instalado ambas herramientas, ingresen a VSCode.



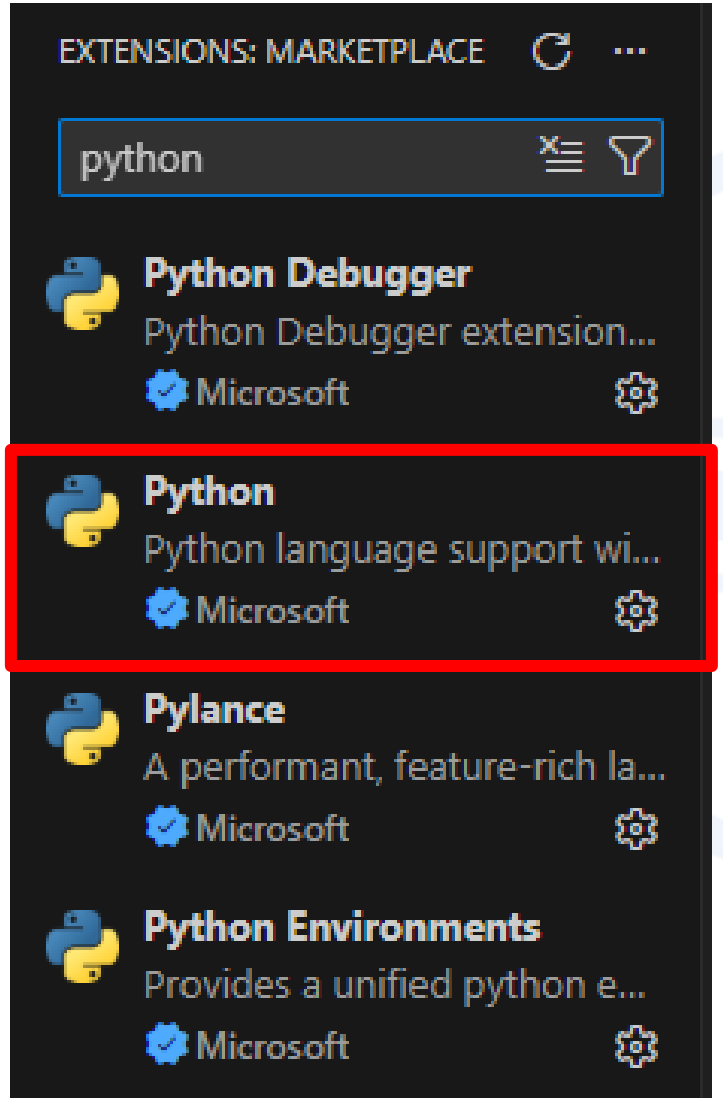
En VSCode

- Identifiquen la siguiente opción en el menú lateral izquierdo.



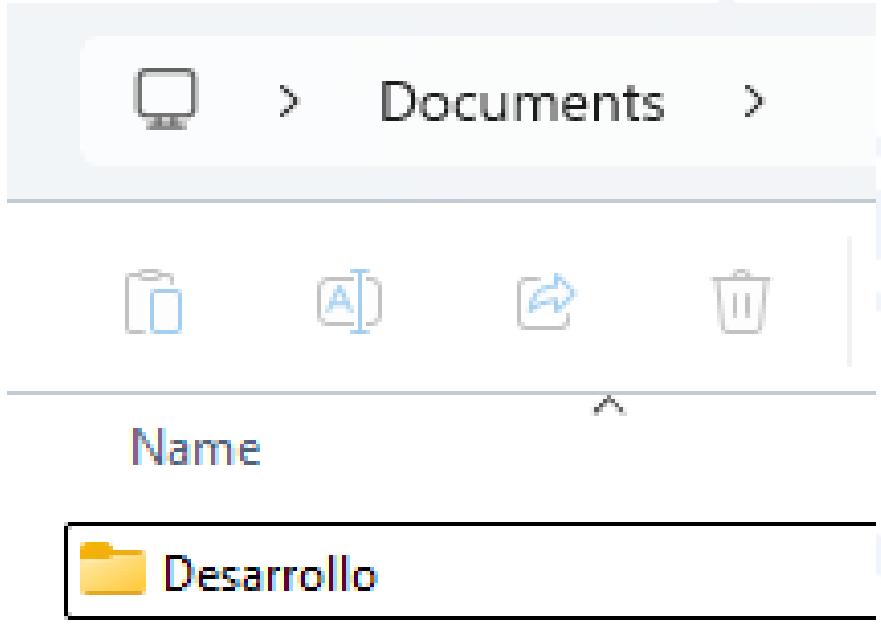
En VSCode

- En el buscador, coloquen “Python”, y den clic sobre instalar del que dice únicamente Python.
- Esto procederá con la instalación de las demás dependencias requeridas.



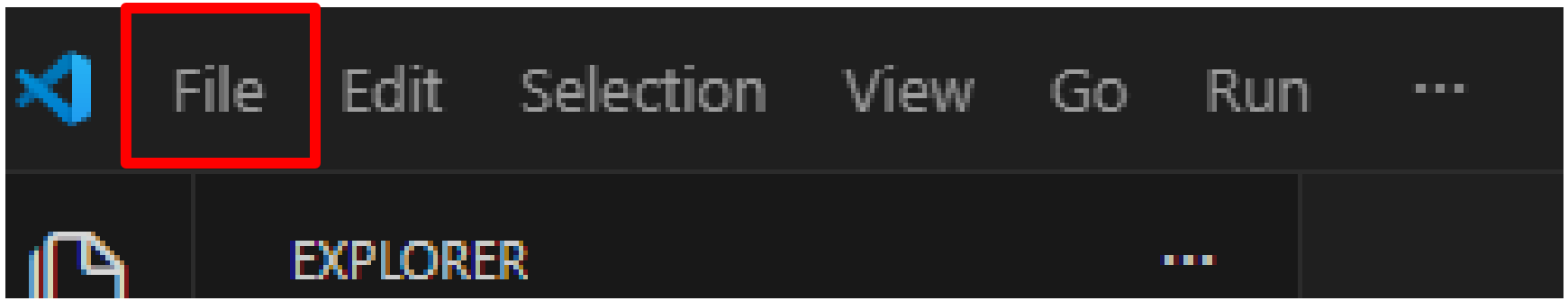
En su computadora

- Ingresen a alguna carpeta de su preferencia, en mi caso fue “Documentos”.
- Creen una carpeta llamada “Desarrollo” o del nombre de su preferencia. Para evitar conflictos, no coloquen espacios.



En VSCode

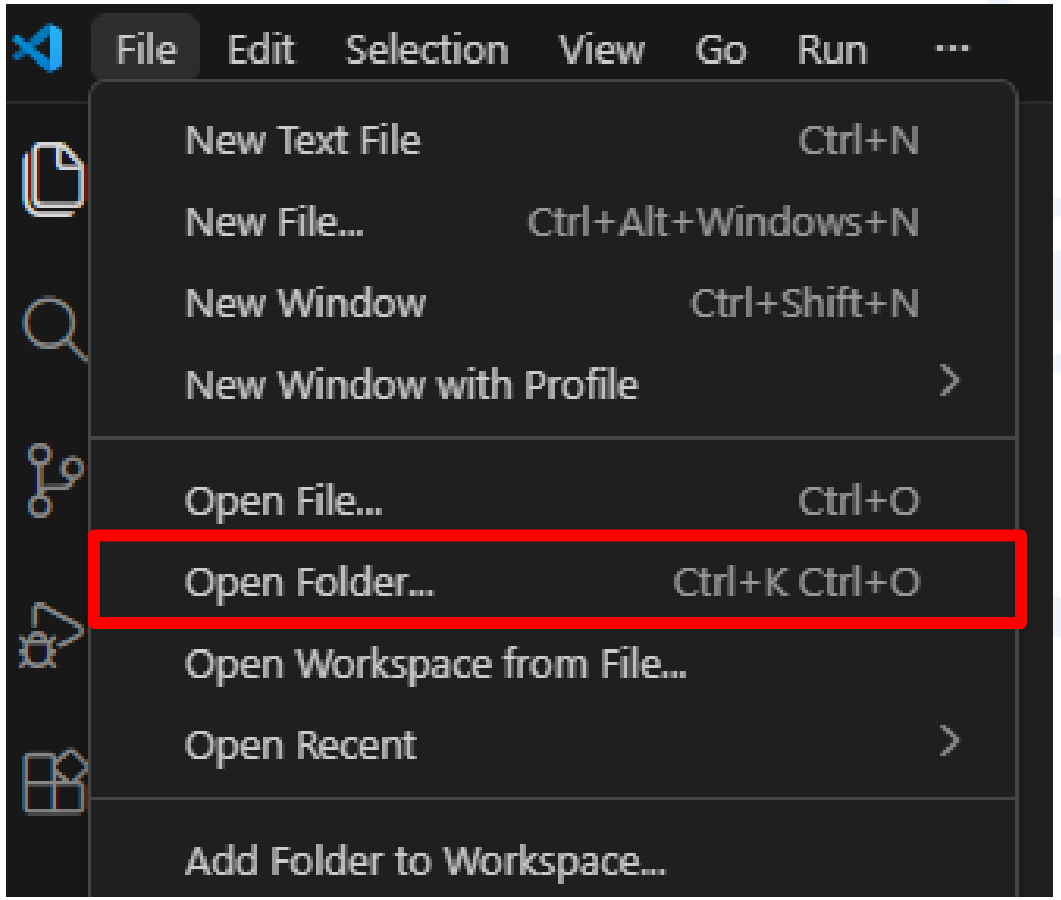
- Identifiquen la opción que se encuentra en la parte superior izquierda de la herramienta, y den clic.



En VSCode

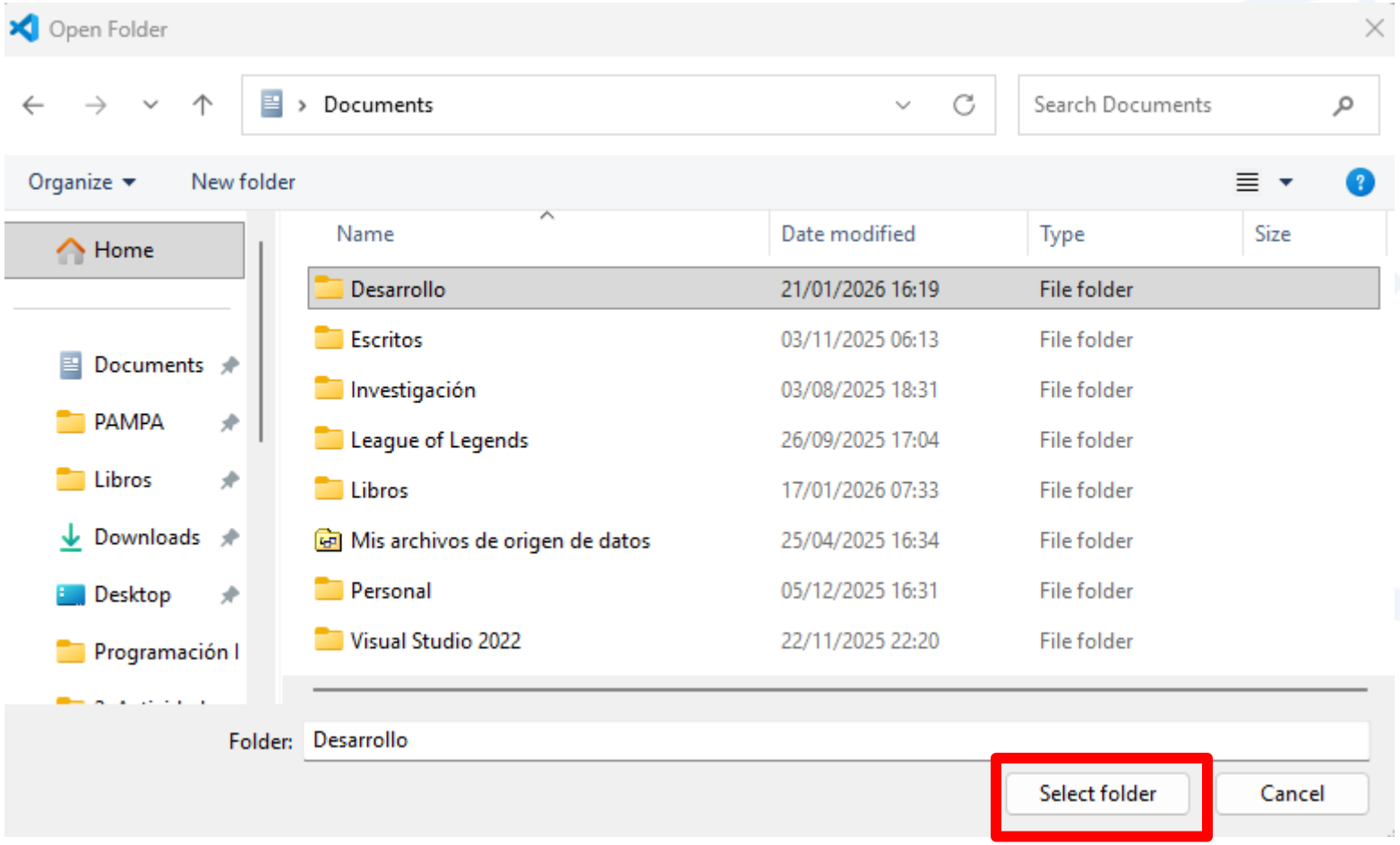
- Una vez dieron clic, se debió haber abierto un submenú, selección la opción de "**Open folder**" o "**Abrir carpeta**".

- Den clic.



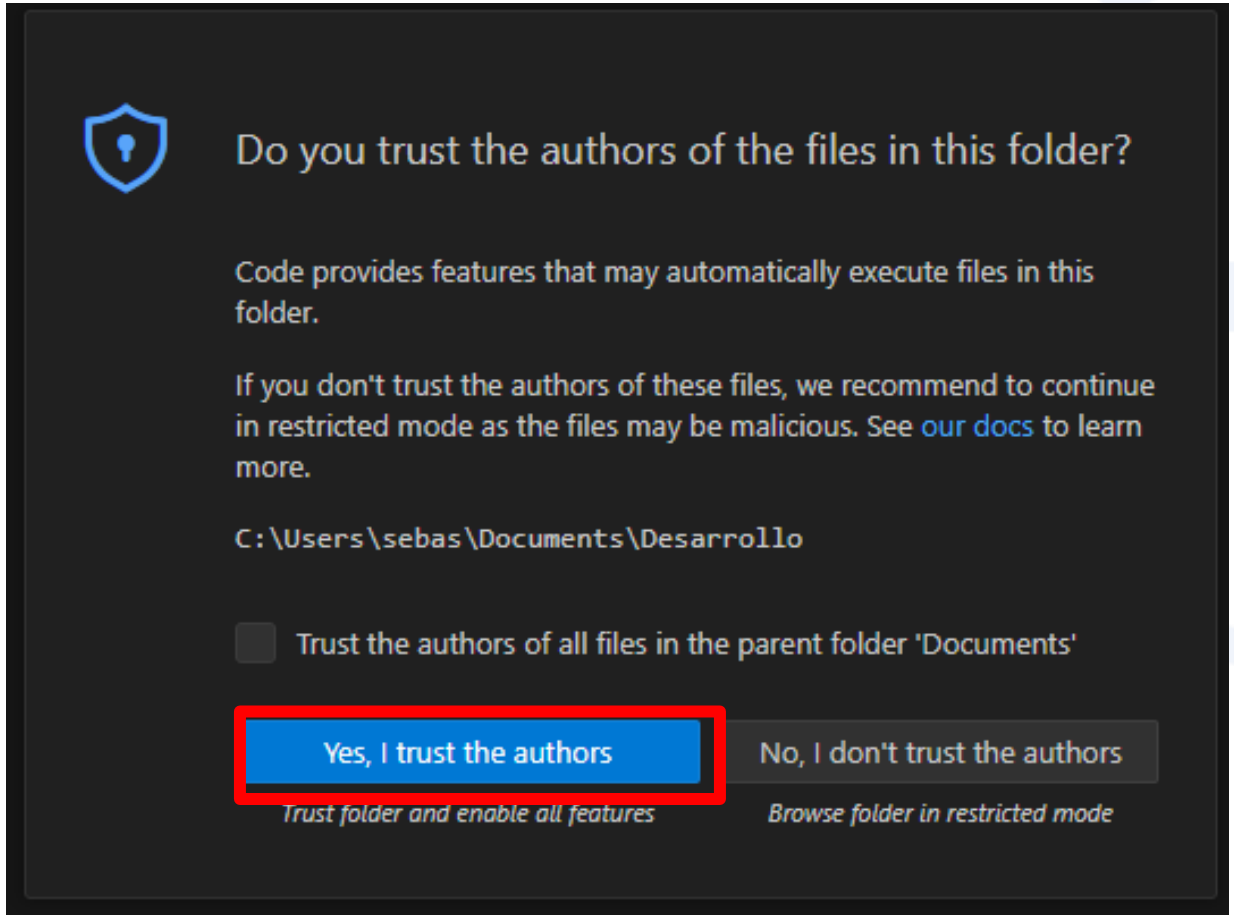
En VSCode

- Se debió haber abierto un explorador de archivos.
- Ubiquen la carpeta que crearon anteriormente y den clic en “Seleccionar carpeta”.



En VSCode

- Den clic en “Yes, I trust the authors” o “Sí, confío en los autores”.

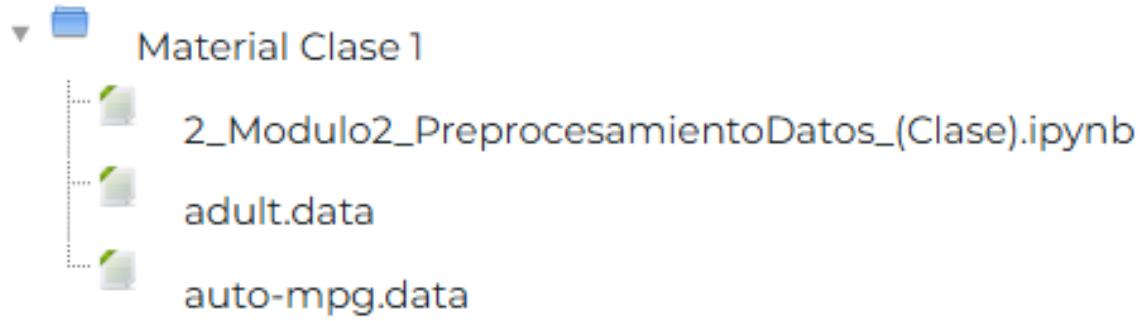


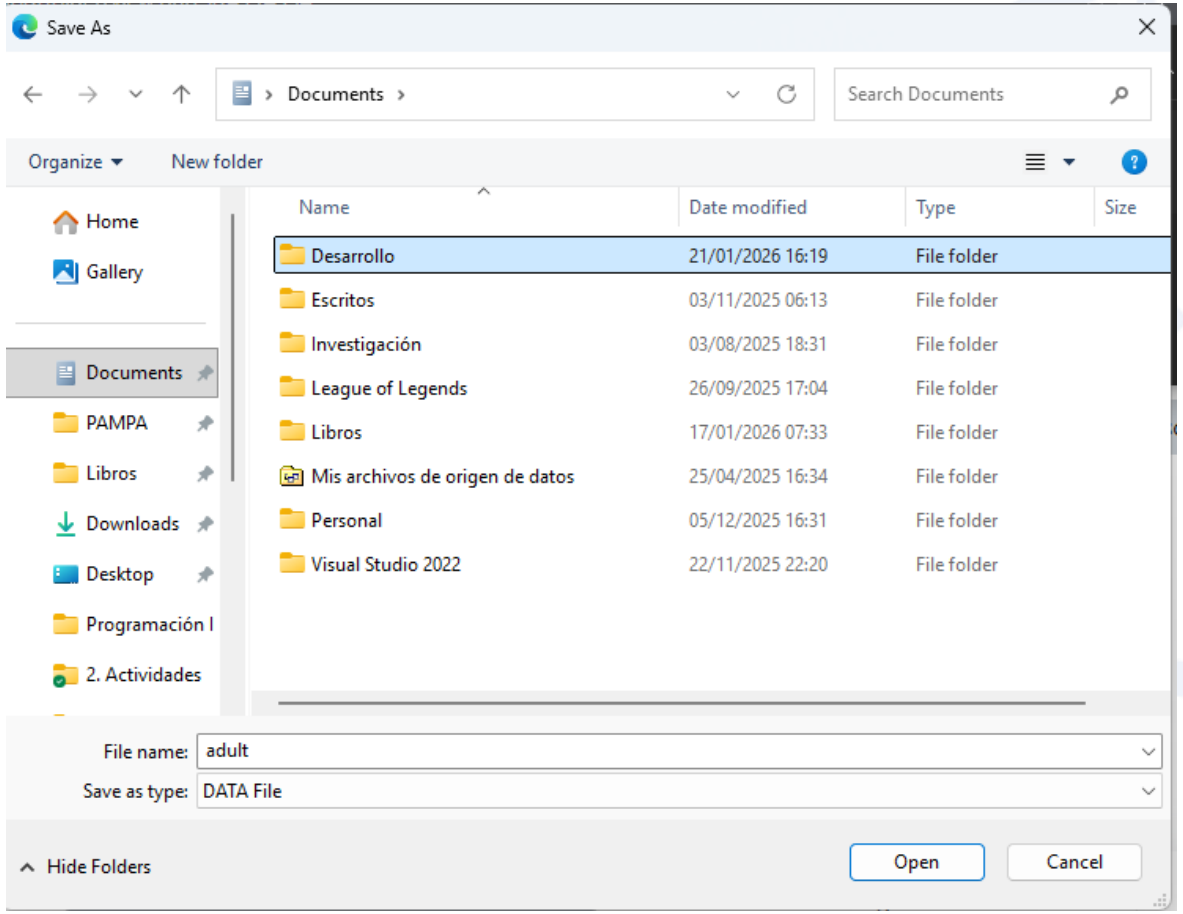
En su aula virtual

- Ingresen a los contenidos de la clase dos y ubiquen la carpeta “Material Clase 2”.

En su aula virtual

- Descarguen los archivos y guárdenlos en la carpeta creada anteriormente.

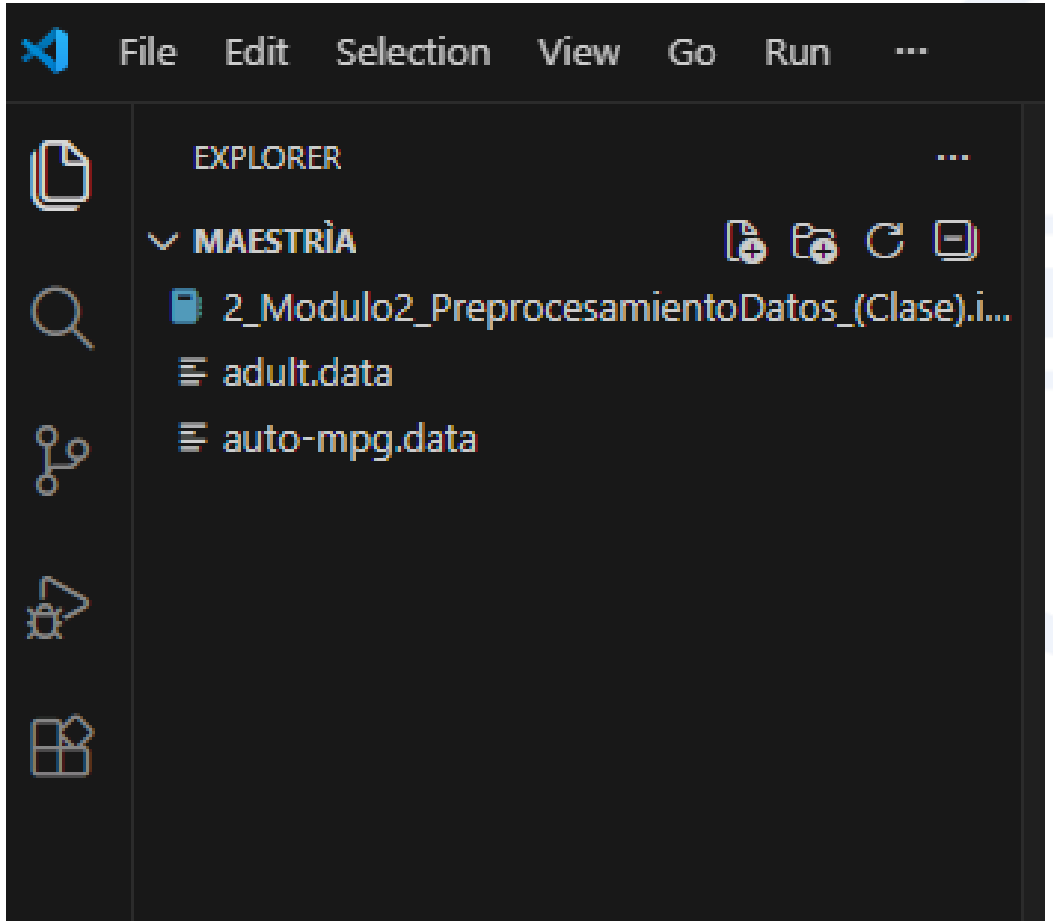




En VSCode

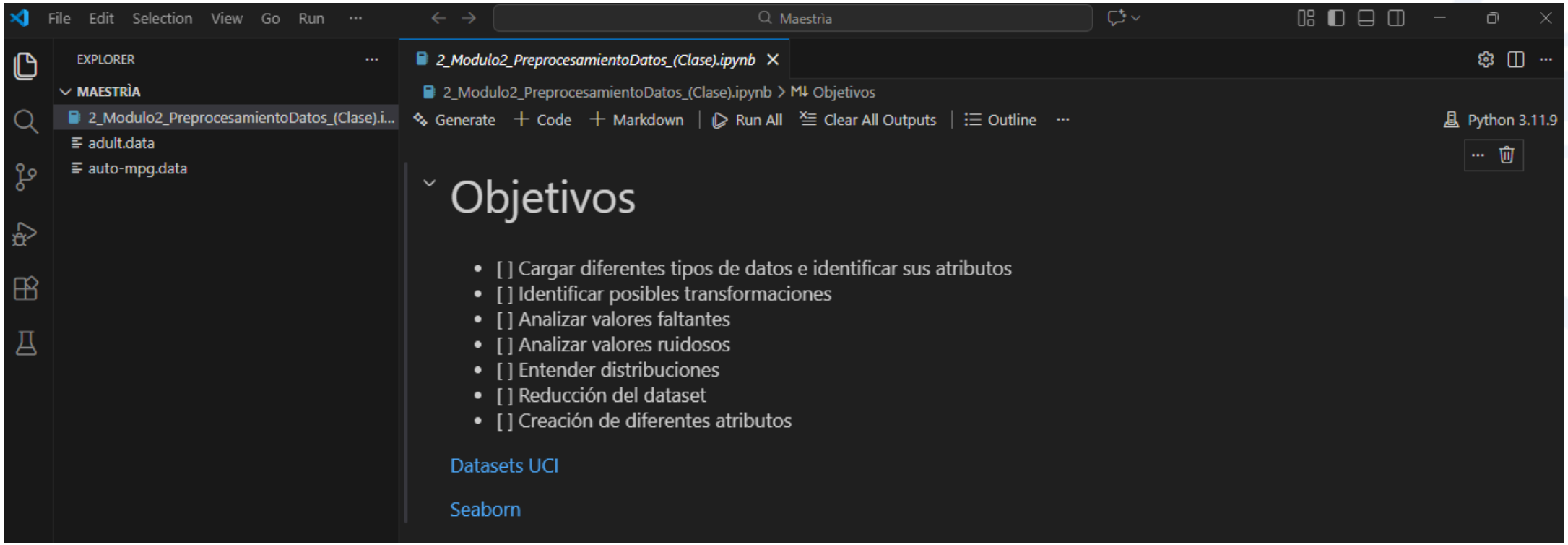
- Abran VSCode, y deberían encontrar los archivos descargados.

- Den clic sobre “2_Modulo2_Preprocesmaiento Datos_(Clase).ipynb



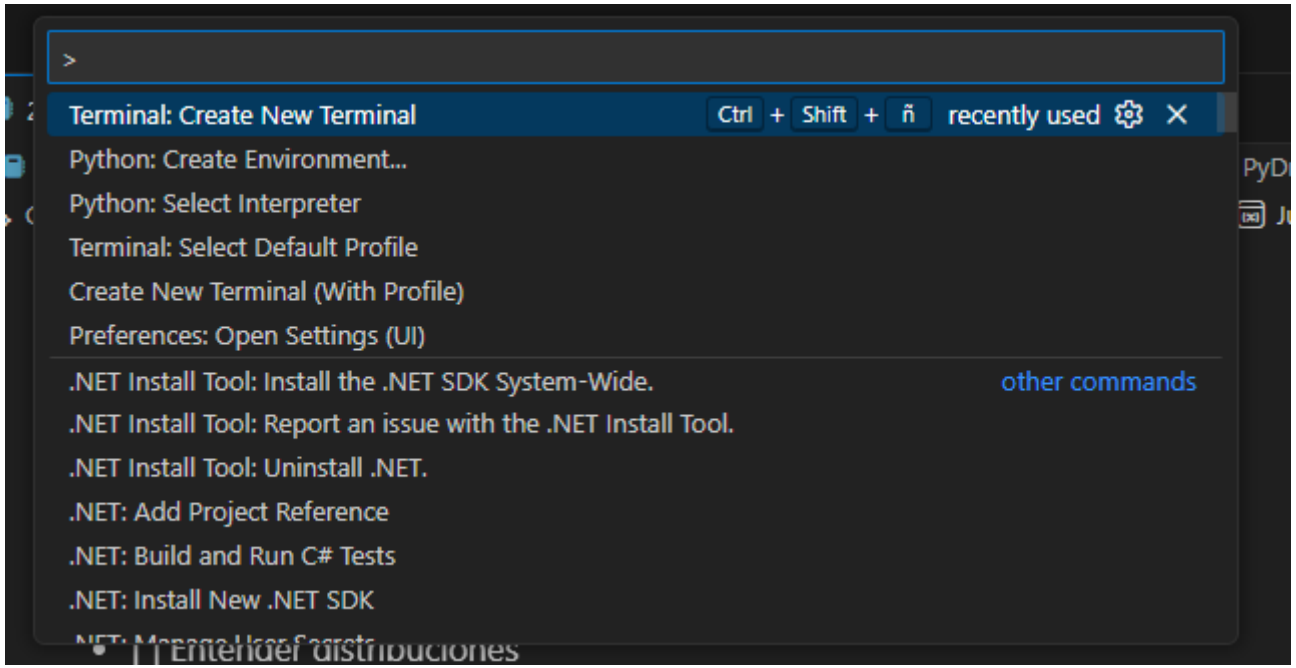
En VSCode

- Les debió haber aparecido lo siguiente:



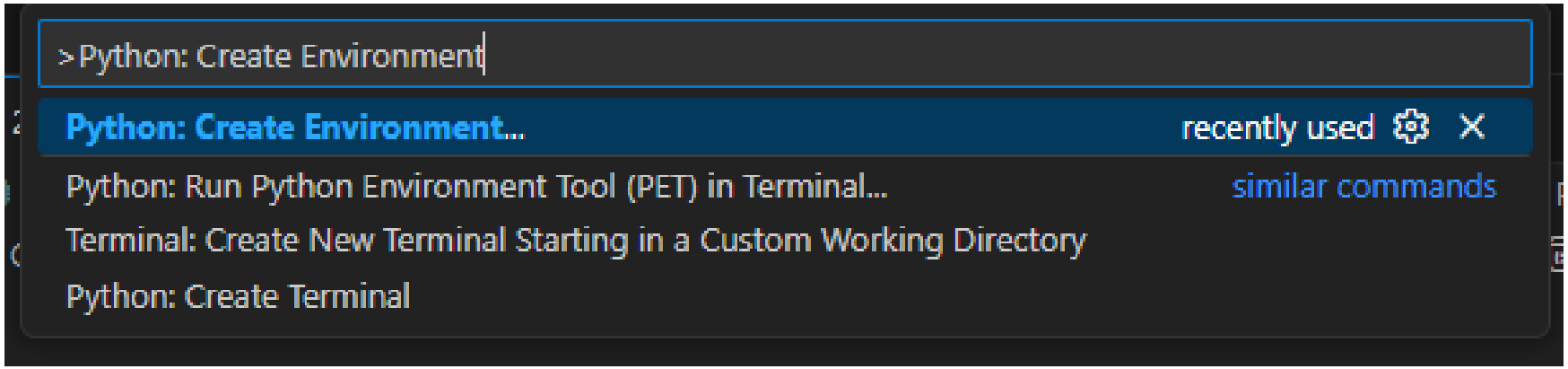
En VSCode

- Presionen las teclas **shift + control + p** al mismo tiempo, debería aparecer lo siguiente o similar.



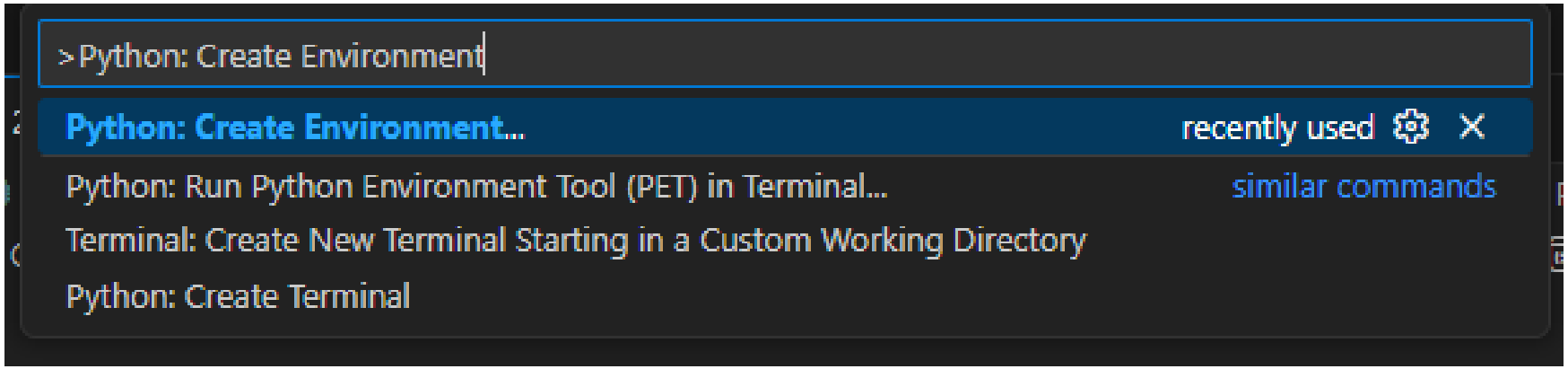
En VSCode

- En el buscar ingresen lo siguiente y seleccionen la opción subrayada.



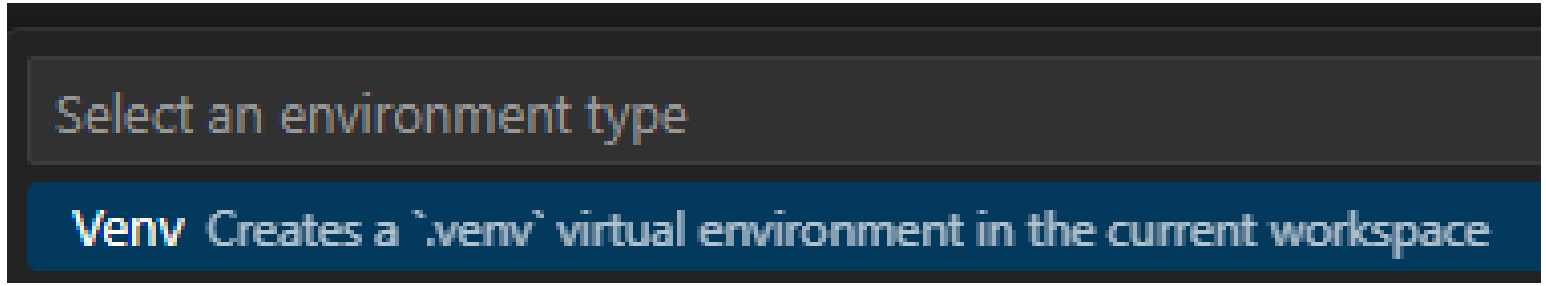
En VSCode

- En el buscar ingresen lo siguiente y seleccionen la opción subrayada.

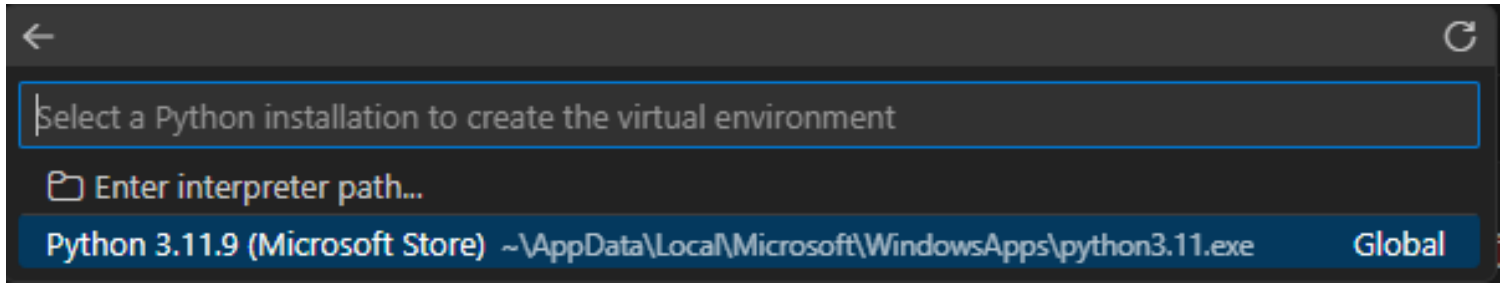


En VSCode

- Primero deberán seleccionar la siguiente opción:

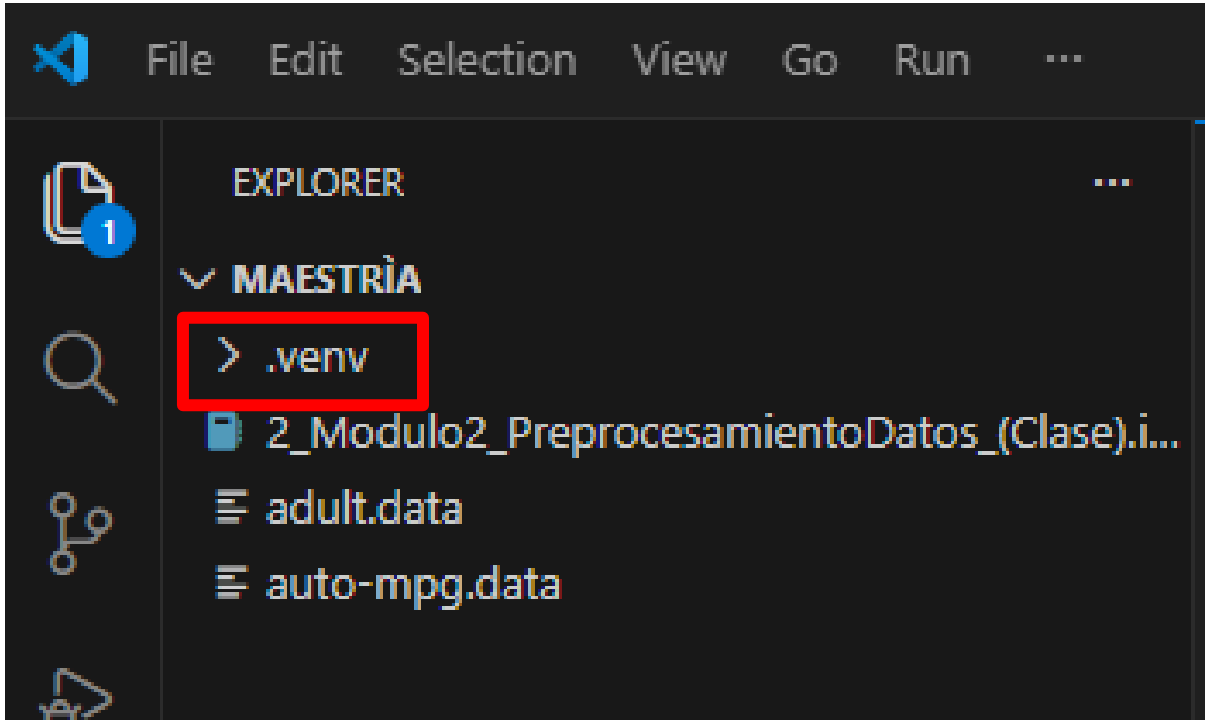


- Después deberán seleccionar lo siguiente:



En VSCode

- Esperen un momento y podrán identificar que en su VSCode apareció la siguiente carpeta.



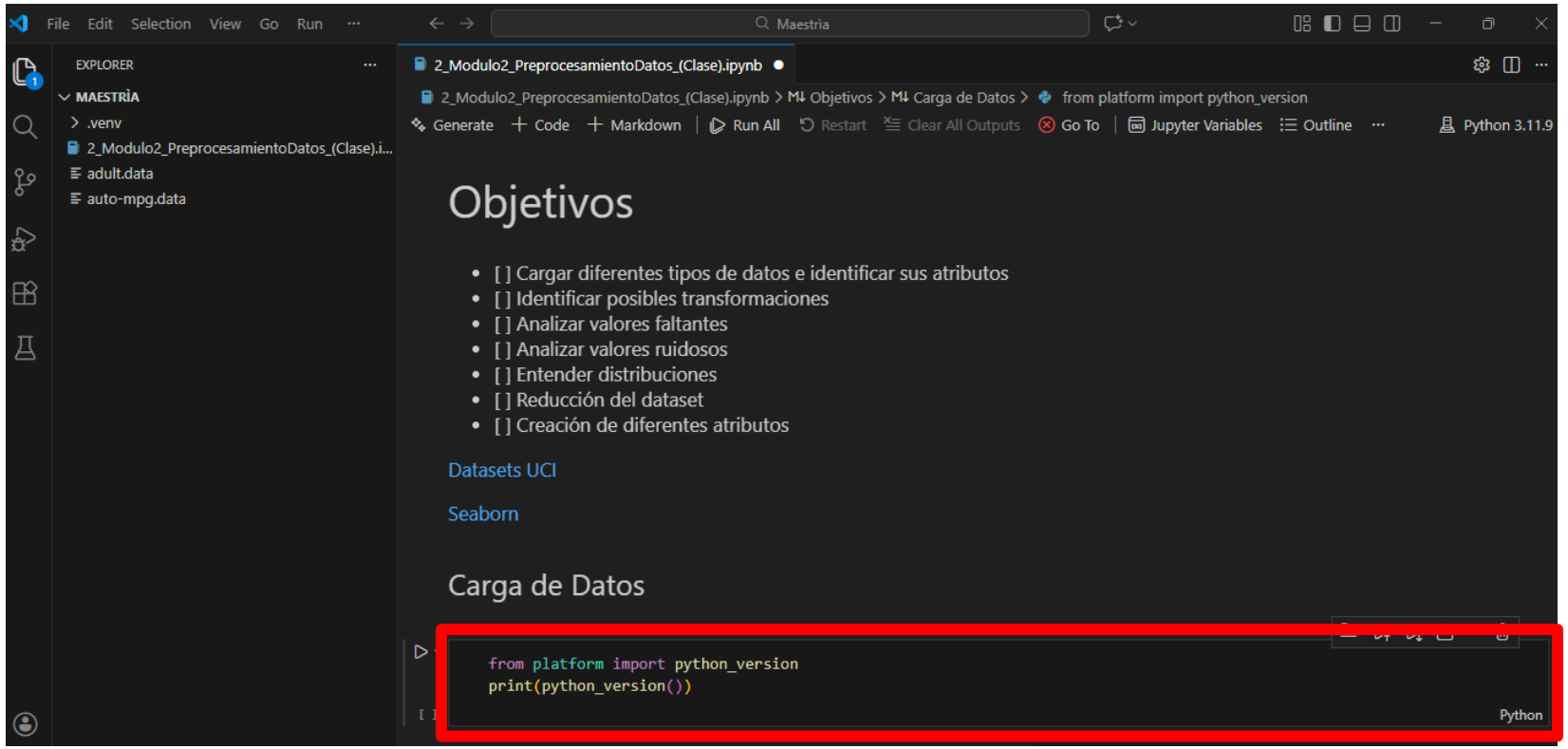
Virtual Environment

- Lo que acaban de crear se conoce como un virtual environment, y permitirá que todas las librerías que ustedes requieran instalar, se realice única y exclusivamente allí.

- Es decir, los paquetes no se instalarán en sus computadoras propiamente, sino en un entorno que funcionará exclusivamente para esta asignatura.

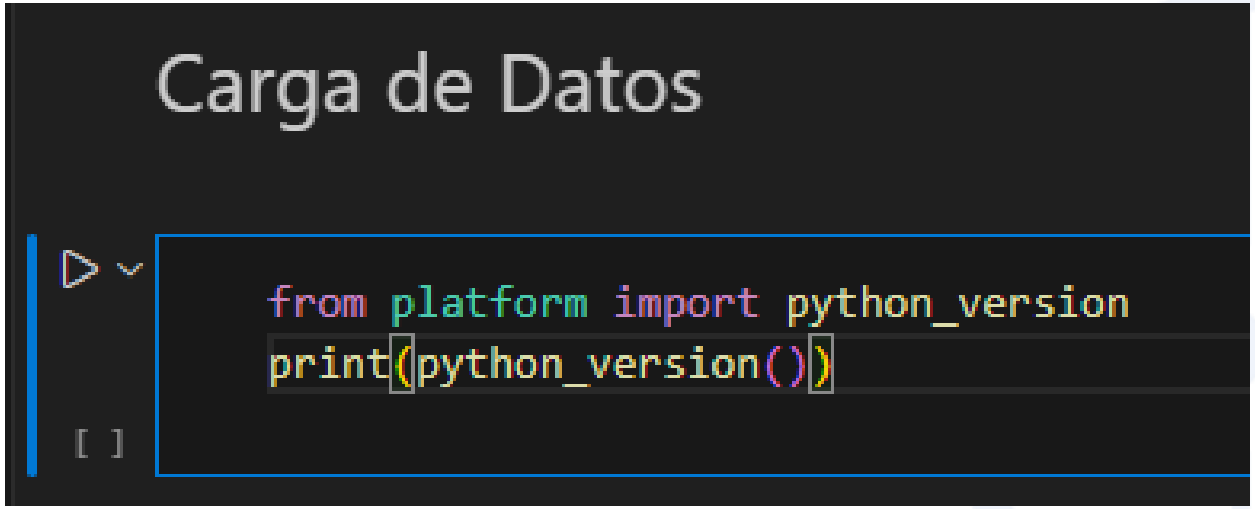
En el archivo abierto

- Ubiquen la celda seleccionada y denle clic.

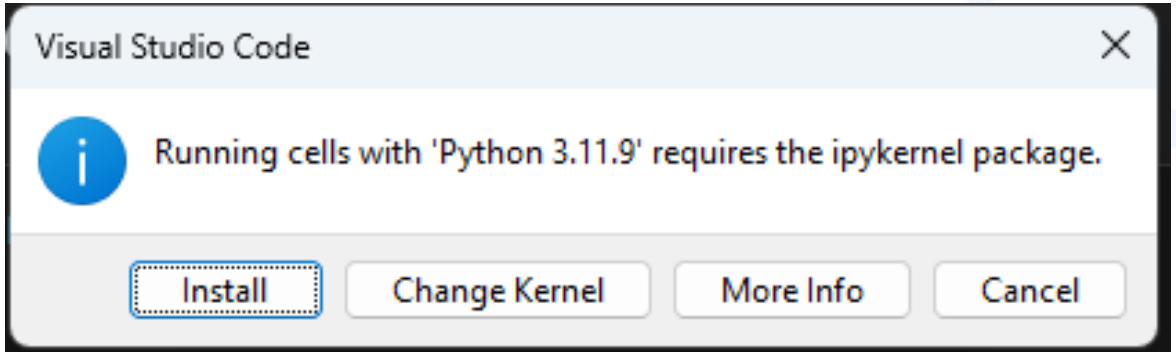


En el archivo abierto

1. Una vez allí, presionen las teclas **shift + enter**.

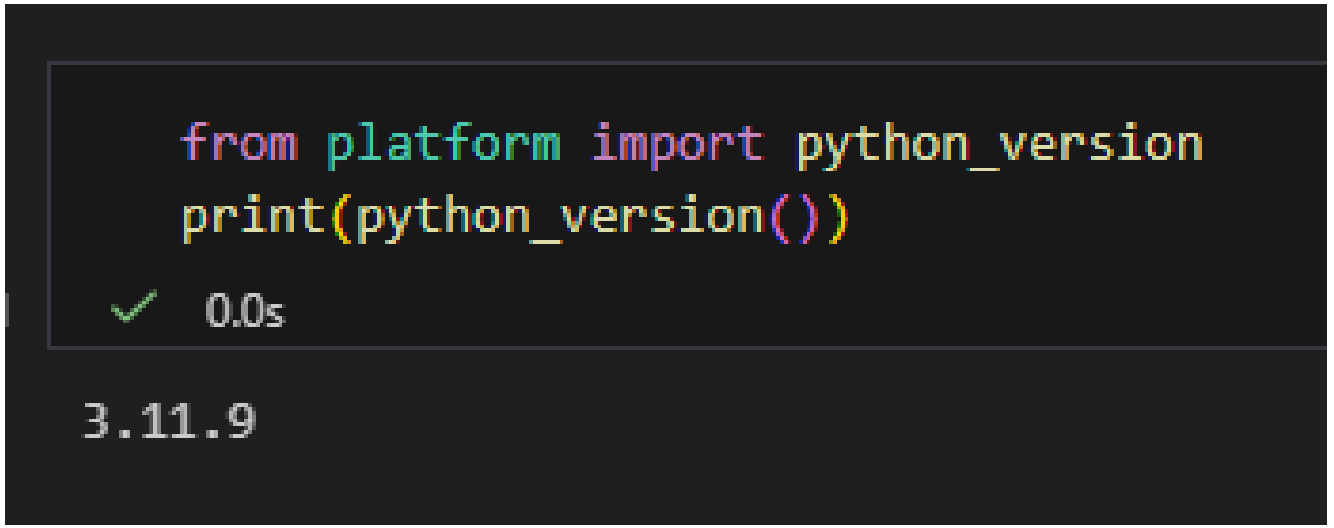


2. Den clic en “Install” o “Instalar”.



En el archivo abierto

- Esperen un momento a que se termine de instalar y, cuando vean algo similar a la captura a continuación, significa que están preparados para realizar la actividad.



Pontificia Universidad Católica del Ecuador
Seréis mis testigos

Actividad Práctica

07

Muchas gracias

Mgtr. Sebastián Tamayo