

Curso ANÁLISIS DE DATOS CON HERRAMIENTAS DE BIG DATA

36 horas



Universidad de
América

Código SNIES 1715



Este curso aporta al estudiante las herramientas para el procesamiento de grandes volúmenes de datos que no es posible analizar con herramientas tradicionales como hojas de cálculo. El enfoque es 75% práctico y 25% teórico con ejercicios reales de proyectos de Big Data.

Para el desarrollo del curso, se utilizará el lenguaje Python, librería pandas para la manipulación de datos y plotly para visualizar resultados; el enfoque será 100% en la nube con la herramienta gratuita Google Colaboratory, lo que permitirá desde cualquier computador acceder al material sin necesidad de un computador en específico.

PRESENTACIÓN

JUSTIFICACIÓN



Las herramientas tradicionales para procesamiento de datos funcionan muy bien a una escala de 20.000 a 50.000 filas si hablamos de hojas de cálculo, pero cuando la información supera este límite es muy ineficiente tratar de analizarla extrayendo patrones al ver las tablas. Así mismo, tareas repetitivas como un cálculo estadístico básico exige que el usuario seleccione manualmente columnas lo que puede generar error para obtener resultados consolidados, cuando hablamos de dos o tres archivos podría ser eficiente con mucha práctica pero si hablamos de 100 o 1000 archivos donde se necesita realizar el cálculo es algo casi titánico.



Este curso brindará a los estudiantes herramientas computacionales que les permitan convertir datos en información e información en conocimiento, automatizar tareas donde se requieren unir muchos archivos filtrarlos hacer cálculos y mostrar los resultados; se utilizará la librería Pandas, la más utilizada para lectura y procesamiento de datos del lenguaje python y plotly como herramienta de visualización interactiva.



Se realizará una introducción de cómo el Big Data se convirtió rápidamente en una herramienta clave para el desarrollo mundial; posteriormente se realizará una introducción con las generalidades del lenguaje de programación donde se profundizará en el uso de la librería pandas y finalmente cómo mostrar los resultados con visualizaciones interactivas que servirán tanto en sus proyectos de investigación como en el desarrollo de sus carreras profesionales.

OBJETIVO

GENERAL

Brindar herramientas a los estudiantes para el manejo, análisis y visualización de estructuras de datos con herramientas del lenguaje de programación Python.



ESPECÍFICOS

- + Hacer una introducción a los criterios del panorama del Big Data y cómo estas herramientas están al alcance de cualquier persona dado su contexto open source.
- + Hacer un repaso general del lenguaje de programación Python en el entorno de trabajo de Google Colaboratory.
- + Enseñar el uso de la librería pandas, para cargar datos de diferentes orígenes, filtrar, transformar y guardar la información.
- + Enseñar el uso de la librería plotly para la presentación interactiva de resultados.



PERFIL DEL ASPIRANTE

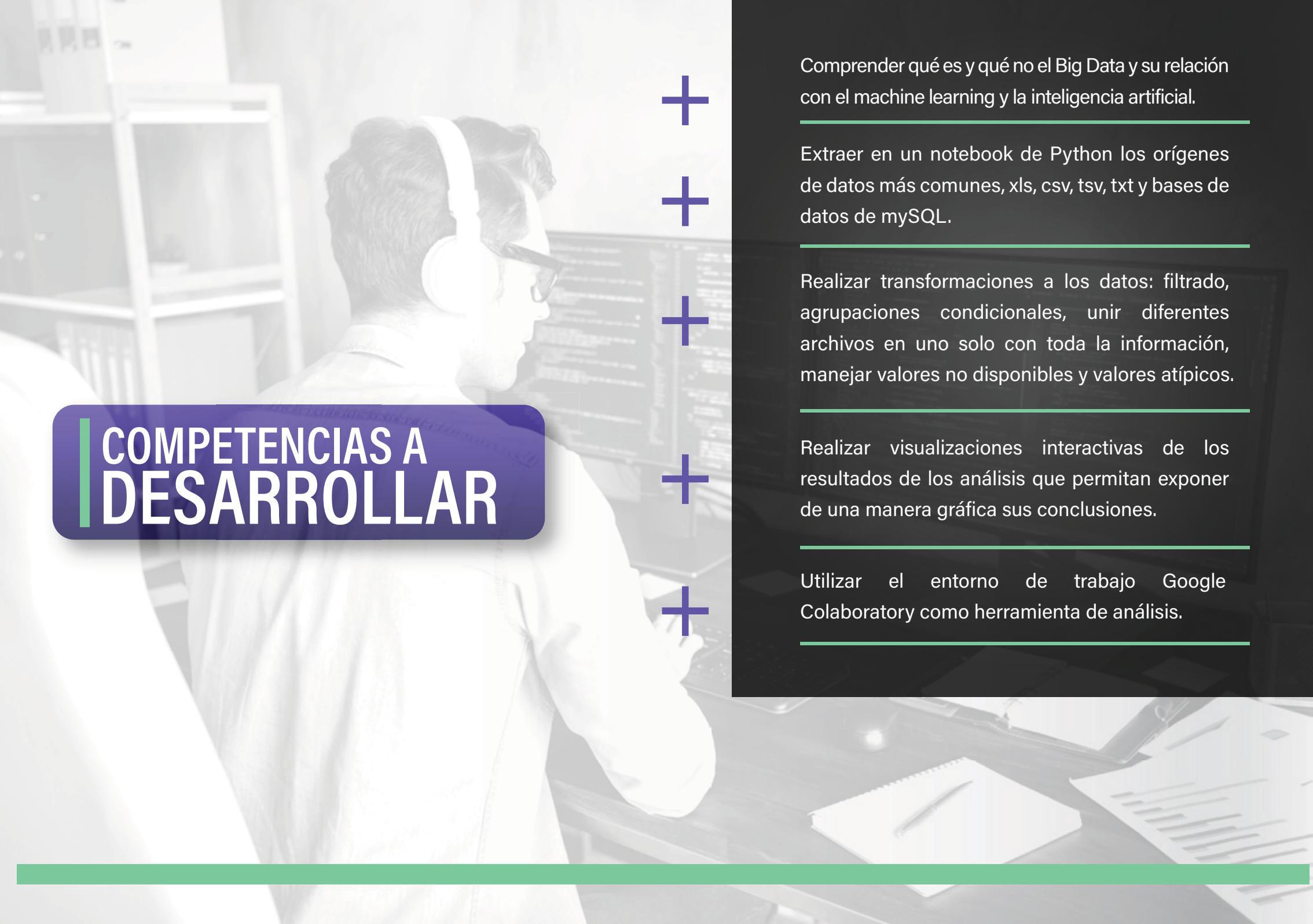
El curso está orientado para todas aquellas personas que posean un conocimiento básico de programación, que quieran manejar de manera eficiente volúmenes de datos que con herramientas tradicionales como las hojas de cálculo no se puedan analizar y que quieran adentrarse en el ámbito del análisis de datos.



METODOLOGÍA

Se propone una metodología virtual de seis horas a la semana, dividida en tres clases de dos horas donde los primeros 30 minutos son teóricos y se explica el tema a tratar, una hora donde se desarrollará de manera simultánea con los estudiantes el ejercicio de la aplicación con la herramienta Colaboratory, para que de esta manera realicen las preguntas y expresen las dificultades que pudieran tener en la resolución del problema; en los últimos 30 minutos se hace una introducción al tema de la siguiente clase y se propone un ejercicio para reforzar los conocimientos adquiridos.

Los set de datos y los notebooks con los que se trabajará en la clase estarán preparados con los enunciados para que el estudiante solo proponga la solución y los tenga como referencia para resolver problemas con las herramientas adquiridas en el curso.



COMPETENCIAS A DESARROLLAR



Comprender qué es y qué no el Big Data y su relación con el machine learning y la inteligencia artificial.



Extraer en un notebook de Python los orígenes de datos más comunes, xls, csv, tsv, txt y bases de datos de mySQL.



Realizar transformaciones a los datos: filtrado, agrupaciones condicionales, unir diferentes archivos en uno solo con toda la información, manejar valores no disponibles y valores atípicos.



Realizar visualizaciones interactivas de los resultados de los análisis que permitan exponer de una manera gráfica sus conclusiones.



Utilizar el entorno de trabajo Google Colaboratory como herramienta de análisis.



CONTENIDO TEMÁTICO

1) Módulo 1 Big Data (4 horas)

- ¿Qué es Big Data y por qué ha influido tanto en la última década?.
- 5 V del Big Data.
- Herramientas para el procesamiento de grandes volúmenes de datos.
- Importancia de las visualizaciones en la presentación de conclusiones y hallazgos de análisis de datos.
- Computación en la nube como factor de éxito en el Big Data.

2) Módulo 2 Repaso de conceptos Python (6 horas)

- Uso de Colaboratory de Google con Python.
- Tipos de datos.
- Entrada y salida de información por medio de consola.
- Bloques condicionales.
- Bloques de interacción.
- Instalación de librerías externas.

3) Módulo 3 Librería Pandas (16 horas)

- Generalidades, origen de la librería y principales usos.
- Tipos de datos soportados por Pandas.
- Transformación de tipos de datos.
- Carga de datos de diferentes orígenes.
- Selección de filas y columnas.
- Filtrado de datos.
- Filtrado condicional de datos.
- Tratamiento de datos faltantes.
- Agrupaciones condicionales de datos.
- Herramientas básicas de visualización de Pandas.
- Combinación lógica de diferentes fuentes de datos en una sola fuente.
- Manejo de campos de tipo fecha.
- Manejo básico de series temporales.
- Repaso estadística básica con Pandas.

4) Módulo 4 Librería plotly (6 horas)

- Generalidades, origen de la librería y principales usos.
- Visualizaciones de puntos.
- Visualizaciones de líneas.
- Visualizaciones de barras e histogramas.
- Visualizaciones circulares.
- Visualizaciones especiales en 3D y mapas de calor.

5) Módulo 5 Proyectos propuestos (4 horas)

- Análisis exploratorio de datos.
- Preparación de datos para algoritmos de machine learning.
- Detección de anomalías.



Curso ANÁLISIS DE DATOS CON HERRAMIENTAS DE BIG DATA

36 horas

- www.uamerica.edu.co
www.uniamerica.edu.co
- **Oficina de Mercadeo y Admisiones**
educacion.continua@uamerica.edu.co
PBX: (60 1) 3376680 Ext. 249, 149 y 156
PBX: (60 1) 6580658 Ext. 670, 677, 656 y 650
☎ (+57) 3108680919 ☎ (+57) 3108699046
☎ (+57) 3147269388 ☎ (+57) 3108809757
- **EcoCampus de Los Cerros:** Avenida Circunvalar No 20 -53
Tel: (60 1) 3376680 / Fax. (60 1) 3362941
Sede Norte: Calle 106 No. 19-18
Tel: (60 1) 6580658
Bogotá D.C., Colombia.



Universidad de
América

Código SNIES 1715