

Curso de Análisis de Datos con Python y Pandas

Perfil de Ingreso:

- Conocimientos básicos de programación en Python de preferencia (ciclos for, condicionales, listas).
- Interés en el análisis de datos y sus aplicaciones.
- Capacidad para comprender conceptos matemáticos y estadísticos básicos.

Perfil de Egreso:

- Dominio de las habilidades fundamentales de manipulación y análisis de datos con Pandas y matplotlib.
- Manejo eficiente de entornos y librerías.
- Capacidad para realizar análisis exploratorio de datos y visualización básica.
- Conocimiento sobre técnicas de preprocesamiento de datos y limpieza.
- Manejo eficiente de datos así como estructuras de datos tales como diccionarios y listas.

Módulo 1: Introducción a Python y Pandas

Objetivos:

- Introducir a los estudiantes al lenguaje de programación Python y su relevancia en el análisis de datos.
- Presentar la biblioteca Pandas y sus principales funcionalidades para la manipulación y análisis de datos.
- Configurar el entorno de trabajo y familiarizarse con los primeros pasos utilizando Jupyter Notebook.

Temas a Cubrir:

- 1. Introducción a Python:**
 - Importancia de Python en el ámbito del análisis de datos.
 - Principales características y ventajas de Python como lenguaje de programación.
- 2. Presentación de Pandas:**
 - Visión general de Pandas y su utilidad en el análisis de datos.
 - Funcionalidades principales de Pandas, como la manipulación de datos tabulares.
- 3. Configuración del Entorno de Trabajo:**
 - Instalación de Python y configuración del entorno de desarrollo.
 - Introducción a Jupyter Notebook como una herramienta interactiva para el análisis de datos.

Módulo 2: Creación de Entornos y Manejo de Datos

Se busca dar una introducción al manejo de entornos en Python para su uso en la instalación de librerías, módulos y clases. A su vez, una pequeña introducción al manejo de variables así como listas y diccionarios.

Objetivos:

- Introducir conceptos fundamentales sobre la creación de entornos para análisis de datos (Conda, venv).
- Familiarizar a los estudiantes con las operaciones básicas de manipulación de diccionarios y listas en Python (List comprehension, manejo de diccionarios, etc).

Temas a Cubrir:

1. **Entornos en Python:**
 - Introducción a Conda y venv para la gestión de entornos virtuales.
 - Creación de un entorno virtual para el proyecto de análisis de datos.
2. **Manejo de Variables:**
 - Concepto de variables en Python y su asignación.
 - Operaciones básicas con variables: asignación, actualización y eliminación.
3. **Listas y Diccionarios:**
 - Introducción a las estructuras de datos de listas y diccionarios en Python.
 - Operaciones básicas con listas: creación, acceso a elementos, iteración.
 - Operaciones básicas con diccionarios: creación, acceso a elementos, actualización.

Módulo 3: Manipulación de Datos con Pandas

Objetivos:

- Familiarizar a los estudiantes con las técnicas de manipulación de datos utilizando la biblioteca Pandas.
- Enseñar cómo cargar datos en Pandas desde diferentes fuentes y realizar operaciones básicas de manipulación.

Temas a Cubrir:

1. **Carga de Datos en Pandas:**
 - Importación de datos desde archivos CSV, Excel y otras fuentes utilizando Pandas.
 - Exploración inicial de conjuntos de datos recién cargados.
2. **Operaciones Básicas de Manipulación de Datos:**
 - Selección de columnas y filas de un DataFrame.
 - Filtrado de datos basado en condiciones.
 - Ordenamiento de datos según valores de columnas específicas.

3. Conceptos de Índices y Columnas:

- Introducción a los índices y su importancia en Pandas.
- Manipulación de columnas y renombrado de columnas en un DataFrame.

Módulo 4: Preprocesamiento de Datos

Objetivos:

- Familiarizar a los estudiantes con las técnicas de preprocesamiento de datos para la preparación de conjuntos de datos para análisis y modelado.
- Enseñar métodos para limpiar y transformar datos, así como introducir el concepto de feature engineering.

Temas a Cubrir:

1. Limpieza de Datos:

- Identificación y manejo de valores faltantes en los conjuntos de datos.
- Eliminación de valores duplicados y registros erróneos.

2. Transformación de Datos:

- Normalización de datos para llevarlos a una escala común.
- Estandarización de datos para tener una distribución normal.
- Codificación de variables categóricas utilizando técnicas como One-Hot Encoding.

3. Introducción al Feature Engineering:

- Definición y concepto de feature engineering en el contexto del análisis de datos.
- Ejemplos de técnicas de feature engineering, como la creación de nuevas características a partir de las existentes.

Módulo 5: Manejo de DataFrames en Pandas

En esta sesión, explicaremos técnicas básicas y avanzadas para el manejo de DataFrames en Pandas, que son fundamentales para realizar análisis de datos eficientes y completos.

Temas a Cubrir:

1. Merge y Concatenación:

- Combinación de DataFrames utilizando las funciones merge y concat en Pandas.
- Especificación de columnas clave y tipos de unión (inner, outer, left, right) en merge.
- Concatenación vertical y horizontal de DataFrames con concat.

2. Apply:

- Aplicación de funciones a nivel de fila o columna utilizando la función apply.

- Uso de funciones lambda y funciones definidas por el usuario para aplicar transformaciones complejas.
- 3. **GroupBy y Agregación:**
 - Agrupación de datos en Pandas utilizando la función groupby.
 - Aplicación de funciones de agregación a grupos de datos utilizando GroupBy.
 - Utilización de funciones de agregación comunes, como sum, mean, count.
- 4. **Manejo de Fechas y Tiempo:**
 - Conversión de columnas a tipo de dato datetime utilizando funciones de Pandas.
 - Operaciones de filtrado y selección basadas en fechas y tiempos.
- 5. **Ordenamiento de Datos:**
 - Ordenamiento de datos en un DataFrame utilizando la función sort_values.
 - Especificación de columnas clave y configuración de parámetros opcionales como ascending.

Módulo 6: Visualización de Datos

Objetivos:

- Introducir a los estudiantes en las técnicas de visualización de datos para comunicar de manera efectiva la información contenida en un conjunto de datos.
- Familiarizar a los estudiantes con las bibliotecas de visualización de datos en Python utilizando matplotlib.
- Enseñar cómo crear diferentes tipos de gráficos y visualizaciones para explorar y presentar datos de manera efectiva.

Temas a Cubrir:

1. **Importancia de la Visualización de Datos:**
 - Exploración de la importancia de la visualización de datos en el análisis y la presentación de información.
 - Ejemplos de cómo una buena visualización puede ayudar a comprender mejor los datos y tomar decisiones más informadas.
2. **Bibliotecas de Visualización en Python:**
 - Introducción a las bibliotecas populares de visualización de datos en Python, como Matplotlib y Seaborn.
 - Comparación de las características y capacidades de cada biblioteca.

Módulo 7: Limpieza y visualización desde un dataset sucio

En esta sesión final, pondremos en práctica todo lo aprendido en los módulos anteriores mediante la lectura y manipulación de un conjunto de datos real, que puede contener desafíos y problemas típicos encontrados en la práctica del análisis de datos.

Objetivos:

- Aplicar todas las técnicas aprendidas en los módulos anteriores para abordar un conjunto de datos reales y completar tareas de análisis específicas.
- Practicar la carga, limpieza, manipulación, análisis y visualización de datos utilizando Pandas y otras herramientas de Python.
- Desarrollar habilidades para resolver problemas de análisis de datos de manera efectiva y eficiente.

Se espera realizar:

1. **Lectura de Datos Sucios:**
 - Cargar el conjunto de datos proporcionado, que puede contener valores faltantes, duplicados u otros problemas típicos de los datos reales.
2. **Limpieza y Preprocesamiento:**
 - Identificar y manejar valores faltantes, duplicados u otros problemas de calidad de datos.
 - Realizar preprocesamiento según sea necesario, como la normalización, estandarización o codificación de variables.
3. **Análisis Exploratorio:**
 - Utilizar técnicas aprendidas como groupby, agregaciones y visualizaciones para obtener información significativa.
4. **Manipulación de DataFrames:**
 - Aplicar técnicas avanzadas de manipulación de DataFrames, como merge, concat y apply según sea necesario para resolver problemas específicos.
5. **Visualización de Datos:**
 - Crear visualizaciones efectivas para comunicar los hallazgos y resultados del análisis de manera clara y concisa.