TALLER BIG DATA—PYTHON

PRESENTADO POR:

Karen lorena Vergara Suarez

PRESENTADO A:

Álvaro Pérez Niño

ADSO:

2501259

CENTRO DE SERVICIOS Y GESTION EMPRESARIAL

Medellín—Antioquia

ACTIVIDAD

Quiero que imagine que está en el aula de clases donde los estudiantes aprenden en línea usando computadoras y dispositivos, este taller se basa en información recopilada de ese sistema en línea que se recolecto con los estudiantes, este nos ayudará a entender cómo los estudiantes usan la plataforma de aprendizaje y cómo les va en sus estudios, con la información que depositaron.

¿De dónde provienen los datos?

Los datos provienen de un sistema de aprendizaje en línea llamado "Territorio", es una plataforma en línea donde los estudiantes aprenden desde cualquier lugar con acceso a internet, estos datos se recopilan automáticamente cuando los estudiantes interactúan con la plataforma, como cuando ven videos o hacen actividades.

¿Qué Información Contienen los Datos?

Los datos cuentan la historia de 40 estudiantes diferentes, los detalles incluyen cosas como:

\* si trabajan

\*dónde viven

\* cómo les va en diferentes materias

\* cuánto participan en la plataforma

\*y cuán bien se desempeñan en sus tareas y exámenes.

¿Qué Haremos con los Datos?

Primero, observamos los datos y se hacen algunas cuentas básicas para entenderlos mejor, luego, se eliminará información que no nos ayuda mucho o que está incompleta, después de limpiar los datos, se vuelven a analizar para encontrar patrones interesantes que se nos haya pasado por alto.

¿Por Qué Esto Es Importante?

Con esta información que recolectamos, podemos entender mejor cómo los estudiantes se manejan y aprenden en línea, se puede descubrir qué ayuda a los estudiantes a tener éxito en el aprendizaje y qué cosas le pueden dificultar, esto puede ayudar a la institución educativa a tomar decisiones importantes, como, por ejemplo; cómo mejorar la plataforma, cómo ofrecer mejores clases y cómo apoyar a los estudiantes que están teniendo dificultades.

Conclusión

El análisis de datos educativos nos permite adentrarnos en cómo los estudiantes se comportan cuando usan una plataforma de aprendizaje en línea, al explorar aspectos como su información personal y su rendimiento académico, podemos descubrir información valiosa que ayuda a tomar decisiones y planificar estrategias educativas. Este taller nos da la oportunidad de desentrañar los datos, descubrir tendencias y presentar ideas clave que pueden cambiar la forma en que se maneja el aprendizaje y cómo se brindan servicios educativos personalizados.

import pandas as pd

#import numpy as pd

#import matplot as pd

from datetime import datetime

# Cargar los conjuntos de datos desde archivos Excel (reemplaza 'archivo1.xlsx' y 'archivo2.xlsx' con los nombres de tus archivos)

data1 = pd.read\_excel('DatosP01.xlsx')

data2 = pd.read\_excel('DatosP02.xlsx')

# Mostrar las primeras filas de cada DataFrame para verificar la carga exitosa

print("Data1:")

print(data1.head())

print("\nData2:")

print(data2.head())

# Especifica las columnas clave para realizar la unión

key\_columns = ['Id']

# Realiza la unión de los DataFrames basada en las columnas clave

merged\_data = pd.merge(data1, data2, on=key\_columns, how='inner')

# Reemplazar valores incoherentes en la columna 'Sexo'

data1['Sexo'] = data1['Sexo'].replace(['Femenino', 'Masculino', 'Mujer'], ['F', 'M', 'F'])

# Fusionar los conjuntos de datos usando la columna 'Id' como clave

data = pd.merge(data1, data2, on='Id')

# Calcular el promedio de calificaciones

data['Promedio'] = (data['Sistemas'] + data['Matematicas'] + data['Ingles'] + data['Comunicación']) / 4

# Filtrar aprendices aprobados y no aprobados

aprobados = data[data['Promedio'] > 3.5]

no\_aprobados = data[data['Promedio'] <= 3.5]

# Obtener características de aprendices aprobados y no aprobados

data['FechaNacimiento'] = pd.to\_datetime(data['FechaNacimiento'], format='%d/%m/%Y')

# Calcula las edades a partir de la fecha de nacimiento

hoy = datetime.now()

data['Edad'] = hoy.year - data['FechaNacimiento'].dt.year

# Filtra los aprendices aprobados (supongamos que 'Promedio' es la columna con el promedio de calificaciones)

aprobados = data[data['Promedio'] > 3.5]

noaprobados = data[data['Promedio'] <= 3.5]

# Calcula el promedio de edades de los aprendices aprobados

promedio\_edad\_aprobados = aprobados['Edad'].mean()

promedio\_edad\_noaprobados = noaprobados['Edad'].mean()

caracteristicas\_aprobados = aprobados[['Id', 'Nombre', 'Sexo', 'Estrato', 'Trabaja', 'Enfermedad', 'Ubicacion', 'Promedio']]

caracteristicas\_no\_aprobados = no\_aprobados[['Id', 'Nombre', 'Sexo', 'Estrato', 'Trabaja', 'Enfermedad', 'Ubicacion', 'Promedio']]

# Calcular el promedio de edad de los aprendices aprobados y no aprobados

#promedio\_edad\_aprobados = caracteristicas\_aprobados['Edad'].mean()

#promedio\_edad\_no\_aprobados = caracteristicas\_no\_aprobados['Edad'].mean()

# Imprimir resultados

print("Características de los aprendices aprobados:")

print(promedio\_edad\_aprobados)

print("\nCaracterísticas de los aprendices no aprobados:")

print(promedio\_edad\_noaprobados)

print("\nPromedio de edad de los aprendices aprobados:", promedio\_edad\_aprobados)

print("Promedio de edad de los aprendices no aprobados:", promedio\_edad\_noaprobados)

# Reemplazar las comas por puntos en las columnas numéricas

columnas\_numericas = ['Asistencia', 'Entregas\_Completas', 'Sistemas', 'Matematicas', 'Ingles', 'Comunicación']

data2[columnas\_numericas] = data2[columnas\_numericas].apply(lambda x: x.str.replace(',', '.'))

# Convertir las columnas numéricas al tipo float

data2[columnas\_numericas] = data2[columnas\_numericas].astype(float)

# Calcular el promedio ponderado para data2

data2['Promedio'] = (data2['Sistemas'] + data2['Matematicas'] + data2['Ingles'] + data2['Comunicación']) / 4

# Agregar una columna para indicar si el aprendiz aprobó o no en data2

data2['Aprobado'] = data2['Promedio'] > 3.5

# Filtrar aprendices que usaron el LMS en data2

usaron\_lms\_data2 = data2[data2['Uso\_LMS'] == 'Si']

# Obtener características de aprendices que aprueban y no aprueban en data2

aprobados\_data2 = usaron\_lms\_data2[usaron\_lms\_data2['Aprobado']]

no\_aprobados\_data2 = usaron\_lms\_data2[~usaron\_lms\_data2['Aprobado']]