Elevate Labs

DAY 1 - TASK 1

DATA CLEANING AND PREPROCESSING

Introduction

Analyse, preprocess and clean the data of medical appointments done by patients. The data consist of 110527 rows and 14 columns
Let’s analyse each column

Column name	Data type	Description
patient_id	float	Identification of patient
appoinment_id	int	Identification of each appointment
gender	object	Gender of patient (either ‘F’- female or ‘M’- male)
schedule_date	object	Date on the appointment is scheduled
appoinment_date	object	Date of the appointment/ date of visiting doctor
age	int	Age of patient
appoinment_place	object	Place where the appointment takes place
scholarship	int	Whether receiving the ‘Bolsa familia’ aid (either 0 or 1)
hyper_tension	int	Whether patient has hypertension or not (either 0 or 1)
diabetes	int	Whether patient has diabetes or not (either 0 or 1)
alcoholic	int	Whether patient is alcoholic or not (either 0 or 1)
handicapped	int	Whether patient is handicapped or not (0,1,2,3,4)
sms_received	int	Whether patient has received sms verification or not (either 0 or 1)
no_show	object	Whether patient has showed on time for the appointment or not (either Yes or No)

Steps of task

1. Data and basic libraries importing.
    a. We imported basic libraries like NumPy, Pandas, Matplotlib, and Seaborn.
    b. We imported the .csv file of appointment data.

2. Data cleaning and preprocessing.
    a. Column headers name changing to appropriate preferred name:Changing the case of column headers from upper case to lowercase with '_'.
    b. Finding null values and duplicates: Finding any rows with no values and any rows with the same value repeated with isnull() and duplicated() functions using Pandas.
    c. Data type changing: Changing the columns with date values but given an object data type to datetime datatype using pd.to_datetime() function using Pandas.
    d. Column analysing, outlier detection and cleaning: Analysing every column, finding outliers and wrong records and cleaning the data using encoding categorical values and replacing wrong values

Result

The data set is cleaned by removing, replacing and changing wrong entries, outliers, duplicates and wrong data type columns