Data Cleaning Summary - Medical Appointment No Shows

This document provides a detailed summary of the data cleaning process performed on the Medical Appointment No Shows dataset. The dataset was cleaned using Python and Pandas, and various operations were applied to handle missing values, remove duplicates, and standardize data.

1. Dataset Overview

The dataset contains information about medical appointments, including patient details, appointment dates, and whether the patient showed up for the appointment or not. The columns are as follows:
PatientId, AppointmentId, Gender, ScheduledDay, AppointmentDay, Age, Neighbourhood, Scholarship, Hipertension, Diabetes, Alcoholism, Handcap, SMS_received, and No-show.

2. Data Cleaning Steps

2.1. Handling Missing Values

The dataset was checked for missing values using the isnull() method. No missing values were found across any of the columns, so no further action was needed in this regard.

2.2. Removing Duplicates

The dataset was also checked for duplicate rows using the duplicated() method. No duplicate rows were identified, so no rows were removed.

2.3. Handling Age Column

The age column was reviewed, and it was found that some rows had invalid age values, such as age = 0 and age = -1. These were deemed invalid, and rows with age less than or equal to 0 were removed. After this operation, the dataset contained only valid age entries.

2.4. Simplifying Handcap Column

The handcap column contained values other than 0 and 1, such as 2, 3, and 4, which were considered invalid for binary classification. These values were converted to 1, and the column was standardized to have only values 0 and 1, where 1 indicates that the patient is handicapped.

2.5. Standardizing Text Columns

The text columns such as Gender and No-show were standardized to ensure consistency. The Gender column, which had values 'F' and 'M', was kept as is, while the No-show column had values 'Yes' and 'No' which were already standardized.

2.6. Converting Date Columns

The ScheduledDay and AppointmentDay columns were originally in string format, so they were converted to datetime format, retaining only the date part (removing time).

3. Final Dataset Structure

After the cleaning process, the dataset had the following columns:
PatientId, AppointmentId, Gender, ScheduledDay, AppointmentDay, Age, Neighbourhood, Scholarship, Hipertension, Diabetes, Alcoholism, Handcap, SMS_received, No-show.

The final dataset was saved as 'cleaned_medical_appointments.csv' and is ready for further analysis or modeling.