Healthcare Appointment No-Show Prediction

Introduction:

In the healthcare industry, missed appointments (no-shows) result in resource underutilization, increased waiting times, and higher operational costs. Predicting which patients are likely to miss their appointments can help medical facilities take preventive actions such as sending reminders, offering flexible rescheduling, or optimizing appointment allocation. This project aims to use machine learning to analyze patient appointment data and provide insights that can reduce no-show rates.

Abstract:

Using a dataset of over 100,000 patient appointments from a public health database, I developed a machine learning model to predict appointment no-shows. The model is based on features such as age, gender, SMS reminders, number of previous no-shows, days between scheduling and the appointment, and appointment weekday. We trained a Decision Tree Classifier to identify patterns in the data and evaluated it using classification metrics. In addition to prediction, the project includes a Power BI dashboard for visualizing trends and supporting decision-making. The ultimate goal is to enable healthcare providers to improve scheduling efficiency and patient outcomes through data-driven strategies.

Tools Used:

- Python (Pandas, Scikit-learn): For data processing and machine learning

- Jupyter Notebook: For code development and experimentation

- Power BI: For interactive visual analytics and dashboard creation

Steps Involved in Building the Project:

1. Data Acquisition and Cleaning:

Loaded the dataset and removed invalid entries such as negative ages. Standardized column names and date formats.

2. Feature Engineering:

Created additional features including days between scheduling and appointment, appointment weekday, and binary encoding for No-show and separated age into Age groups using bins.

3. Exploratory Data Analysis (EDA):

Identified trends such as higher no-show rates on Mondays and among younger patients. Assessed the impact of SMS reminders and waiting time.

4. Model Development:

Trained a Decision Tree Classifier and evaluated it using a train-test split. Used accuracy, precision, recall, and F1-score to measure performance.

5. Visualization:

Developed Power BI dashboards to show no-show distributions across age groups, weekdays, and SMS_received. Included charts for easier interpretation by non-technical stakeholders.

6. Optimization Insights:

Recommended policy changes such as targeted reminders for high-risk groups and scheduling adjustments to minimize no-shows based on historical data.

Conclusion:

The project demonstrated the effectiveness of data-driven approaches in addressing healthcare inefficiencies. The decision tree model successfully predicted appointment attendance with a reasonable degree of accuracy. Key influencing factors included the presence of SMS_received, patient age, time between scheduling and the appointment, and the day of the week. Insights from Power BI dashboards provided a foundation for targeted operational improvements. Implementing the recommendations could lead to more reliable scheduling, better resource utilization, and improved patient care. Future work could include integrating additional features such as weather conditions, patient distance from the clinic, and historical medical records to enhance model accuracy.