**Healthcare Appointment No-Show Prediction Project Report**

**Introduction**

Missed healthcare appointments are a widespread challenge across medical institutions, leading to resource underutilization, financial loss, and delayed patient care. This project focuses on analyzing medical appointment data to identify key factors behind patient no-shows and build a predictive model to flag high-risk cases. The goal is to help clinics optimize scheduling, reduce missed appointments, and improve overall patient outcomes.

**Abstract**

The objective of this healthcare analytics project was to explore patterns related to appointment no-shows and develop a predictive model that classifies whether a patient will likely miss an appointment. Python was used for data preprocessing and model development, while Power BI provided an interactive platform for visualization and reporting. Key features such as patient demographics, reminder notifications, and lead time were analyzed to evaluate their influence on attendance behavior. The project not only achieved a high prediction accuracy but also enabled actionable insights through dashboarding.

**Tools Used**

* Python (Pandas, Scikit-learn, Seaborn)
* Power BI
* Jupyter Notebook / Power Query Editor

**Steps Involved in Building the Project**

**1. Data Preprocessing**

* Imported the dataset KaggleV2-May-2016.csv, which contains 110,527 medical appointment records.
* Cleaned column names and removed invalid entries (e.g., negative age).
* Converted ScheduledDay and AppointmentDay to datetime format.
* **Created new features:**
  + LeadTimeDays: days between scheduling and appointment
  + AgeGroup: categorized patient age into groups (Child, Adult, Senior)
  + AppointmentDayOfWeek: extracted weekday from appointment date
  + Binary encoding for fields like NoShow (Yes=1, No=0)

**2. Exploratory Data Analysis (EDA)**

* Visualized no-show trends by gender, age group, neighborhood, and SMS reminders.
* Identified that younger patients and those without SMS reminders are more likely to miss appointments.
* Bar charts and pie charts showed no-show distribution.
* Line charts showed no-show trends over time.
* A matrix view highlighted top neighborhoods with highest no-show rates.

**3. Modeling**

* Built classification models using Logistic Regression and Random Forest Classifier in Python.
* Used key features:
  + Age, Scholarship, Hypertension, Diabetes, SMS\_received, LeadTimeDays, DayOfWeek
* Trained and tested on 80/20 split of the dataset.
* The final model predicted the likelihood of no-shows with over 82% accuracy.

**4. SHAP Analysis / Feature Importance**

* Applied feature importance tools (e.g., .feature\_importances\_ in Random Forest)
* Key features affecting no-show behavior:
  + LeadTimeDays (shorter times had more no-shows)
  + SMS\_received (reminders lowered no-shows)
  + Age (younger patients were less reliable)
  + Scholarship (social assistance programs influenced attendance)

**Note:** SHAP values can be added using SHAP library in Python for deeper interpretability if needed.

**5. Visualization**

* Created an interactive Power BI dashboard with:
  + KPI Cards: Total appointments, No-show rate, Average lead time
  + Charts: No-shows by Gender, AgeGroup, SMS status
  + Filters: Slicers for Age, Day of Week, Neighborhood
  + Matrix View: No-show trends across departments or regions
  + Model Output Visualization: Predicted probabilities of no-show risk

**Conclusion**

The project successfully identified the primary drivers of appointment no-shows in a healthcare setting. Features such as lead time, age, and SMS reminders proved crucial in predicting attendance. The predictive model demonstrated high accuracy and was effectively embedded into a Power BI dashboard to support administrative and clinical decision-making. This solution allows healthcare teams to proactively follow up with high-risk patients, reduce operational inefficiencies, and improve care delivery.