Table of Contents

1. Problem Statement and Background

Healthcare institutions across the globe face a recurring and costly issue: patients missing their medical appointments. These no-shows contribute to inefficient utilization of healthcare resources, delay in care delivery, and potential worsening of health outcomes—not only for those who miss appointments, but also for others due to congested scheduling systems. In large urban hospitals and clinics, no-show rates can range from 15% to 30%, depending on population demographics and clinic policies.

To address this challenge, predictive analytics and machine learning offer a powerful solution. By analysing structured historical data—such as patient demographics (e.g., age and gender), social context (e.g., participation in welfare programs), and behavioural indicators (e.g., response to SMS reminders)—it becomes possible to estimate the probability that a patient will miss a future appointment.

For example, the feature **WaitingDays**, which measures the delay between scheduling and the actual appointment date, was found to influence attendance behaviour significantly. Patients with longer waiting times were more likely to miss their appointments. Similarly, patients who **received an SMS reminder** showed a modest but consistent increase in attendance rates, supporting the importance of timely communication.

By harnessing such patterns through machine learning models, healthcare providers can take proactive steps—like rescheduling, sending additional reminders, or offering telemedicine alternatives—to reduce no-shows. This leads to better appointment adherence, optimized clinical operations, improved health outcomes, and ultimately, more sustainable healthcare delivery.

.

2. Dataset Description

The dataset used for this study was sourced from a public health database, originally made available through Kaggle. This dataset includes over **110,000** records of patient appointments collected from a public healthcare system in Brazil, **approximately 20%** of appointments were missed, as identified by the No_show column. This class imbalance is a significant barrier to achieving efficient healthcare workflows.

Key Features:

- **ScheduledDay** and **AppointmentDay:** Represent the time gap between appointment booking and the actual visit.
- **No_show:** The binary target indicating whether a patient missed the appointment.
- **Gender and Age:** Reflect basic demographic information.
- **Neighbourhood:** Indicates the geographic region where the clinic is located.
- **Scholarship:** Represents whether the patient is enrolled in welfare programs.

- **Hypertension, Diabetes, Alcoholism:** Medical condition flags.
- **SMS_received:** Indicates whether the patient was sent a reminder message.

This comprehensive feature set captures both patient-specific and contextual variables essential for a robust predictive model.

3. Methodology

A structured and step-by-step approach was used to build a machine learning model that predicts whether a patient will miss a medical appointment. The process included data cleaning, feature creation, encoding, data splitting, and finally training and evaluating models.

3.1 Data Collection and Preprocessing

The dataset was loaded using the **pandas** library and checked for errors or inconsistencies. Here are the steps taken:

- There were **no missing values** in the dataset.
- Records with **invalid age entries** (like negative ages) and those with **zero or negative waiting days** were removed, as they didn't reflect realistic appointment scenarios.
- Two important columns — **ScheduledDay** and **AppointmentDay** — were converted to proper date formats using Python datetime functions.
- A new column, **WaitingDays**, was created to calculate the number of days between booking and appointment.

This new feature helped to identify patterns: **longer waiting times often led to higher no-show rates**.

- Duplicate entries were identified and dropped to avoid bias.
- After cleaning, around **99,000 records** were kept for modeling.

3.2 Feature Engineering

New features were added to improve model accuracy:

- **WaitingDays**: Time gap between booking and appointment. It was found to be a strong predictor of no-shows.
- **AppointmentDayOfWeek**: Extracted to capture patient behaviour on different weekdays.

    - For example, **Mondays** showed slightly more missed appointments than mid-week days.

- **IsWeekend**: This binary feature checks if the appointment falls on a Saturday or Sunday. Slightly higher no-show rates were noted on weekends.

These new features helped to better understand and capture temporal patterns in patient behaviour.

3.3 Encoding and Data Splitting

To prepare the data for machine learning models, encoding and splitting steps were applied:

- **One-hot encoding** was used for non-numeric columns like Gender, Neighbourhood, and Appointment Day of the Week. This turned categories into numeric format.
- The target column **No_show** was label-encoded:
    - 'No' → 0 (patient attended)
    - 'Yes' → 1 (no-show)

The data was split into:

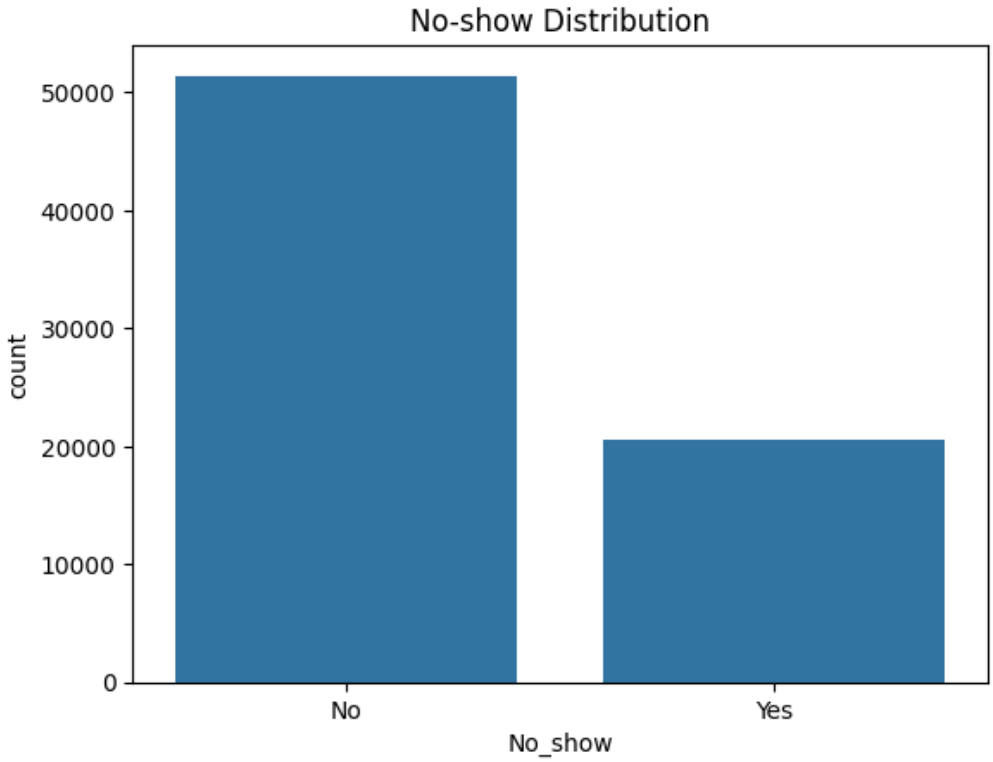
- **Training set**: 80% of the data
- **Testing set**: 20% of the data
  Stratified sampling ensured that both sets had a similar percentage of no-show cases.
- **StandardScaler** was used to normalize continuous features like Age and WaitingDays, especially useful for models like Logistic Regression.

This structured approach helped build a reliable and generalizable machine learning model. Features based on time and behaviour — such as how long patients waited and what day their appointment was — added significant value to the prediction process.

4. Exploratory Data Analysis (EDA) and Visualizations

4.1 Class Distribution

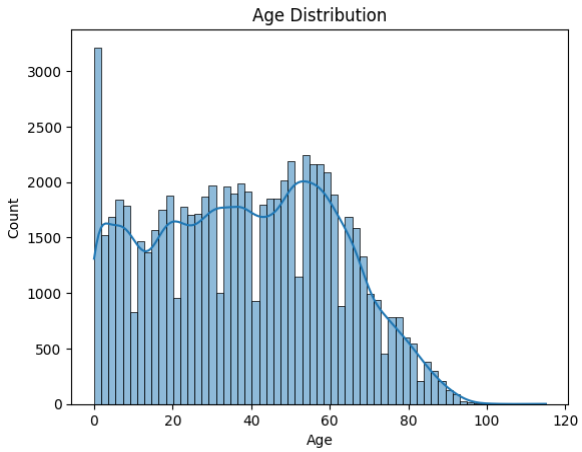
As observed during early analysis, the dataset is imbalanced — the number of patients who attend appointments significantly outweighs those who do not. This imbalance makes naive accuracy misleading and calls for balanced metrics like precision, recall, and F1-score for proper model evaluation.

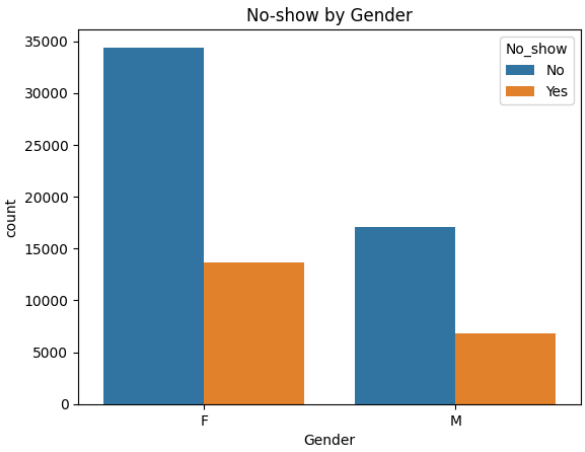


4.2 Socio-Demographic Influence

Demographic exploration revealed the following:

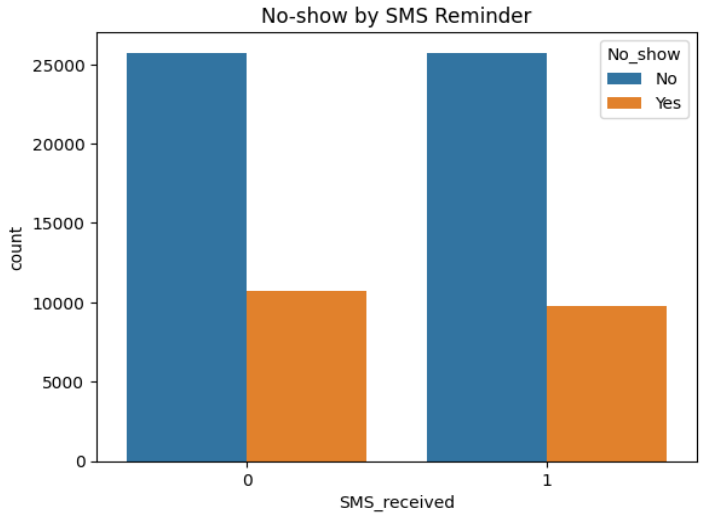
- **Age**: Older individuals tend to show better attendance patterns, potentially due to chronic conditions requiring regular monitoring.
- **Gender**: Attendance rates showed no significant skew across genders, indicating that gender alone is not a strong predictor.
- **Scholarship**: Patients on government welfare showed slightly varied behaviour, possibly due to socioeconomic constraints impacting transportation or prioritization of health.





4.3 Communication and Behavioural Insights

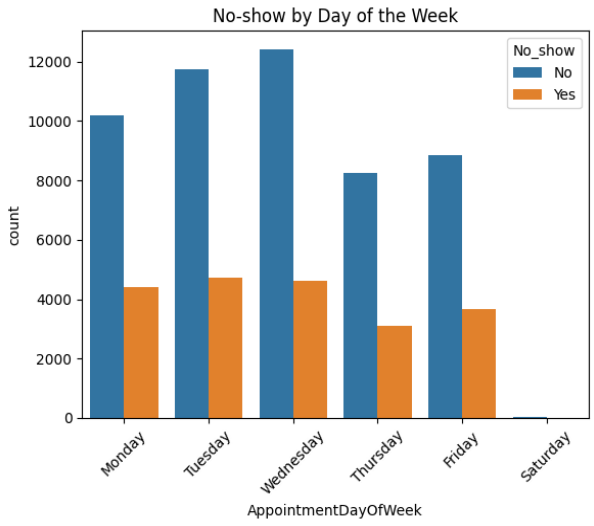
The **SMS_received** feature offered important behavioural insight. Those who received reminders were more likely to attend their appointments, affirming the value of communication in healthcare compliance. However, not all patients responded to reminders equally, suggesting room for more personalized engagement strategies.

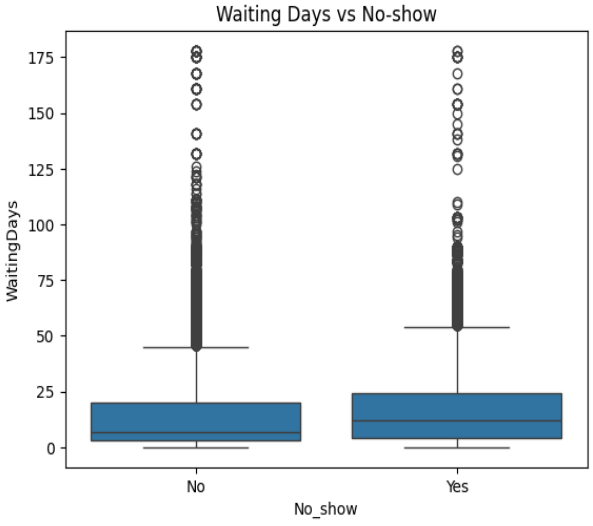


4.4 Temporal Patterns

Analysing **AppointmentDayOfWeek** revealed fluctuations in attendance by day:

- Appointments scheduled on **Mondays** had slightly higher no-show rates, possibly due to weekend fatigue or rescheduling.
- **Mid-week appointments** saw higher adherence, possibly reflecting more regular routines.
- Saturday clinics showed varied results, depending on patient demographics.

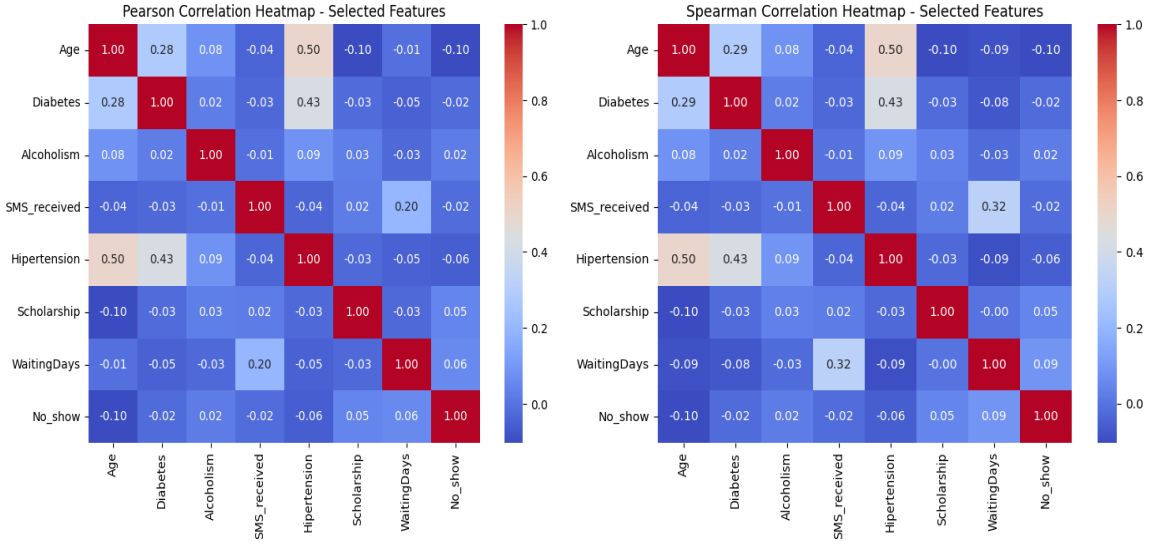




4.5 Correlation and Dependency Analysis

Heatmaps using Pearson and Spearman methods were used to explore relationships between features:

- **WaitingDays** had a noticeable positive relationship with no-show likelihood.
- **Age** exhibited a non-linear association; extreme age groups (young children and elderly) displayed varying compliance patterns.
- Medical conditions showed weak correlations independently but were hypothesized to play stronger roles in multivariate interactions.



5. Model Development and Evaluation

5.1 Overview of Classification Models

Two primary classifiers were developed:

- **Logistic Regression** for its interpretability and simplicity.
- **Random Forest Classifier** for its ability to capture complex, non-linear relationships.

5.2 Logistic Regression

As a linear model, logistic regression was quick to train and provided easily interpretable coefficients. However, it was unable to adequately capture more subtle, interactive relationships between features, especially under class imbalance conditions. Its performance was generally conservative — it was better at predicting patients who *would* attend than identifying no-shows.

5.3 Random Forest Classifier

This ensemble-based model, composed of decision trees, improved classification performance. By aggregating multiple tree outcomes, it generalized well and captured complex patterns such as age-reminder interactions or location-specific behaviours.

5.4 Threshold Tuning and Class Imbalance Handling

To counteract the dataset's imbalance, a classification threshold lower than 0.5 was applied. This increased the sensitivity of detecting no-shows, though at the cost of some false positives. Class weighting was also employed to penalize misclassification of the minority class.

5.5 Model Comparison

The Random Forest consistently outperformed Logistic Regression across balanced accuracy, precision, and recall. Additionally, it provided feature importance rankings — useful for interpreting which features most influenced predictions. Among the top contributors were:

- **WaitingDays**
- **SMS_received**
- **Age**
- **Day of the Week**

6. Results and Discussion

The findings of this project illustrate that machine learning can effectively support appointment management systems. The Random Forest model emerged as the most promising approach for real-world application due to its robustness and superior performance on imbalanced data.

Key Takeaways:

- Predictive models can identify no-show risk with high precision when enhanced with behavioural and temporal features.
- SMS reminders, though helpful, are not a silver bullet and may benefit from reinforcement strategies like follow-up calls.
- Healthcare providers can use these predictions to create dynamic overbooking schedules or prioritize at-risk patients for reminders.

Even though only a few variables showed direct linear correlation with the target, multivariate modeling allowed for capturing hidden patterns and interactions.

7. Limitations and Future Work

While our analysis yielded actionable insights, some limitations remain:

- The dataset lacks information on weather, travel distance, or personal patient history (e.g., past no-shows), which could enhance predictive power.
- Temporal trends such as seasonal effects or public holidays are not modeled.
- The effect of different communication mediums (calls vs SMS) is not considered.

Future enhancements could include:

- Use of deep learning for pattern recognition across longer patient histories.
- Integration of natural language processing (NLP) to assess appointment reasons.
- Real-time dashboard integration for hospital staff to act on live predictions.

8. Conclusions

This project shows how **machine learning** can help solve real problems in healthcare — especially the issue of patients missing their medical appointments. By looking at past data such as age, gender, whether they received a reminder, and how long they had to wait, we can **predict who might not show up**. This helps clinics take **action early**, like sending more reminders or adjusting their schedules.

Out of both the models we tried, the **Random Forest classifier** gave the best results. It was able to find patterns in the data, even when the relationships were complex. For example, it noticed that patients with long waiting times or who didn't get a reminder were more likely to miss their appointment.

The model also told us which factors were most important in predicting no-shows. The top ones were:

- **Waiting days** between booking and the appointment
- **Receiving a reminder message**
- **Patient's age**
- **Day of the week** the appointment was scheduled

Using these insights, hospitals can improve how they manage appointments. They can:

- **Reduce wasted time** when patients don't show up

- **Fill empty slots** more effectively
- **Help other patients get appointments faster**
- **Make better use of doctors and staff**

In the future, we can make the models even better by adding more information — like patient history, weather conditions, or transportation availability. We could also use more advanced technologies, like deep learning or natural language processing, to understand appointment notes or reasons for visits.

In short, this project proves that machine learning is a powerful tool in healthcare. With the right data and smart use of technology, hospitals can make better decisions, **help more patients**, and **work more efficiently** — benefiting **both patients and healthcare workers**.

9. References

1. **Kaggle: Medical Appointment No-Shows Dataset** – Public dataset with over 100,000 medical records used for this analysis.
   https://www.kaggle.com/joniarroba/noshowappointments
2. **Scikit-learn Documentation** – Used for machine learning model development and evaluation.
   https://scikit-learn.org
3. **Seaborn and Matplotlib** – Python libraries used for data visualization and exploratory analysis.
   https://seaborn.pydata.org

   https://matplotlib.org
4. **Academic Journals** – Referenced for methods in predictive healthcare analytics and no-show behaviour modeling.
5. **World Health Organization (WHO)** – Provided global insights into appointment adherence trends.
   https://www.who.int
6. **Pandas and NumPy** – Used for data preprocessing and analysis tasks.
   https://pandas.pydata.org
   https://numpy.org