Literature Review, Data Description, and Project Approach

Emily Chan

Student ID:501216133

Supervisor: Dr. Ceni Babaoglu, PhD

Date of Submission: 13 October 2025

Table of Contents

[Revised Abstract 3](#_Toc211848074)

[Literature Review 3](#_Toc211848075)

[Introduction 3](#_Toc211848076)

[Machine Learning Techniques for No-Show Prediction 3](#_Toc211848077)

[Key Predictive Features 4](#_Toc211848078)

[Temporal Patterns and Operational Insights 4](#_Toc211848079)

[Gaps and Justification for Current Study 4](#_Toc211848080)

[Data Description 5](#_Toc211848081)

[Data Overview 5](#_Toc211848082)

[Descriptive Statistics 5](#_Toc211848083)

[Constraints and Limitations 5](#_Toc211848084)

[Preprocessing Steps 6](#_Toc211848085)

[Project Approach 6](#_Toc211848086)

[Theme and Problem Definition 6](#_Toc211848087)

[Research Questions 6](#_Toc211848088)

[Proposed Methodology 7](#_Toc211848089)

[Model Selection 7](#_Toc211848090)

[Feature Engineering 7](#_Toc211848091)

[Model Evaluation 7](#_Toc211848092)

[Tools and Environment 7](#_Toc211848093)

[Visual Representation 8](#_Toc211848094)

[GitHub Repository Link 8](#_Toc211848095)

[References 8](#_Toc211848096)

# Revised Abstract

Missed medical appointments disrupt healthcare operations, increase costs, and compromise patient outcomes. This capstone project explores the use of machine learning classification techniques to predict appointment attendance using a synthetic dataset that includes patient demographics, scheduling intervals, and time slot allocations. Three research questions guide the study:

1. What patient and appointment related features are most predictive of medical appointment no-shows?
2. Can a machine learning model accurately classify whether a patient will attend their appointment based on historical data?
3. Do appointment attendance rates vary by time of day or day of the week and how can scheduling be adjusted accordingly?

To address these questions, the project applies logistic regression, decision trees, and ensemble methods such as Random Forest and XGBoost. Model performance will be evaluated using metrics including accuracy, precision, recall, and ROC-AUC. By integrating temporal features and operational insights, the study aims to support data driven strategies and improve clinical efficiency.

# Literature Review

## Introduction

Missed medical appointments, or no-shows, continue to challenge healthcare systems worldwide by disrupting clinical workflows, reducing operational efficiency, and compromising patient outcomes. In response, machine learning (ML) has emerged as a promising tool for predicting and mitigating no-show behaviour. This literature review synthesizes recent research on ML-based appointment prediction, focusing on algorithmic approaches, key predictive features, temporal patterns, and operational insights. It also identifies gaps that this capstone project aims to address.

## Machine Learning Techniques for No-Show Prediction

A variety of ML algorithms have been applied to predict appointment attendance, including logistic regression, decision trees, support vector machines (SVM), and ensemble methods such as Random Forest and XGBoost. These models vary in complexity, interpretability, and performance.

* **Deina et al. (2024)** introduces Symbolic Regression and Instance Hardness Threshold to address class imbalance, outperforming traditional resampling techniques like SMOTE and NearMiss with sensitivity scores above 0.94.
* **Tuan et al. (2025)** developed a personalized ML framework incorporating geolinked clinical, socioeconomic, and climate data, achieving an AUC of 0.852. Their work highlights the value of contextual features in improving predictive accuracy.
* **Toffaha et al. (2024)** compared twelve classification algorithms and found that interpretable models such as FIGS and TAO Tree offered competitive performance with reduced runtime, making them suitable for real-time clinical applications.

These studies demonstrate the feasibility of ML for no-show prediction while revealing trade-offs between accuracy, interpretability, and computational efficiency.

## Key Predictive Features

Across studies, several features consistently emerge as strong predictors of no-show behaviour:

* **Patient age:** Younger patients tend to miss appointments more frequently
* **Scheduling lead time:** Longer intervals between booking and appointment increase no-show risk
* **Appointment type:** First-time visits have higher no-show rates than follow-ups
* **Attendance history:** Prior behaviour is a strong indication of future attendance

**Leviva-Araos et al. (2025)** emphasized the importance of semi-automated feature selection to streamline model development while maintaining predictive accuracy. Their framework integrated service capacity management and overbooking strategies, demonstrating the operational relevance of predictive modeling.

## Temporal Patterns and Operational Insights

Temporal factors such as time of day, day of the week, and seasonality are often underexplored in predictive models. Yet they hold significant potential for improving scheduling efficiency.

* **Russotto et al. (2025)** found that no-show rates varied significantly by appointment timing, with younger patients and longer wait times contributing to high absenteeism.
* **Oikonomidi et al. (2023)** conducted a systematic review showing that targeted reminders (e.g. text messages, phone calls) were most effective when timed around high-risk periods.

Despite these findings, few studies have directly integrated temporal scheduling data into predictive models. This project addresses that gap by leveraging the slots.csv file to extract time-based features.

## Gaps and Justification for Current Study

While prior research has demonstrated the feasibility of ML-based no-show prediction, several gaps remain:

* **Limited integration of temporal scheduling features** into predictive models.
* **Underutilization of synthetic datasets** for scalable experimentation and reproducibility.
* **Lack of operational insights** to inform scheduling strategies and resource allocation.

This capstone project addresses these gaps by incorporating temporal features (e.g. time slots, day of week) and evaluating multiple classification models using synthetic dataset designed for healthcare scheduling analysis. The inclusion of time-based data enables a more granular understanding of attendance behavior, which can inform practical interventions such as dynamic scheduling and targeted reminders.

# Data Description

## Data Overview

The dataset selected for this project is the [Medical Appointment Scheduling System dataset](https://www.kaggle.com/datasets/carogonzalezgaltier/medical-appointment-scheduling-system), publicly available on Kaggle. It is a synthetic dataset designed to simulate realistic clinical appointment data making it suitable for predictive modeling and classification tasks in healthcare operations. The dataset comprises three interrelated CSV files.

* **appointments.csv:** contains records of scheduled appointments, including patient ID, appointment date, scheduling interval, appointment duration, and attendance status (show/no-show). There are 111,488 rows with no duplicates.
* **patients.csv:** includes demographic information such as patient ID, sex, date of birth, and insurance provider. There are 36,697 rows with no duplicates.
* **slots.csv:** Provides available time slots for appointments, including date, time and availability status. There are 104,360 rows with no duplicates.

This structure enables the integration of demographic, temporal, and operational features, supporting a comprehensive analysis of appointment attendance behavior.

## Descriptive Statistics

Preliminary exploration of the dataset reveals the following characteristics:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Description | Summary Statistics | Notes |
| Patient Age | Age of patient | Mean:57, Range: 15-100 | Minimum age is 15, no patients 0-14 included |
| Scheduling Interval | Days between booking and appointment | Mean: 7, Range: 1-30 | Varies from 1 day in advance to several weeks |
| Attendance Status | Attended vs No-Show | Attended: 77%, No-Show 23% | Class imbalance favoring attended visits |
| Time of Day | Appointment time slot | Range: 8:00am - 5:45pm at 15-minute intervals | Well-distributed; supports temporal pattern analysis |
| Day of Week | Weekday of appointment | Range: 31 Dec 2014 – 1 Dec 2024 | Broad distribution; supports weekday pattern analysis |

These features are critical for feature engineering and model training, particularly in identifying predictors of no-show behaviour.

## Constraints and Limitations

While the dataset is rich and well-constructed, several constraints must be considered:

* **Synthetic Nature:** As a simulated dataset, it may not capture all nuances of real-world clinical behavior, such as socioeconomic or psychological factors.
* **Missing Contextual Variables:** Factors like weather, transportation access, or patient health status are not included, which may limit model generalizability.
* **Class Imbalance**: Preliminary analysis suggests a disproportionate number of attended appointments, which may require resampling techniques such as SMOTE or cost-sensitive learning.
* **Temporal Granularity:** Although time slots are included, the dataset does not specify appointment urgency or provider type, which could influence attendance.
* **Age Range Limitation:** The dataset excludes patients under the age of 15, which prevents analysis of pediatric appointment behaviour and may limit applicability to family or pediatric clinics.

These constraints should be carefully considered during model development and interpretation to ensure realistic and actionable insights.

## Preprocessing Steps

To prepare the dataset for modeling, the following preprocessing steps will be undertaken:

* **Data Cleaning:** Removal of duplicates, handling of missing values, and validation of data formats.
* **Feature Engineering:**
  + Derivation of scheduling intervals (days between booking and appointment)
  + Categorization of age into groups (e.g. pediatric, adult, senior)
  + Encoding of categorical variables (e.g. sex, insurance type)
* **Normalization:** Scaling of numerical features to ensure consistent input for ML algorithms.
* **Temporal Feature Extraction:** Identification of day of week and time of day patterns from the slots.csv file.

These steps will ensure the dataset is suitable for training robust classification models and for exploring temporal trends in appointment attendance.

# Project Approach

## Theme and Problem Definition

This project is grounded in the theme of predictive analytics and classification, with a specific focus on forecasting medical appointment attendance. The problem under investigation is the high rate of missed appointments in clinical settings, which leads to inefficient resource allocation and diminished quality of care. By leveraging machine learning techniques, the project aims to develop a predictive model that can classify whether a patient is likely to attend their scheduled appointment.

## Research Questions

3 research questions guide this project:

1. What patient and appointment related features are most predictive of medical appointment no-shows?
2. Can a machine learning model accurately classify whether a patient will attend their appointment based on historical data?
3. Do appointment attendance rates vary by time of day or day of the week, and how can scheduling be adjusted accordingly?

These questions are designed to address both individual-level predictors and system-level scheduling patterns, ensuring the project delivers actionable insights for healthcare operations.

## Proposed Methodology

To address the research questions, the project will implement a supervised learning framework using classification algorithms. The methodology includes the following components:

### Model Selection

* **Logistic Regression:** Serves as a baseline model for binary classification.
* **Decision Trees:** Offers interpretable rule-based predictions.
* **Random Forest:** An ensemble method that improves accuracy and handles feature interactions.
* **XGBoost:** A gradient boosting algorithm known for its performance and scalability in structured data tasks.

### Feature Engineering

* Derivation of scheduling intervals (days between booking and appointment)
* Categorization of age into groups (e.g. pediatric, adult, senior)
* Encoding categorical variables such as sex and insurance type
* Extraction of temporal features (e.g. day of week, time of day) from the slots.csv file.

### Model Evaluation

* Performance will be assessed using metrics such as **accuracy, precision, recall, F1-score,** and **ROC-AUC**.
* **Cross validation** will be applied to ensure robustness and generalizability.
* **Confusion matrices** will be analyzed to understand prediction errors and model behaviour.

## Tools and Environment

The following tools will be used to implement the methodology:

* **Python:** Primary programming language for data processing and modeling
* **Pandas and NumPy:** For data manipulation and cleaning
* **Scikit-learn and XGBoost:** For implementing machine learning models
* **Matplotlib and Seaborn:** for data visualization and exploratory analysis
* **Google Colab:** For cloud-based development and documentation
* **GitHub:** for version control and collaboration with the faculty supervisor

# Visual Representation

*A screenshot of a computer

AI-generated content may be incorrect.*

*Figure1. Project Methodology for Predicting Medical Appointment No-Shows*

# GitHub Repository Link

[CIND820 - Medical No Show Prediction](https://github.com/emchan4/CIND820-medical-no-show-prediction)

# References

Deina, C., Fogliatto, F. S., da Silveira, G. J. C., & Anzanello, M. J. (2024). Decision analysis framework for predicting no-shows to appointments using machine learning algorithms. BMC Health Services Research, 24, Article 37. https://doi.org/10.1186/s12913-023-10418-6

Tuan, W.-J., Yan, Y., Abou Al Ardat, B., Felix, T., & Chen, Q. (2025). Predicting missed appointments in primary care: A personalized machine learning approach. Annals of Family Medicine, 23(4), 294–301. https://doi.org/10.1370/afm.240316

Salazar, L. H. A., Parreira, W. D., Fernandes, A. M. R., & Leithardt, V. R. Q. (2022). No-show in medical appointments with machine learning techniques: A systematic literature review. Information, 13(11), 507. https://doi.org/10.3390/info13110507

Alshaya, S., McCarren, A., & Al-Rasheed, A. (2019). Predicting no-show medical appointments using machine learning. In Advances in Data Science, Cyber Security and IT Applications (pp. 211–223). Springer. https://doi.org/10.1007/978-3-030-36365-9\_18

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: How can we know it works? Journal of the American Medical Informatics Association, 26(12), 1651–1654. https://doi.org/10.1093/jamia/ocz130

Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: A survey. Journal of Electrical Systems and Information Technology, 10, Article 40. https://doi.org/10.1186/s43067-023-00108-y