Emily Chan

Student Number: 501216133

Supervisor: TBD

Submission Date: 22 Sept 2025

Table of Contents

[Problem Statement and Context 3](#_Toc209426769)

[Research Questions and Justification 3](#_Toc209426770)

[Research Question 1 3](#_Toc209426771)

[Research Question 2 3](#_Toc209426772)

[Research Question 3 4](#_Toc209426773)

[Dataset Selection and Rationale 4](#_Toc209426774)

[Proposed Methodology and Tools 5](#_Toc209426775)

[Techniques 5](#_Toc209426776)

[Tools 5](#_Toc209426777)

[References 6](#_Toc209426778)

## Problem Statement and Context

Missed medical appointments are a persistent challenge in the healthcare system worldwide. They lead to inefficient use of resources, increased operational costs and delayed care for patients. Clinics struggle to anticipate no-shows, disrupting scheduling and lowering service quality. Understanding the behavioural patterns behind appointment attendance is crucial for improving healthcare access and efficiency.

The theme of predictive analytics and classification offers a powerful framework to address this issue. By analyzing historical appointment data and patient demographics, we can build models that predict whether a patient is likely to attend their scheduled appointment. This approach enables proactive interventions, such as reminders or rescheduling strategies, to reduce no-show rates.

The dataset used in this project is called Medical Appointment Scheduling System and it includes detailed records of appointments, patient profiles and time slots. Prior research has shown that factors such as age, gender, waiting time and scheduling intervals can influence attendance behaviour. For example, studies have found that longer wait times between scheduling appointment dates are associated with high no-show rates (Dantas et al., 2018; Oikonomidi et al., 2023). By applying classification techniques such as logistic regression, decision trees and ensemble models, this project aims to replicate and extend these findings in a new context.

This issue is directly relevant to healthcare operations and aligns with the predictive analytics theme. The insights generated can support data driven decision making in clinical settings and contribute to more efficient appointment management systems.

## Research Questions and Justification

### Research Question 1

What patient and appointment related features are most predictive of medical appointment no-shows?

This question seeks to identify the key variables such as age, sex, scheduling interval and waiting time that influence whether a patient attends their appointment. Understanding these predictors is essential for building effective classification models and guiding operational decisions.

### Research Question 2

Can a machine learning model accurately classify whether a patient will attend their appointment based on historical data?

This question builds on the first by applying the identified features to a classification task. It requires the use of supervised learning algorithms to predict binary outcomes. The question is relevant because it tests the feasibility of deploying predictive models in real-world healthcare settings. It also allows for performance evaluation using metrics such as accuracy, precision and recall which are essential for validating the model’s utility.

### Research Question 3

Do appointment attendance rates vary by time of day or day of the week and how can scheduling be adjusted accordingly?

This question adds temporal dimension, examining whether specific time slots or days correlate with higher no-show rates. Understanding these patterns can help clinics optimize their scheduling practices by allocating high risk slots to patients with a lower likelihood of missing appointments or by implementing targeted interventions during vulnerable periods. This question complements the predictive modeling approach by adding operation insights that can improve resource utilization and service delivery.

Together, these questions form a cohesive framework for analyzing appointment attendance behaviour using predictive analytics. They address both individual-level predictors and system-level scheduling patterns, ensuring the project delivers actionable insights for healthcare operations.

## Dataset Selection and Rationale

The dataset selected for this project is the [Medical Appointment Scheduling System dataset](https://www.kaggle.com/datasets/carogonzalezgaltier/medical-appointment-scheduling-system), available publicly on Kaggle. It is a synthetic dataset designed to simulate realistic clinical appointment data. The dataset includes 3 interrelated CSV files:

* **appointments.csv**: Contains detailed records of scheduled appointments, including attendance status, appointment date, scheduling interval, and appointment durations.
* **patients.csv**: Includes demographic information such as patient ID, sex, date of birth and insurance provider.
* **slots.csv**: Provides the available time slots for appointments, including date, time and availability status.

The inclusion of the time-based slots.csv file enables analysis of temporal patterns, such as variations in attendance by time of day or day of the week. This supports the third research question and adds an operational layer to the predictive modeling framework.

This dataset was selected for its relevance, richness and feasibility for predictive analytics and classification tasks. It offers a comprehensive view of patient behaviour and appointment logistics, enabling the development of models to predict attendance outcomes. Similar predictive modeling frameworks have been successfully applied to primary care datasets, demonstrating the feasibility and relevance of this approach (Leiva-Araos et al., 2025). The inclusion of both temporal and demographic features supports robust feature engineering and model training.

The dataset is well structured and manageable in the 14 week project time line. It aligns with the course’s emphasis on applying machine learning techniques to practical problems and supports the development of a classification model with measurable performance metrics.

Potential limitations include:

* The absence of certain behavioural or contextual variables (e.g. Weather, transportation access) that may influence attendance.
* Possible class imbalance between attended and missed appointments, which may require resampling techniques.
* Limited geographic or institutional scope, which may affect generalizability.

Despite these limitations, the dataset is highly suitable for the project’s objectives and provides a strong foundation for meaningful analysis and model development.

The dataset’s structure makes it ideal for a capstone project, synthetic yet realistic, and manageable for iterative development within the course timeline.

## Proposed Methodology and Tools

To address the research questions and solve the problem of predicting medical appointment attendance, this project will apply a combination of classification techniques, feature engineering and model evaluation strategies using industry standard tools.

### Techniques

* **Classification Algorithms**: The project will implement and compare multiple classification models including:
  + **Logistic Regression:** A baseline model for binary classification.
  + **Decision Trees**: For interpretable rule-based predictions.
  + **Random Forest & XGBoost:** Ensemble methods to improve accuracy and handle feature interactions.
* **Feature Engineering**:
  + Derive new features such as scheduling intervals, waiting times and age groups.
  + Encode categorical variables (e.g. Insurance type and sex)
  + Handle missing values and normalize numerical features.
* **Model Evaluation:**
  + Use metrics such as accuracy, precision, recall, F1-score and ROC-AUC to assess model performance.
  + Apply cross validation to ensure robustness
  + Analyze confusion matrices to understand prediction errors

### Tools

* **Python**: Primary programming language for data processing and modeling.
* **Pandas and NumPy:** For data manipulation and cleaning
* **Scikit-learn & XGBoost:** For implementing machine learning models.
* **Matplotlib & Seaborn:** For data visualization and exploratory analysis.
* **Google Colab:** For documenting the workflow and sharing results in a cloud-based environment.
* **GitHub:** For version control and collaboration with faculty supervisor.

These tools offer flexibility, scalability, and reproducibility, ideal for a capstone level predictive analytics project. Techniques such as ensemble models and resampling have shown strong performance in predicting no-shows (Deina et al., 2024). Additionally, the methodology supports temporal analysis to explore patterns in appointment attendance across different times and days, directly addressing the third research question. This approach ensures predictive accuracy while informing practical scheduling improvements.

Overall, the proposed methodology integrates predictive modeling with operational insights, ensuring that the project not only identifies key attendance predictors but also informs practical scheduling improvements in clinical settings.

## References

Booth, W. C., Colomb, G. G., & Williams, J. M. (2016). The craft of research (4th ed.). University of Chicago Press.

Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). SAGE Publications.

Dantas, L. F., Fleck, J. L., Oliveira, F. L. C., & Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. Health Policy, 122(4), 412–421. https://doi.org/10.1016/j.healthpol.2018.01.002

Jansen, D. (2023). What is a research question? GradCoach. https://gradcoach.com/what-is-a-research-question/

Liu, X. (2017). Definition of research topic. In The Sage encyclopedia of communication research methods (Vol. 4, pp. 1463–1465). SAGE Publications. https://doi.org/10.4135/9781483381411

University of Southern California. (n.d.). Organizing your social sciences research paper: Research designs. USC Libraries. https://libguides.usc.edu/writingguide/researchdesigns

Oikonomidi, T., Norman, G., McGarrigle, L., Stokes, J., van der Veer, S. N., & Dowding, D. (2023). Predictive model-based interventions to reduce outpatient no-shows: a rapid systematic review. Journal of the American Medical Informatics Association, 30(3), 559–569. https://doi.org/10.1093/jamia/ocac242

Leiva-Araos, A., Contreras, C., Kaushal, H., & Prodanoff, Z. (2025). Predictive optimization of patient no-show management in primary healthcare using machine learning. Journal of Medical Systems, 49(7). https://doi.org/10.1007/s10916-025-02143-w

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