Supporting Document

Daily Activity Forecasting – Methodology & Challenges

---

1. Project Objective

The objective of this assignment is to forecast **daily step counts for the next 365 days** using wearable activity data, while incorporating available clinical information in an **interpretable and healthcare-appropriate** manner.
The solution emphasizes:

- Clean time-series preprocessing
- Transparent modeling choices
- Explainability over black-box accuracy

---

2. Data Description

2.1 Time-Series Data (File A)

- High-frequency wearable step count records
- Fields include timestamps (start, end) and step count
- Data is irregular and event-based

2.2 Clinical Data (File B)

- Patient-level categorical attributes such as:
  - Age (derived from birth year)
  - Smoking status
  - Therapy information
- Clinical records are largely static and lack reliable event-level timestamps

---

3. Data Preprocessing & Aggregation

3.1 Timestamp Handling

- All timestamps were converted to Python datetime objects
- Timezone was standardized to **UTC**
- Records were sorted chronologically

3.2 Daily Aggregation

- Raw step events were aggregated into **daily step counts**
- A **continuous daily index** was created
- Missing days were explicitly filled with zero activity to avoid temporal gaps

This produced a clean, gap-free daily time series suitable for forecasting.

---

4. Feature Engineering & Clinical Fusion

4.1 Time-Series Features

- Lag feature:
  - steps_t-1 (previous day's step count)
- These features capture short-term temporal dependency in activity behavior

4.2 Clinical Feature Integration

Clinical attributes were integrated as **static daily features**, including:

- Age
- Smoking status (is_smoker)
- Therapy exposure (is_on_therapy)

Each daily record contains the same clinical context for the patient.

---

5. Modeling Approach

5.1 Model 1 — Baseline Time-Series Model

- **Prophet** was used as a univariate baseline
- Inputs:
  - Date (ds)
  - Daily step count (y)
- Prophet captures:
  - Trend
  - Weekly and yearly seasonality
- Performance was evaluated using:
  - RMSE
  - MAE
- A temporal hold-out window was used for validation

5.2 Model 2 — Multivariate Explainable ML Model

- **Explainable Boosting Machine (EBM)** from interpretml
- Inputs:
  - Lagged step feature
  - Clinical features (age, smoking status, therapy)
- Temporal splitting was used to avoid data leakage
- EBM was selected for:
  - Non-linear modeling capability
  - Built-in explainability suitable for healthcare use

---

6. Evaluation & Comparison

- Both models were evaluated on unseen data
- Metrics used:
  - RMSE
  - MAE
- The multivariate EBM model showed **lower error** compared to the univariate baseline, demonstrating the value of incorporating clinical context.

---

7. Explainability

- Global feature importance from EBM was analyzed
- Key insights:
  - Previous day activity is the strongest predictor
  - Age and therapy exposure influence activity levels
  - Smoking status shows a smaller but consistent effect
- These effects align with intuitive clinical and behavioral expectations

---

8. Forecast Output

The final output is a **365-day daily forecast** with the following schema:

- Date
- Predicted_Steps
- Trend_Component
- Exogenous_Impact

This format ensures both usability and interpretability of the forecast.

---

9. Challenges & Design Trade-offs

The primary challenge in this assignment was the **absence of reliable event-level timestamps** in the provided clinical data (e.g., therapies, diagnoses, and side effects).
To avoid incorrect temporal assumptions and data leakage, clinical features were integrated as **static patient context** rather than dynamic event-based variables.

Additionally, the limited duration of historical step data constrained the complexity of temporal feature engineering. Model choices and validation strategies were therefore designed to prioritize **robustness, interpretability, and reproducibility** over aggressive optimization.

---

10. Conclusion

This project demonstrates an end-to-end, explainable time-series forecasting pipeline for healthcare activity data.
The solution balances predictive performance with transparency, making it suitable for real-world healthcare applications where interpretability and data integrity are critical.