The Adherence-AI Project: A Research Journal

Lead Researcher: Nanda | **Project Start Date:** August 2025 | **Status:** Completed

1. Introduction & Background

The Problem: Medication non-adherence is one of the silent, persistent challenges in modern healthcare. For patients managing chronic conditions like anemia, failing to follow a prescribed regimen can lead to severe health complications, hospital readmissions, and a diminished quality of life. The core issue is that non-adherence is deeply personal and multi-faceted; it's driven by a complex web of clinical, social, and psychological factors that are difficult to untangle.

Our Motivation: This project was born from a simple but powerful idea: what if we could use machine learning to look beyond the surface-level clinical data and understand the *human story* behind non-adherence? Our goal was to build a tool that could not only predict which patients were at risk but also provide actionable insights to help healthcare providers intervene effectively and compassionately. This journal documents our journey—our steps, our experiments, our stumbles, and our ultimate success.

2. Project Objectives

Our mission was guided by a clear set of goals:

1. **Develop a Predictive Model:** Build a robust machine learning model to accurately classify patients as either 'Adherent' or 'Non-Adherent'.
2. **Achieve Realistic Performance:** Move beyond simplistic metrics to establish a reliable, real-world accuracy baseline.
3. **Identify Key Predictive Factors:** Uncover the most influential drivers of medication adherence within our dataset.
4. **Generate Actionable Insights:** Translate the model's findings into practical strategies that could be implemented in a clinical setting.

3. Methodology & Steps Taken: The Project Log

This section details our end-to-end research process, documenting the experiments and decisions made along the way.

Phase 1: Data Exploration and Cleaning

Our journey began with a dataset of 500 anonymized anemia patients. The data was rich but imperfect, containing missing values across several columns. Our first task was foundational: create a clean, complete dataset. We implemented a standard imputation strategy, filling missing numeric values with the **median** and categorical values with the **mode**. This ensured that our subsequent models would have a solid foundation to learn from.

Phase 2: Baseline Modeling and the "69% Wall"

With clean data in hand, we trained our first set of models. We chose two powerful, well-regarded algorithms: **Random Forest** and **XGBoost**. Both models performed admirably, but they quickly converged on a similar performance, hitting a consistent accuracy of around **69%**.

Observation: Even after extensive hyperparameter tuning using GridSearchCV, neither model could break past this performance ceiling. This was a critical finding. It told us that the limitation wasn't in our models' ability to learn, but in the information contained within the raw features themselves. The signal was clear, but it wasn't strong enough. To do better, we needed better features.

Phase 3: The Feature Engineering Experiments

This was the most crucial phase of the project, involving two distinct experiments.

- Experiment A: General Feature Engineering (The False Start)
  Our first attempt was to create logical but general features. We engineered a frailty_index (age * comorbidities) and a financial_burden score. The hypothesis was that these composite features would provide a stronger signal.
  - **Result:** Failure. The model's performance *decreased* to ~63-68%.
  - **Lesson Learned:** We concluded that these features, while logical, might have been adding more noise than signal. They were too generic and didn't capture the specific nuances of adherence behavior. This was a valuable lesson in humility; a good idea on paper doesn't always translate to better performance.
- Experiment B: Targeted Feature Engineering (The Breakthrough)
  We went back to our baseline model's feature importance list. The top predictors were consistently health_literacy_score, provider_consistency, social_support_index, belief_in_medication, and income_bracket. This sparked our breakthrough hypothesis: the relationships between these key features are more important than the features themselves.
  This led to the creation of two new, highly targeted features:

1. **patient_readiness_score:** A composite score that unified a patient's psycho-social state by combining health literacy, social support, belief in the medication, and provider consistency.
2. **literacy_x_income:** An interaction term that explicitly captured the compounded risk faced by patients with both low health literacy and low income.

Phase 4: Final Model Selection

We re-ran our Random Forest and XGBoost models on the dataset enhanced with these new, targeted features. The results were immediate and conclusive. Both models saw a significant performance jump, validating our hypothesis. We selected the **Random Forest** as our champion model, as it achieved identical performance to XGBoost while being simpler and more interpretable.

4. Results and Observations

The journey of experimentation yielded a clear winner. The targeted feature engineering was the key that unlocked the next level of performance.

Model Performance Comparison:

Model Version	Accuracy	Recall (Non-Adherent)	Key Takeaway
Baseline (Tuned RF)	69%	0.61	Solid, but hit a performance ceiling.
General Features (RF)	68%	0.54	Worse performance; features added noise.
Targeted Features (RF)	**72%**	**0.61**	**Winner! Higher accuracy and balanced performance.**

Final Model Classification Report (Random Forest with Targeted Features):

	precision	recall	f1-score	support
0	0.74	0.61	0.67	46
1	0.71	0.81	0.76	54
accuracy			0.72	100

5. Model in Action: Predictive Test Cases

To demonstrate the model's intelligence, we tested it on three hypothetical patient profiles:

Name	Prediction (1=Adherent)	Confidence (Adherent)	Confidence (Non-Adherent)
Saanvi (High Risk)	0	23.5%	76.5%
Rohan (Low Risk)	1	77.6%	22.4%
Priya (Ambiguous)	1	56.4%	43.6%

Analysis:
The model correctly identified the high-risk and low-risk cases with high confidence. Most impressively, it classified the ambiguous case, Priya, as "Adherent" but with low confidence. This nuance is critical, as it flags her as a "borderline" patient who, despite a positive prediction, requires extra monitoring and support from her healthcare team.

6. Discussion

This project was a powerful lesson in the art and science of feature engineering. Our most significant finding was that **not all features are created equal**. Our initial, generic features failed because they were based on general assumptions. In contrast, our final, successful features were born from a **data-driven hypothesis**. By listening to what our baseline model told us was important, we were able to craft features that amplified the existing signals in the data.

The journey reaffirmed a core principle of data science: a deep understanding of the problem and an iterative, experimental approach will always outperform a brute-force application of complex algorithms.

7. Conclusion

We successfully developed a machine learning model that predicts medication adherence with **72% accuracy**. The project achieved all its objectives, culminating in a tool that is not only predictive but also deeply insightful. The final model confirms that medication adherence is a profoundly human issue, where factors like a patient's understanding, trust, and socioeconomic context are paramount.

Recommendations for Future Work:

- **Deployment:** The model could be integrated into a clinical dashboard to provide real-time risk scores for patients.
- **Refinement:** Incorporating more data, such as records of past appointment attendance or pharmacy refill data, could further enhance the model's accuracy.
- **Intervention Studies:** Use the model's predictions to run a pilot study, offering targeted educational and social support to high-risk patients to measure the real-world impact on adherence rates.

8. Tools, Tech Stack, and Datasets

- **Technology:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn
- **Dataset:** The project utilized a synthetic but realistic dataset of 500 anonymized anemia patients, sourced from https://raw.githubusercontent.com/nandarishik/Ferry-Internship/main/realistic_medication_adherence_data.csv.