# AI Project Milestone Report

**Course:** CS 4100: Artificial Intelligence (Fall 2025) – Practical Track

## Team Members

- Luke Christenson

- Ipek Ozcan

## 1. Current Project Status

We have made significant progress on our fitness data analysis project, which aims to analyze relationships between workout performance and lifestyle/health factors. We have completed the initial data preparation, exploratory data analysis, and established baseline models. Key accomplishments include:

### Completed Tasks

- Dataset acquisition and cleaning for the Workout & Fitness Tracker dataset
- Comprehensive exploratory data analysis of all key variables
- Implementation of data preprocessing pipeline with robust error handling
- Feature engineering including derived metrics (BMI, workout efficiency)
- Initial baseline models for classification and regression approaches
- Visualization framework for data exploration and analysis results

### In Progress

- Enhancement of correlation analysis visualizations
- Development of advanced feature relationship heatmaps
- Implementation of lifestyle impact analysis framework
- Improvement of model evaluation metrics and visualizations

## 2. Progress Analysis

### Data Preparation

We successfully implemented data loading and preprocessing pipelines that handle the complexities of the fitness dataset. We've addressed challenges such as:

- Column name variations in the dataset
- Appropriate handling of missing values
- Encoding of categorical variables for modeling
- Creation of derived features (workout efficiency score, BMI categories)

**Initial Analysis**

Our exploratory data analysis revealed several promising findings:

- Strong correlations between sleep duration and workout efficiency
- Notable influence of hydration levels on performance metrics
- Clear patterns in how resting heart rate affects workout outcomes
- Significant variation in calorie burn efficiency across workout types

**Baseline Models**

We have implemented two types of models:

1. **Classification Model**: Predicts workout efficiency categories (Low, Medium, High)
   - Logistic Regression baseline: ~85% accuracy
   - Random Forest implementation: ~92% accuracy
2. **Regression Model**: Predicts calories burned during workouts
   - Linear Regression baseline: $R^2$ of ~0.68
   - Random Forest implementation: $R^2$ of ~0.82

These results provide a solid foundation for further refinement and optimization.

## 3. Challenges Encountered

1. **XGBoost Integration**: Initial attempts to incorporate XGBoost led to dependency conflicts. We addressed this by making XGBoost optional and implementing fallback models.

2. **Correlation Visualization**: Some correlation matrices contained NaN values which affected visualization quality. We're currently implementing improved handling of these cases.

3. **Feature Importance Interpretation**: Understanding the relative importance of lifestyle factors requires more sophisticated analysis than initially anticipated. We're developing enhanced visualization techniques to address this.

## 4. Next Steps

For the remainder of the project, we plan to:

1. **Enhance Visualization Framework**:
   - Improve correlation matrices with significance highlighting
   - Develop feature relationship heatmaps with statistical annotations
   - Create detailed lifestyle factor impact visualizations
2. **Advanced Analysis**:
   - Implement comprehensive lifestyle factor analysis component
   - Develop quartile analysis to detect non-linear relationships
   - Create composite lifestyle scoring for holistic analysis

3. **Model Refinement**:
   - Optimize hyperparameters for classification and regression models
   - Enhance evaluation metrics with detailed performance breakdowns
   - Improve interpretability of model outputs
4. **Documentation**:
   - Create comprehensive final report with detailed findings
   - Develop actionable recommendations based on lifestyle impact analysis
   - Prepare project presentation with key visualizations

## 5. Updated Timeline

- **Weeks 5-6** (Current): Complete enhancement of visualization framework and lifestyle analysis
- **Week 7**: Finalize model refinement and comprehensive evaluation
- **Week 8**: Complete documentation, final report, and presentation preparation

## 6. Conclusion

The project is progressing well, with key components of data preparation, exploratory analysis, and baseline modeling already in place. We have identified promising relationships between lifestyle factors and workout performance that warrant further investigation. The next phase will focus on enhancing our analysis framework, improving visualizations, and refining models to provide deeper insights into how lifestyle and health factors influence fitness outcomes.