# **AAI-540 ML Design Document Template**

## **Team Info**

Project Team Group #: 1

Authors: [Peter Ogunrinde](mailto:pogunrinde@sandiego.edu), [Christopher Alleyne](mailto:calleyne@sandiego.edu), Jory Hamilton

Business Name: HealthTrack AI

Publication Date: 9/15/2025

## **Team Workflows**

GitHub Project Link: <https://github.com/pogunrinde/AAI-540_FinalProject>

Asana Board Link: <https://app.asana.com/1/952672460738672/project/1211354276436627/list/1211354282389281>

Team Tracker Link: <https://docs.google.com/document/d/1x5c1JYJv4ZOtk6_y2sbXwwClwE8IJLRs/edit?usp=sharing&ouid=116064852923920598003&rtpof=true&sd=true>

## **Project Scope**

**Project Background: Maternal Health Risk (Rural Medicine)**

For this project, we propose using the **Maternal Health Risk dataset**, which includes health indicators from women in rural areas of Bangladesh (and comparable populations). The dataset classified maternal health outcomes into risk categories.

**Model Objective:** The objective is to **predict maternal health risk levels (low, mid, high)** based on patient health indicators, with the goal of supporting doctors in remote or underserved regions in making faster, more accurate risk assessments.

**Type of ML Problem:** This is a **supervised multi-class classification problem**, where the target variable is categorical (maternal health risk level).

**Technical Background:**

**Evaluation:** The model will be evaluated with metrics suitable for multi-class classification, including **accuracy**, **precision, recall, and F1-score (macro-averaged)**, as well as a **confusion matrix** to analyze misclassifications. From a clinical perspective, false negatives (predicting low risk when the patient is actually high risk) are the most critical errors to minimize.

**Data Source:** The **Maternal Health Risk dataset** is sourced from the UCI Machine Learning Repository. It contains medical records such as age, blood pressure, heart rate, and body mass index.

**Data Preparation:** Preparation steps will include handling missing values (if any), normalizing numerical variables (e.g., blood pressure, BMI), encoding categorical features, and possibly balancing classes if certain risk categories are underrepresented.

**Data Exploration:** Exploratory Data Analysis (EDA) will examine variable distributions, correlations (e.g., between age and systolic blood pressure), and class imbalances.

**Main Features (Hypothesized):** Age, systolic blood pressure, diastolic blood pressure, blood sugar, body temperature, and heart rate.

**Model Choice:** I plan to start with interpretable models such as **Decision Trees** and **Logistic Regression** for baseline understanding, and then compare to ensemble methods like **Random Forests** or **Gradient Boosted Trees (XGBoost/LightGBM)** for improved accuracy.

**Goals vs Non-Goals:**

**Goals:**

* Build a predictive model to classify maternal health risks (low, mid, high).
* Ensure the model is interpretable for healthcare practitioners.
* Implement monitoring for accuracy and class distribution drift.
* Document a reproducible ML pipeline.

**Non-Goals:**

* Replacing clinical decision-making or providing final diagnoses.
* Integrating with hospital EHR systems in this project scope.
* Full hyperparameter optimization beyond baseline models.
* Developing a mobile application interface (future scope).

### **Solution Overview**

The ML system will use AWS SageMaker pipelines for preprocessing, training, and deployment. Predictions can be integrated into a dashboard for healthcare workers or exposed via an API for mobile health applications.

Architecture Components:

Data Storage: AWS S3 (raw + processed).

Preprocessing & Feature Engineering: SageMaker Processing jobs.

Model Training: SageMaker Training jobs (Random Forest / XGBoost).

Model Registry: SageMaker Model Registry.

Deployment: Batch inference endpoints.

Monitoring: Data drift detection (SageMaker Model Monitor).

CI/CD: GitHub --> AWS CodePipeline --> SageMaker.

**Data Sources:**

What is your data source? Kaggle Maternal Health Risk Data Set

What is your data volume? Over 11,000 records, labeled for 3 classes (low, mid, high).

Why did you select this data set? Public, labeled dataset addressing a critical health issue.

Any risks (bias, sensitive features, etc)? Relatively small dataset --> may require cross-validation, data augmentation (SMOTE).

**Data Engineering:**

How will you store this data? AWS S3 (raw, processed, curated).

What data pre-processing do you need to do before you feed it into your ML system? Missing value handling, normalization of continuous variables, SMOTE oversampling if imbalance is detected.

**Training Data:**

How will you split your data into training, test and validation? 70% training, 15% testing, 15% validation

Will you use any data labeling techniques? Already provided in the dataset.

**Feature Engineering:**

Which fields from your data will you use or exclude? Diastolic BP, blood sugar, body temperature, heart rate, Age.

Which fields will be combined or bucketed? None

What other data transformations will you apply to your data? Standardization/normalization

**Model Training & Evaluation:**

How will you train your model? Preprocessing, validation & continuous pipeline.

What algorithm will you use? Logistic Regression, Random Forest, XGBoost.

What parameters will you use? Hyperparameter tuning via SageMaker HPO.

How will you evaluate your model? Recall, Precision, Accuracy, F1-score, ROC-AUC. Additionally, these metrics will help us identify overfitting (e.g., too many epochs, model memorizing the training data).

**Model Deployment:**

What instance size will you use? ml.m5.large

Will your model function as a batch or real time model? Why? Batch inference

**Model Monitoring:**

How will you monitor your model? Monitoring accuracy/F1 on validation data

How will you monitor your infrastructure? AWS CloudWatch (CPU, memory, endpoint latency)

How will you monitor your data? Model Monitor for input/output schema drift.

**Model CI/CD:**

What checkpoints will your CI/CD pipeline contain? Data validation --> preprocessing --> model training --> evaluation --> registry --> deployment.

What tests will your CI/CD pipeline contain? Unit tests (data prep), integration tests (pipeline), acceptance tests (performance thresholds).

**Security Checklist, Privacy and Other Risks:**

Will this store or process Personal Health Information (PHI)? No direct identifiers

Will this store or process Personal Identifiable Information (PII)? No direct identifiers

Will user behavior be tracked and stored? No direct identifiers

Will this store or process credit card information? None

If you answered yes to any of the above questions, please justify.

What S3 buckets will this application read from or write to? Encrypted, role based IAM access.

What data bias should be considered? Age group, blood pressure/sugar, BMI, body temperature and heart rate.

Will your model have potential for bias along sensitive features (race, ethnicity, gender, age, religion, disability, sexual orientation, or other personal attributes): Age, ethnicity, basic medical chart.

Are there any ethical concerts with the data or business problems that should be addressed? None

**Future Enhancements:**

Provide at least 3 ways you would improve your ML system if you had more time or additional resources.

* Further gather data to expand the original dataset
* Incorporating datasets specific to the U.S. for deployment more specific for our region
* Cross-referencing datasets that investigate other aspects that ours does not (e.g., social data, as in are there health outcomes related to social groups within a population).