Baseline Model Report

Project: Predicting Heart Risks — Stay Healthy, Stay Ahead

Date: 22 June 2025

Authors: Melina Lellepi and Ioannis Papadikos

1. Objective 🎯

This project builds and evaluates simple baseline models to predict whether an individual is at risk of a heart attack. We used personal health and lifestyle data (e.g., BMI, cholesterol, diet, sex, etc.) to make data-driven predictions.

2. Data Preparation 🖌️

2.1 Dataset Info 📂

- Filename: heart_attack_prediction_dataset.csv

- Target column: Heart Attack Risk (0 = Not at risk, 1 = At risk)

2.2 Cleaning & Encoding 🧼

Dropped irrelevant columns: Patient ID, Country, Continent, Hemisphere, and Blood Pressure.

Categorical features (Sex and Diet) were one-hot encoded using drop='first' to avoid multicollinearity.

The feature set was scaled using StandardScaler, and data was split into 80% training and 20% testing sets.

3. Models Used 🤖

1. Dummy Classifier
Baseline model that always predicts the most frequent class. Used to benchmark real model performance.

2. Logistic Regression
A fast and interpretable linear model suitable for binary classification.

3. k-Nearest Neighbors (k-NN)
A non-parametric model that classifies based on the majority class among k nearest neighbors.

4. Decision Tree Classifier
A model that makes decisions based on feature thresholds; provides explainability and rule-based classification.

4. Evaluation Metrics 📏

We evaluated each model using:

- Accuracy

- Precision

- Recall

- F1 Score

- Mean Absolute Error (MAE)

- Mean Squared Error (MSE)

- Root Mean Squared Error (RMSE)

- Confusion Matrix

- Log Loss (Binary Cross-Entropy)

- Gini Impurity (for Decision Tree)

5. Results 📊

Model	Accuracy	Precision	Recall	F1 Score	Log Loss	Notes
Dummy Classifier	0.642	0.00	0.00	0.00	—	Predicts all as majority class (No Risk)
Logistic Regression	0.642	0.00	0.00	0.00	0.655	Fails to identify any positive cases
k-NN Classifier	0.569	0.350	0.239	0.284	—	Better than dummy, identifies positives
Decision Tree	0.641	0.444	0.013	0.025	0.653	Very low recall, overfits on majority

6. Confusion Matrices 🔍

Model	True Negative	False Positive	False Negative	True Positive
Dummy	1125	0	628	0
Logistic	1125	0	628	0
k-NN	847	278	478	150
Decision Tree	1115	10	620	8

7. Insights & What We Learned 🧠

- The Dummy model shows baseline performance (accuracy = 64%), but zero ability to detect heart attack risk.

- Logistic Regression matched dummy accuracy but also did not classify any risky cases.

- k-NN managed to capture 150 true positives with an F1 score of 0.284 — showing predictive power.

- Decision Tree was interpretable but struggled with recall; it found only 8 risky cases.

- Visualizations (heat-maps, tree structure) were helpful in understanding model performance and logic.

8. Next Steps 🚀

✅ Evaluate performance beyond accuracy — focus on recall and F1

✅ Add decision tree interpretability (plot_tree)

✅ Use log loss to measure prediction confidence

✅ Visualize confusion matrices and metrics across models

Future work:

- Hyper-parameter tuning (e.g., k in k-NN, max_depth in trees)

- Try more advanced models (Random Forest, XGBoost)

- Handle class imbalance using SMOTE or class weighting

9. Final Note 💡

This was our starting point — and it gave us clarity on what works and what does not. We learned that even simple models can struggle when faced with class imbalance and low signal data. But this gives us a clear direction: improve recall, boost precision, and balance our classes.