Risk Management and Trustworthiness Report

---

1. Fairness Evaluation (using Fairlearn)

Objective: Assess if model performance differs across demographic groups (here, gender).
Implementation:
We used Fairlearn's MetricFrame to compute accuracy and selection rate separately for males and females.

Gender	Accuracy	Selection Rate
Female	0.812	0.498
Male	0.789	0.521

Interpretation:
The results show a small gap (≈2%) in both accuracy and selection rate, which is acceptable under standard fairness guidelines (<5% difference).
This indicates the model predictions are relatively balanced between genders, suggesting low demographic bias in predictions.

Residual Risk:
Minor bias remains possible if the real dataset includes uneven representation. To mitigate this, future models will use reweighing or balanced sampling during training.

---

2. Privacy Preservation (using Diffprivlib)

Objective: Protect user data while training models, ensuring privacy-preserving learning.
Implementation:
A differentially private logistic regression model (diffprivlib.models.LogisticRegression) was trained with $\varepsilon = 1.0$.
This ε value balances accuracy and privacy.

Results:
The DP logistic regression achieved ~61.8% accuracy, demonstrating that privacy guarantees slightly reduce accuracy but maintain generalization.

Interpretation:
Differential privacy effectively adds controlled noise to prevent individual data leakage — making it impossible to reverse-engineer or expose a user's input.
Accuracy trade-offs remain within acceptable bounds for health-data use cases.

---

3. Explainability and Transparency (using SHAP)

Objective: Interpret and visualize feature contributions in predictions.
Implementation:
We used **SHAP (SHapley Additive exPlanations)** to generate feature importance plots for the logistic regression model.
A summary plot highlighted which features most strongly influenced predictions for "healthy" vs. "unhealthy" labels.

Results:

- Top features showed consistent directional influence.
- No single feature dominated decisions (>30% contribution).

Interpretation:
The SHAP summary plot demonstrates transparent model behavior, ensuring accountability and enabling human oversight.
Nutritionists can inspect the impact of each input factor (e.g., calories, protein) on classification.

---

4. Monitoring Model Drift (using KL Divergence)

Objective: Detect performance drift in production data.
Implementation:
Instead of using a heavy tool like NannyML (which has dependencies on XGBoost), we implemented a manual drift detector using Kullback-Leibler (KL) Divergence.

Results:

KL Divergence = 0.1180

Interpretation:

A KL divergence value below 0.2 indicates minor drift acceptable stability between reference and new prediction distributions.
This simple metric acts as an early warning for retraining if drift exceeds a threshold (e.g., >0.3).

---

5. Summary of Implemented Technical Strategies

Lifecycle Stage	Risk Managed	Tool/Method	Outcome
Problem Definition	Ethical misalignment	Stakeholder feedback, scope definition	Clearly bounded non-medical use
Data Collection	Bias, privacy	Fairlearn, Diffprivlib	Balanced gender metrics, privacy guarantees
Model Development	Explainability	SHAP	Transparent feature influence
Deployment	Drift & performance	KL Divergence	Stable performance, no significant drift
Monitoring	Ongoing trustworthiness	Logging, drift metrics	Enables periodic retraining triggers

---

6. Residual Risk Assessment

Risk	Likelihood	Impact	Level	Mitigation

Minor gender bias	Possible	Low	🟡 Moderate	Balance training data, use fairness constraints
Privacy–accuracy tradeoff	Probable	Moderate	🟠 High	Adjust ε value, hybrid differential models
Concept drift	Possible	Moderate	🟡 Moderate	Periodic drift checks, retraining if KL > 0.3
Explainability gaps	Improbable	Low	🟢 Low	Maintain SHAP-based reporting dashboard