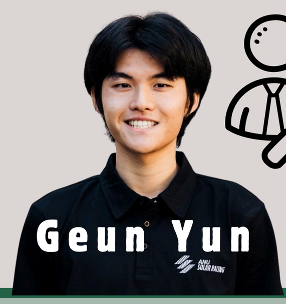
SHIELD:

SHAPLEY AND INFORMATION-THEORY BASED FRAMEWORK FOR EQUITABLE LEARNING VIA DISSIMILAR FEATURE GROUPING

Check out the full paper!


Geun Yun


Prof. Suominen


Prof. Barnard

Check out the repository!

Motivation

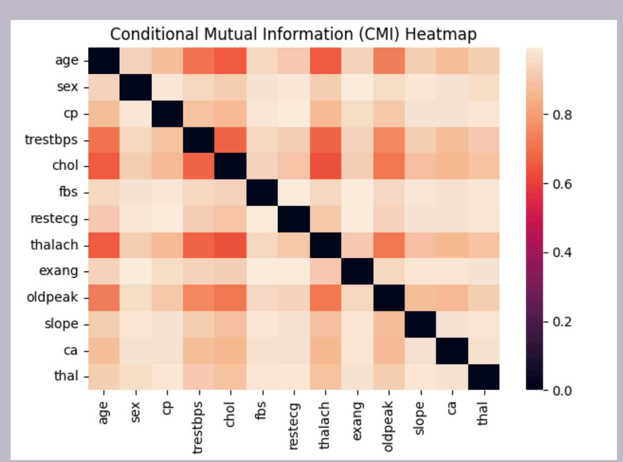
- Problem: Clinical ML models can be accurate, yet opaque and unfair.
- Need: Clinicians and patients need transparent, fair outcomes backed by an equitable decision-making process..
- Gap: Limited integration of Information Theory with SHAP for equitable learning and holistic fairness metrics.
- Solution: SHIELD groups dissimilar features to balance attribution and enable equitable, transparent decisions.

Background

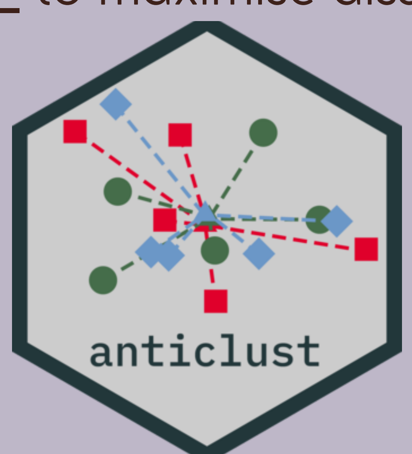
- Conditional mutual information: How much new info a feature adds about the label after knowing others → measurement of dissimilarity.
- SHAP: Splits a prediction into feature contributions, showing which feature positively/negatively contributes to the outcome by how much.
- Bias quadrant: Two-axis fairness map, where x shows explanation diff, and y shows prediction diff between groups, so closer to origin indicates more equal treatment.

Keysteps

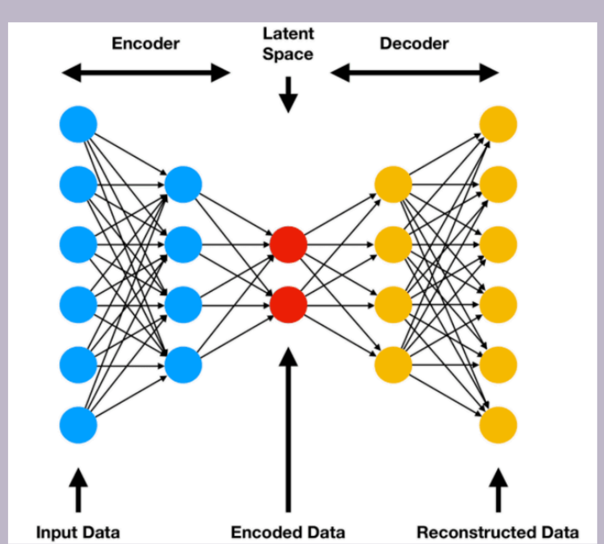
COMPARE features' similarities



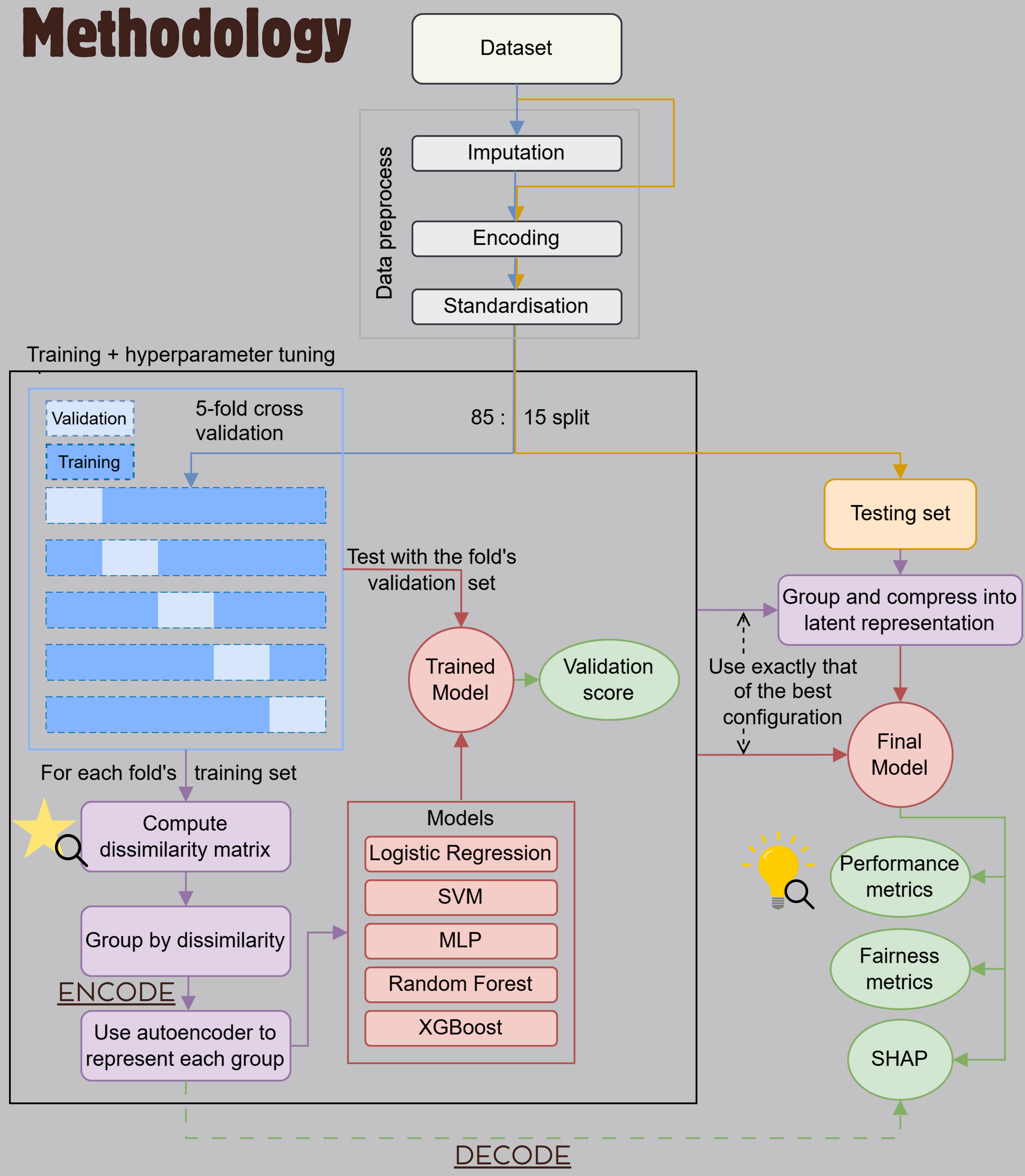
GROUP to maximise dissimilarity



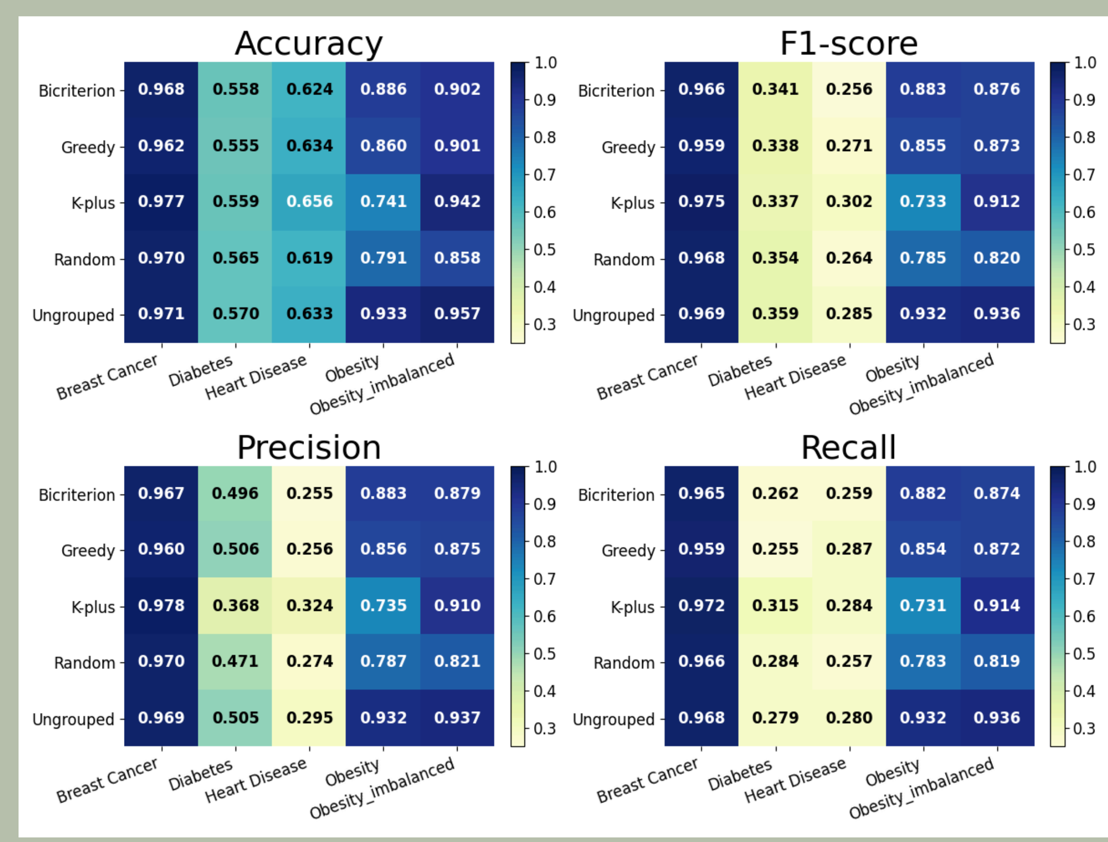
ENCODE to make groups trainable



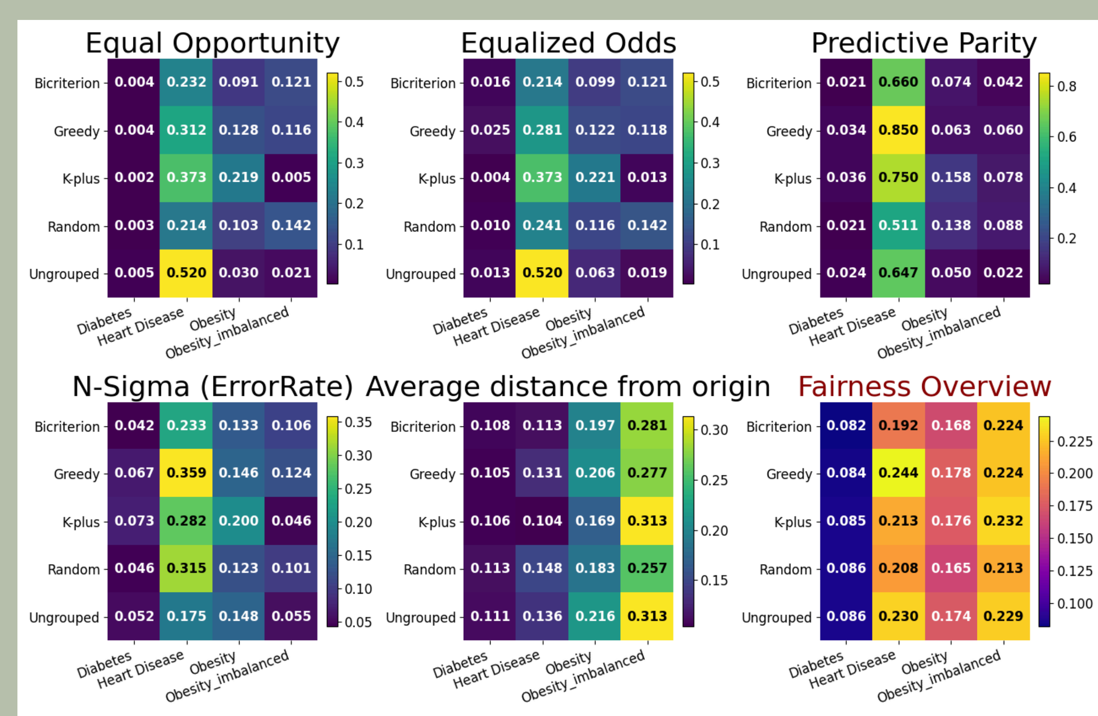
Methodology



Performance Metrics

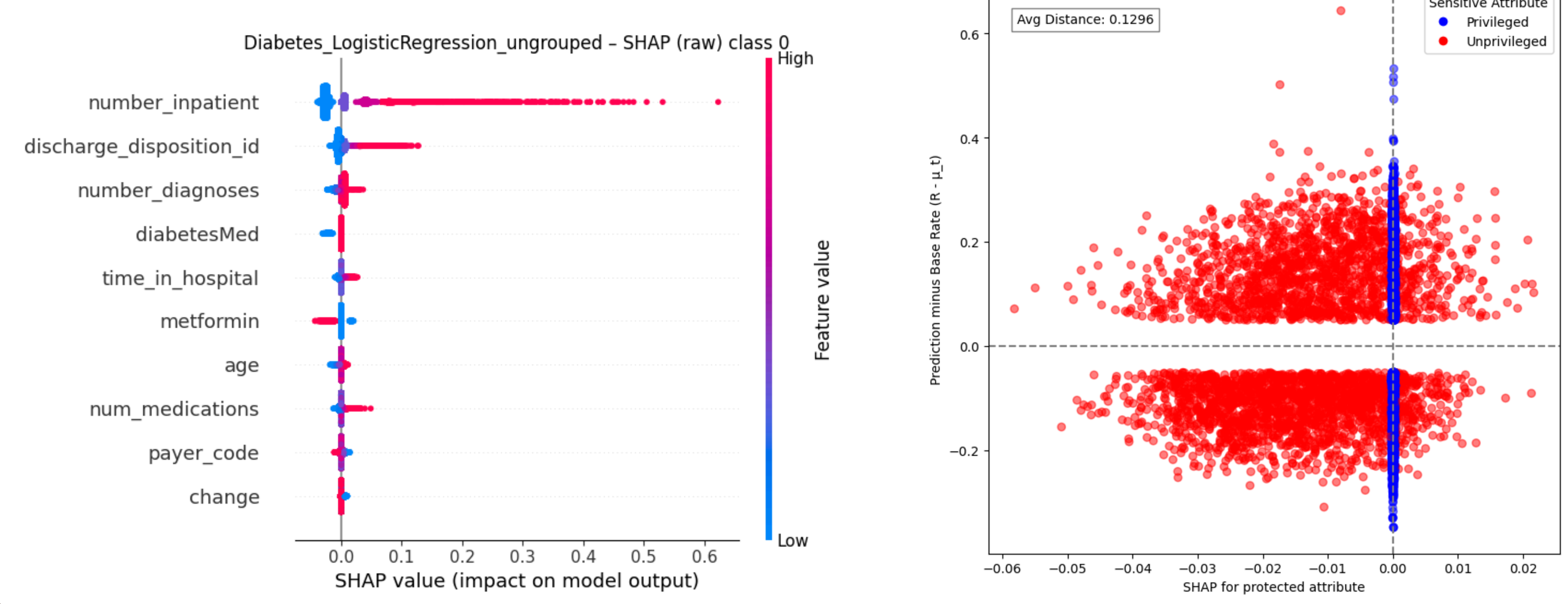


Fairness Metrics

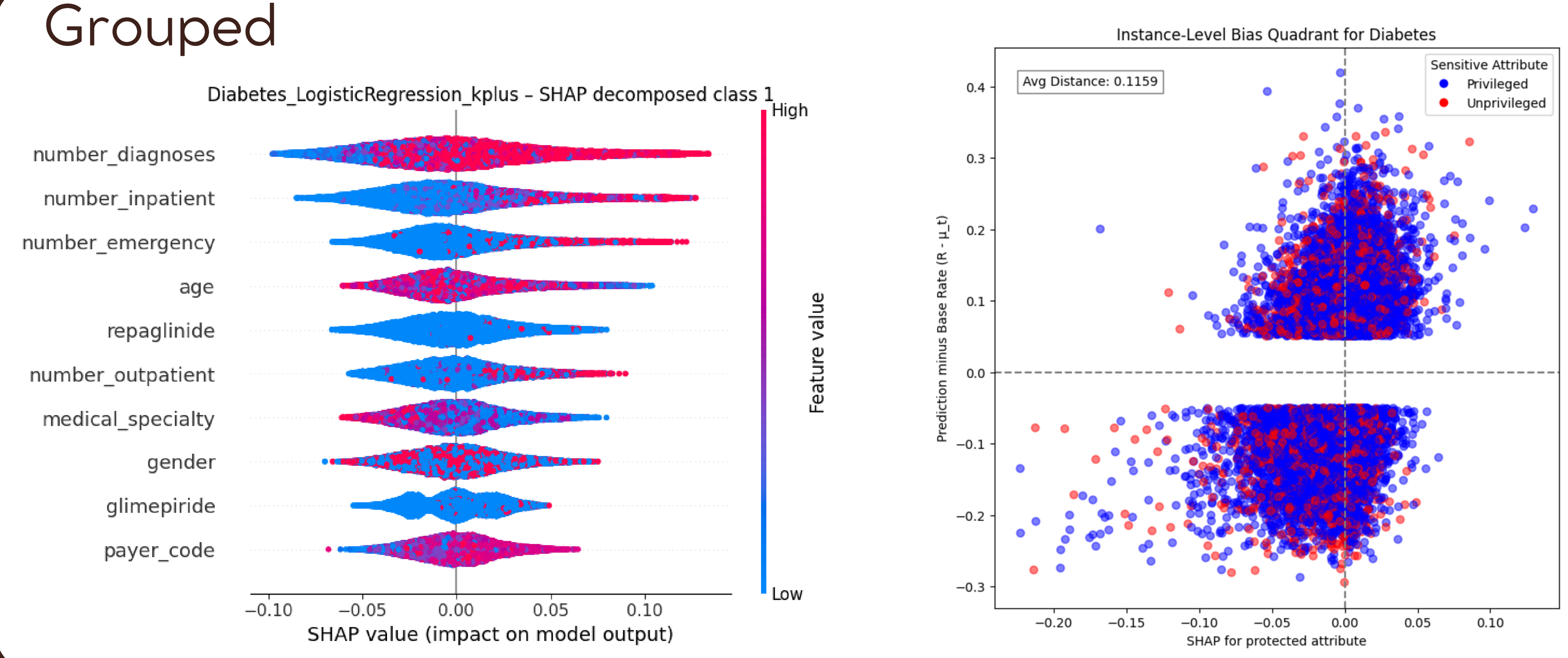


SHAP Analaysis with Bias Quadrant (Diabetes Dataset)

Ungrouped



Grouped



Performance-fairness Trade-off

When compared with ungrouped cases, grouping led to

- ↓ Accuracy and f1-score by 3.43% and 5.16%
- ↑ Avg distance from origin of bias quadrant by 9.47%
- ↑ Fairness overview score by 2.42%.

Grouping Impact on Equitable Learning

The SHAP plots illustrate grouped cases distribute the feature importance more evenly unlike ungrouped, where few of them dominates decisions.

The bias quadrants also show how grouping mitigates the influence of sole membership of protected attribute, like sex, on the outcome as the points are more mixed and less systematically divided.

Implications on Clinical Research

Hence, SHIELD allows more efficient sampling of feature space and participants for researchers. It is also appealing for participants, since each datapoint's contribution to the decision can be more strongly assured. This can all happen while maintaining high predictive performance and even better in fairness metrics.

Ethical approval for the use of datasets was granted by the ANU HREC - Protocol H/2025/0248, "Explainable Health Informatics".