SHAPLEY AND INFORMATION-THEORY BASED FRAMEWORK FO

EQUITABLE LEARNING VIA DISSIMILAR FEATURE GROUPING













Motivation 35



- 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.

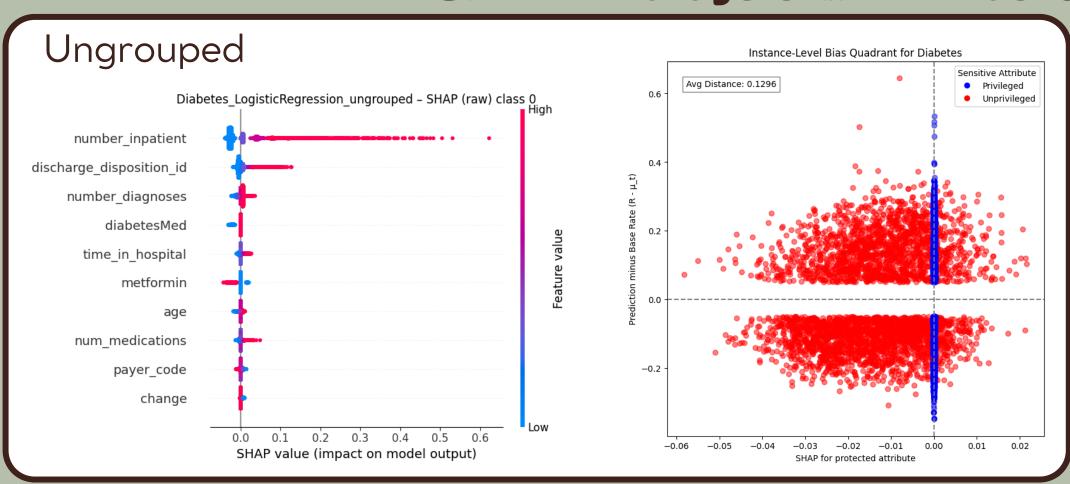
Background

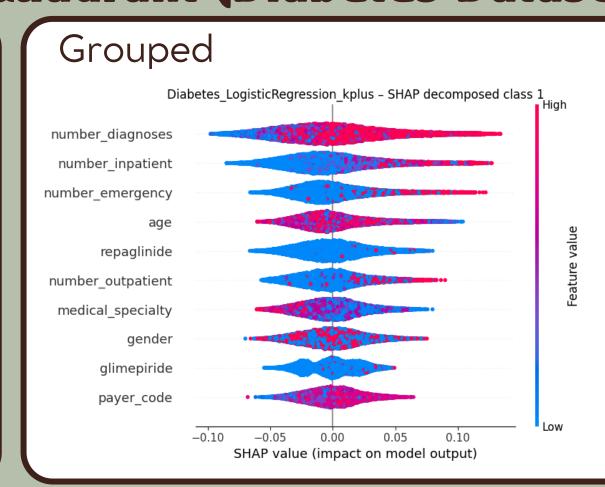


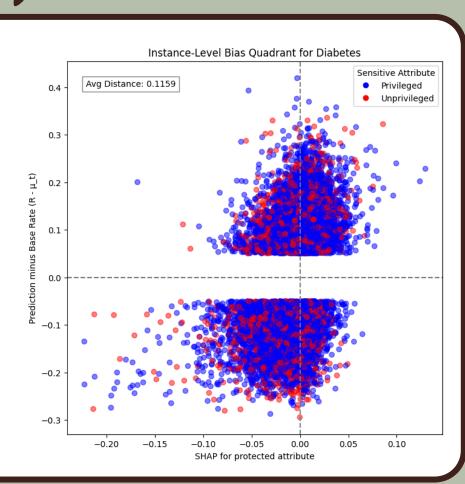
- Conditional mutual information: How much new info a feature adds about the label after knowing others → measurement of dissimilarity.
- SHAP: Splits a prediction into feature contributions, showing which feature positively/negatively contributes to the outcome by how much.
- Bias quadrant: Two-axis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

Performance Metrics Methodology Keysteps **Dataset COMPARE** features' similarities **Imputation** 0.962 0.555 0.634 0.860 0.901 Encoding Standardisation Training + hyperparameter tuning Precision 5-fold cross 85: 15 split Validation 0.256 0.856 validation **GROUP** to maximise dissimilarity Training Testing set Test with the fold's validation set Group and compress into latent representation **Fairness Metrics** Validation Use exactly that Trained of the best Model score anticlust configuration Final **Predictive Parity Equal Opportunity Equalized Odds** Model Bicriterion - 0.004 0.232 0.091 0.121 For each fold's training set Bicriterion - 0.016 0.214 0.099 0,121 Bicriterion - 0.021 0.660 0.074 0.042 ENCODE to make groups trainable Models Compute 0.750 0.158 0.078 dissimilarity matrix Logistic Regression Performance metrics SVM Group by dissimilarity MLP Fairness N-Sigma (ErrorRate) Average distance from origin Fairness Overview metrics Random Forest ENCODE 0.108 0.113 0.197 0.281 **XGBoost** Use autoencoder to SHAP represent each group **DECODE Encoded Data**

SHAP Analaysis with Bias Quadrant (Diabetes Dataset)







Performance-fairness Trade-off —



When compared with ungrouped cases, grouping led to

- \$\rightarrow\$ 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.