Fairness and Interpretability in Machine Learning Models

Introduction and Problem Statement

Problem:
Careless use of machine learning models can do more harm than good. The canonical example being COMPAS.

- Decisions might explicity depend on group membership.
- Decisions might be biased but decisions are hidden.
- Datasets are biased.

Approaches:

- Introduce new loss functions.
- Probabilistic inference.
- Explaining model decision.

Goal: Explore some fairness aware classifiers, evaluate them on datasets in fairness research and use interpretable machine learning methods to explain how the models achieve fairness.

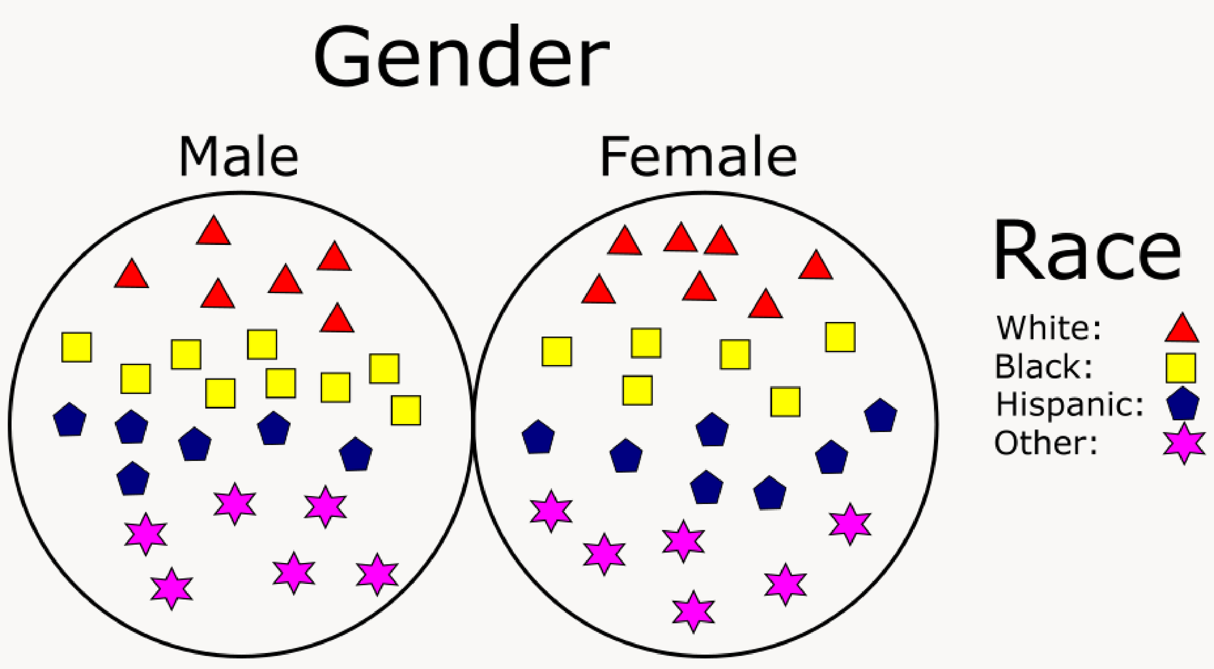
Method

The thesis has focused on implementing three models:

- Fair Bayesian Network with Latent Fair Decisions proposed by Choi et al. [1]
- Fair Tree Classifier proposed by Barata and Veenman [2]
- Naive Bayes trained with and without sensitive attributes.

Datasets:

- Adult dataset
- Compas dataset
- Synthetic dataset



Interpretable Machine Learning:

- Global Model Agnostic Methods
  - Feature Importance.
- Local Model Agnostic Methods
  - Individual Conditional Expectation Plots
  - Counterfactual Explanation Generation

Conclusion

While there exists many models out there that claim to achieve fairness when classifying individuals. This might not always be the case.

- Methods rely on predictions being independent of sensitive groups.
- Just predicting randomly achieves this.
- The model might learn a good prediction function but keeps predictions independent of the sensitive attributes (with a fairness-accuracy tradeoff)
- The might learn a prediction function that just predict randomly the outcome (leading to predictions that are independent of sensitive attributes)
- When a model is evaluated we would like to know which of the above cases are present.

How the models performed:

- Fair Bayesian Network with Latent Fair Decision achieves fairness by affirmative actions.
- The Fair Tree Classifiers achieves fairness by predicting randomly.

How to achieve true fairness:

- Explain the decision that the model makes using interpretability methods.
- Alternatively, researching loss functions for machine learning algorithms that reflect fairness.

Results

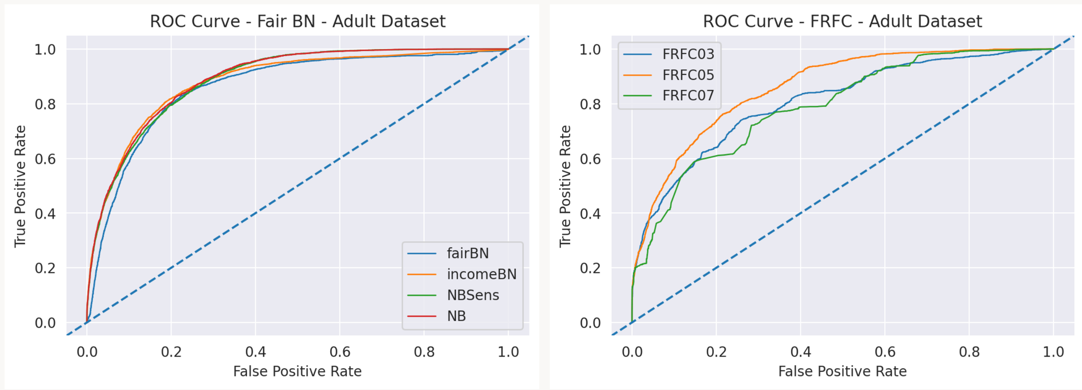


Figure: Roc-Curve of Fair Bayesian Network and Fair Tree Classifier on the Adult Dataset

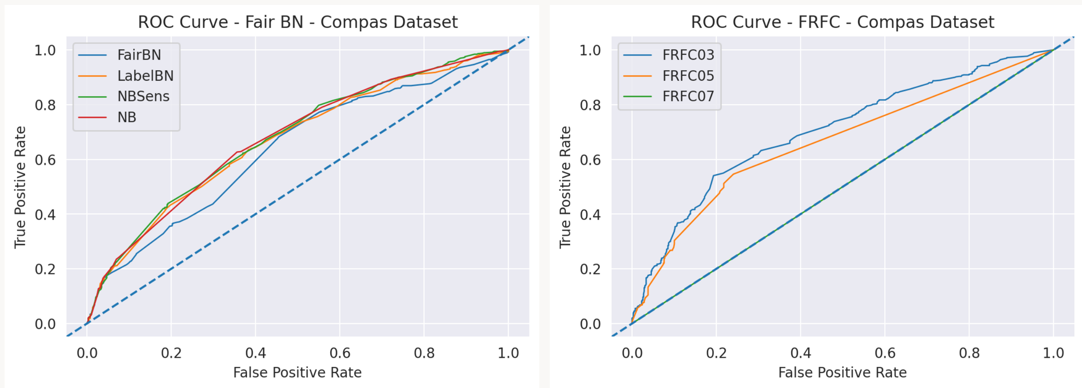


Figure: Roc-Curve of Fair Bayesian Network and Fair Tree Classifier on the Compas Dataset

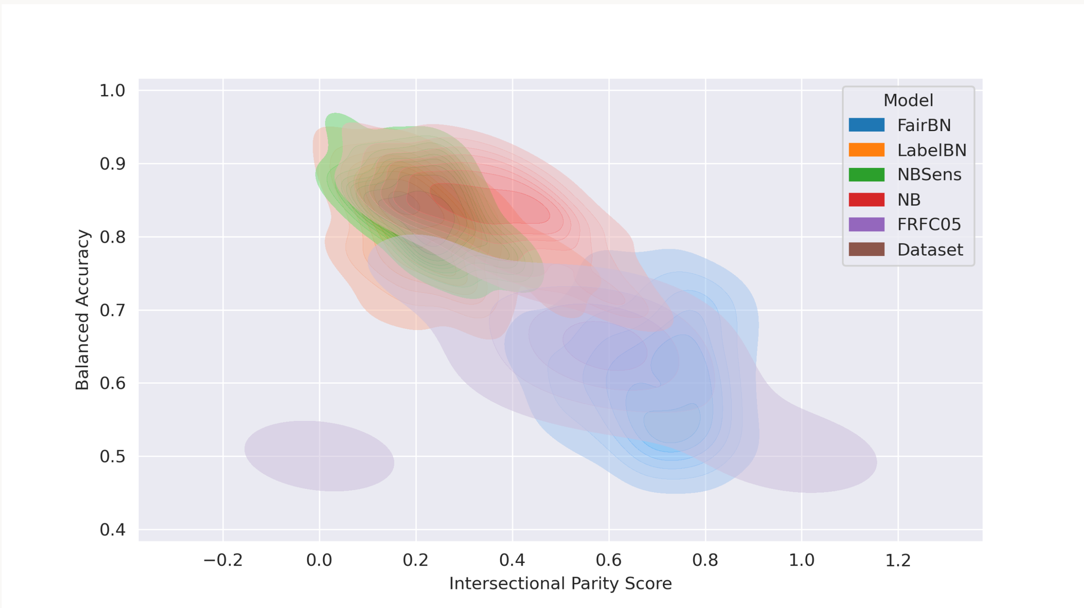


Figure: Balanced Accuracy vs Intersectional Parity Score

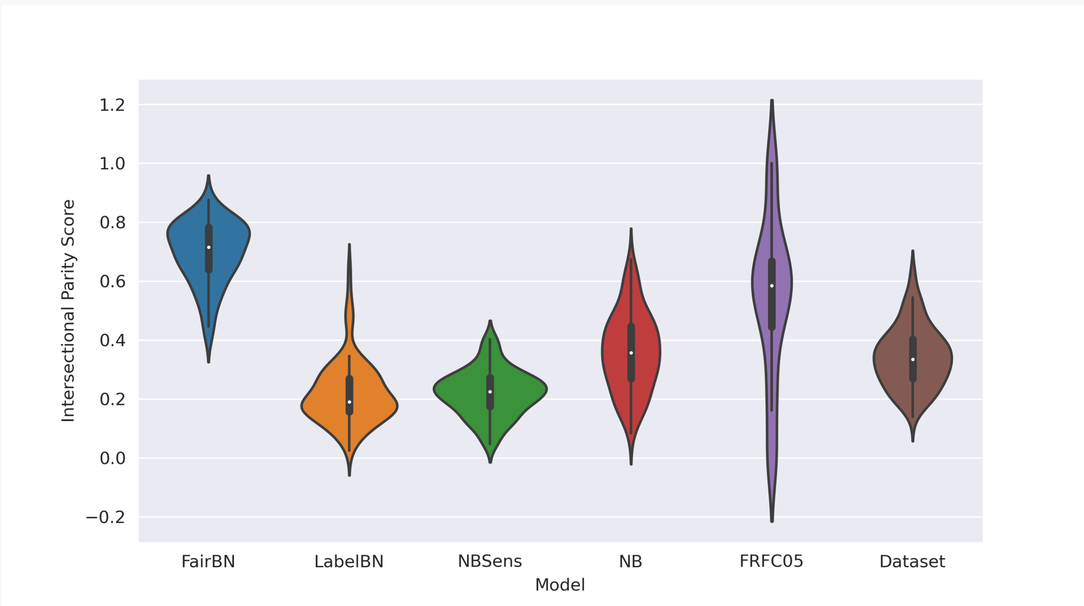


Figure: Intersectional Parity Score for models.

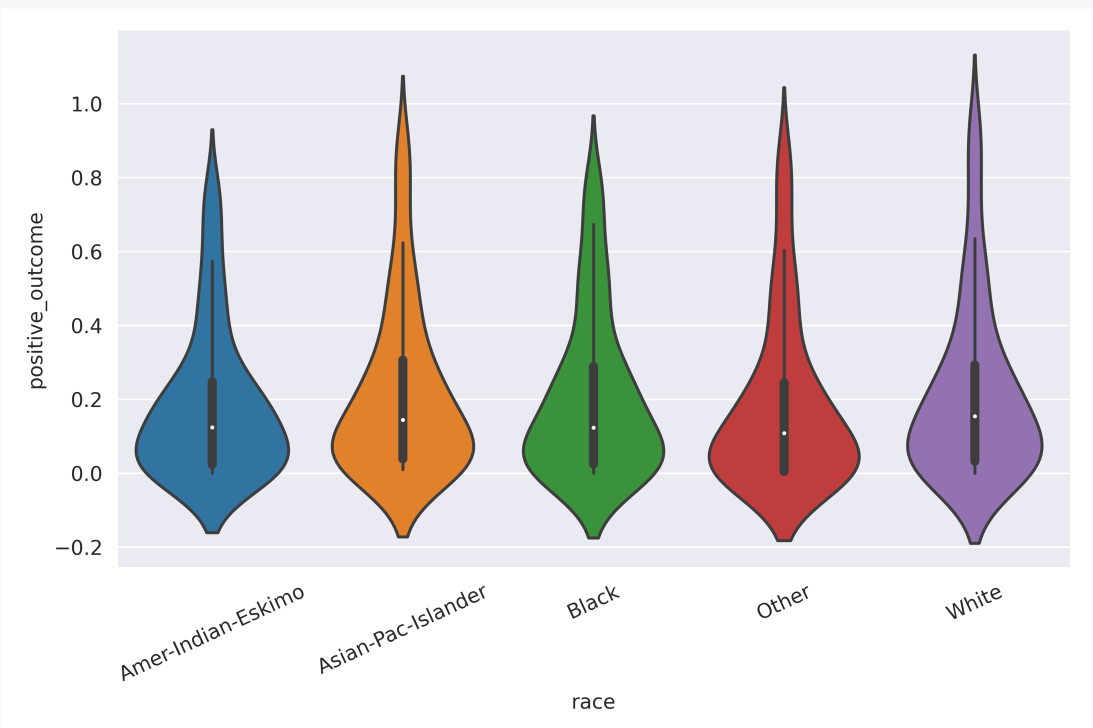


Figure: ICE Plot of Fair Bayesian Network

Naive Bayes

As a baseline method, we train Naive Bayes models both with and without the sensitive attributes in the datasets.

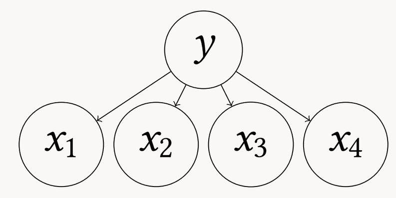


Figure: Bayesian network with 4 features representing the Naive Bayes classifier

Fair Bayesian Network with Latent Fair Decisions

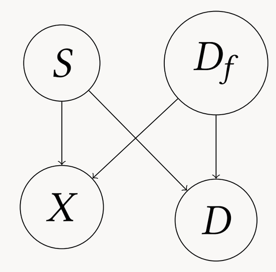


Figure: Bayesian network structures that represent the proposed fair latent variable approach from [1]

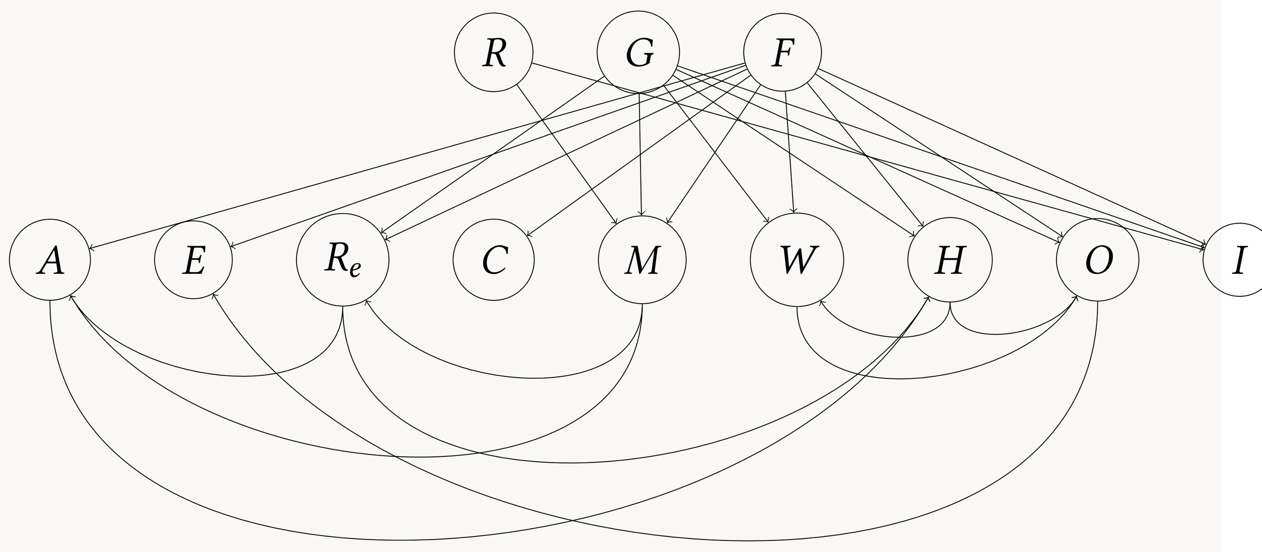


Figure: latentFairClassifier trained on Adult Dataset. The nodes are R: race, G: gender, F: latent fair labels, A: Age, E: Education, R_e: Relationship, C: Capital gain, M: Marital Status, W: Work class, H: Hours-per-week, O: Occupation and I: Income.

The Fair Bayesian Network proposed by Choi et al. [1] tries to simulate the discrimination process. It assumes that the labels D in the dataset are biased and generated through some distribution dependent on the sensitive attributes S and the latent true and fair labels D_f

$$P(D, D_f, S) = P(D|S, D_f)P(S)P(D_f)$$

While the non-sensitive features X are generated through some distribution

$$P(X, D_f, S) = P(X|S, D_f)P(S)P(D_f)$$

Fair Tree Classifier

The fair tree classifier proposed by Barata and Veenman [2]. They introduce a new splitting criterion that evaluates splits in terms of the Area under curve (AUC) wrt the predicted value and the sensitive attribute.
The AUC wrt the true labels Y can be calculated as

$$AUC_Y(\hat{Y}, Y) = \frac{\sum_{t_0 \in Y_-} \sum_{t_1 \in Y_+} \mathbf{1}[\hat{Y}_{t_0} < \hat{Y}_{t_1}]}{|Y_-| \cdot |Y_+|}$$

where Y_- and Y_+ are the set of indexes for negative and positive instances in the true labels. When calculating the AUC wrt the sensitive attribute, denoted AUC_s, the authors have derived the following formula

$$AUC(\hat{Y}, S) = \max(1 - AUC(\hat{Y}, S), AUC(\hat{Y}, S)) \tag{1}$$

The max operator maps the bounds to the range $[0.5, 1]$. The authors then introduce the splitting criterion used in their tree algorihm, Splitting Criterion AUC for Fairness (SCAFF). Which is calculated as

$$SCAFF(\hat{Y}, Y, S, \Theta) = (1 - \Theta) \cdot AUC_Y(\hat{Y}, Y) - \Theta \cdot AUC_S(\hat{Y}, S)$$

Where Θ is a parameter for balancing accuracy and fairness.

References

[1] YooJung Choi, Meihua Dang, and Guy Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12051–12059, 2021.

[2] António Pereira Barata and Cor J. Veenman. Fair tree learning, 2021.

Bjørn Christian Weinbach
bc.weinbach@stud.uis.no
University of Stavanger

Universitetet i Stavanger