1 Fairness Metrics (Verma and Rubin 2018)

Independence $\hat{Y} \perp A$

- Statistical Parity/Demographic Parity: $P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$
- Conditional Statistical Parity: $P(\hat{Y} = 1 \mid E = e, A = a) = P(\hat{Y} = 1 \mid E = e, A = b)$
  E is a set of legitimate features that may affect the outcome.

Separation $\hat{Y} \perp A \mid Y$

- Equalized Odds: $P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$
  `mlr3`: `fairness.equalized.odds`
- Equal Opportunity/ False negative error rate balance: $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$
  `mlr3`: `fairness.tpr`
- Predictive Equality/ False positive error rate balance: $P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$ or
  $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$
  `mlr3`: `fairness.fpr`, `fairness.tnr`
- Treatment Equality: $\frac{\mathrm{FN}}{\mathrm{FP}}\Big|_{A=a} = \frac{\mathrm{FN}}{\mathrm{FP}}\Big|_{A=b}$

Sufficiency $Y \perp A \mid \hat{Y}$

- Predictive parity/ outcome test: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$
  `mlr3`: `fairness.ppv`
- Equal true negative rate: $P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$
  `mlr3`: `fairness.npv`
- Equal false omission rate[1]: $P(Y = 1|\hat{Y} = 0, A = a) = P(Y = 1|\hat{Y} = 0, A = b)$
  `mlr3`: `fairness.fomr`
- Equal false discovery rate[1]: $P(Y = 0|\hat{Y} = 1, A = a) = P(Y = 0|\hat{Y} = 1, A = b)$
- Conditional use accuracy equality: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \wedge P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$

Score-based

- Calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$
- Well-calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b) = s$
- Balance for positive class: $E(S \mid Y = 1, A = a) = E(S \mid Y = 1, A = b)$
- Balance for negative class: $E(S \mid Y = 0, A = a) = E(S \mid Y = 0, A = b)$

Other

- Overall Accuracy Equality: $P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$
  `mlr3`: `fairness.acc`

[1] not officially defined in any of the three papers, but following the same principles as all confusion matrix based metrics

2 Fairness Methods

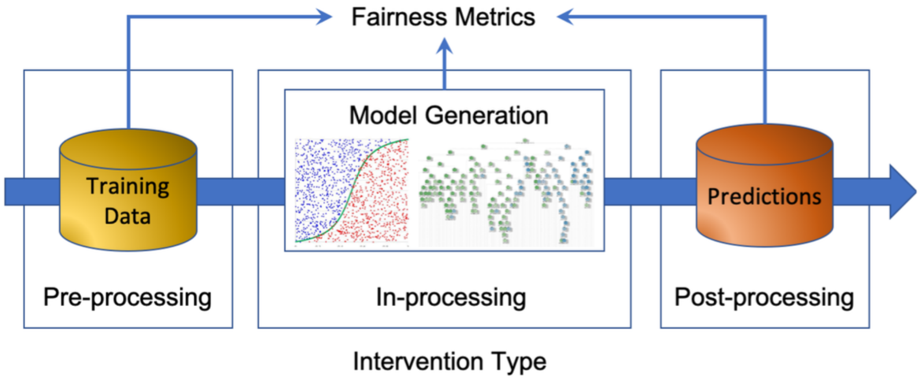

Figure 1: Fairness methods can be applied at different stages of the machine learning pipeline(Caton and Haas 2024).

- Preprocessing: Resampling, Transformation, etc.
- Inprocessing: Regularisation and Constraint Optimisation, Adversarial Learning, etc.
- Postprocessing: Thresholding, Calibration, etc.

3 Sources of bias and the feedback loop

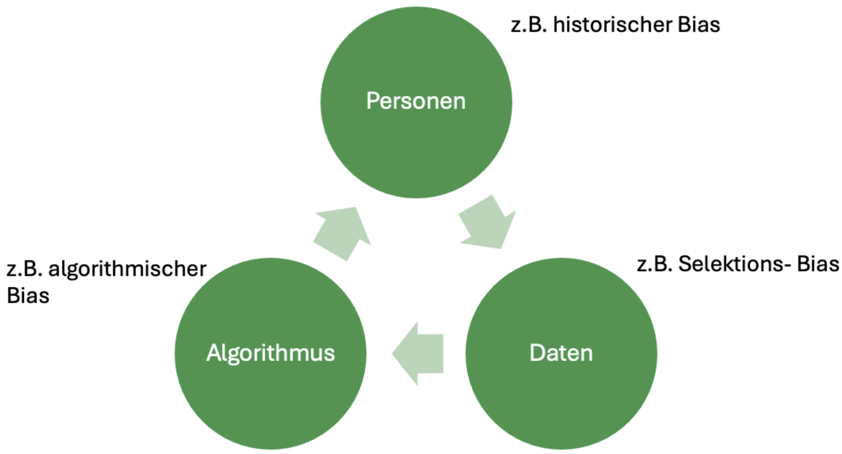

Figure 2: Bias can come into the process at any stage of the data, algorithm, and user feedback loop(Mehrabi et al. 2022).

References

Caton, Simon and Christian Haas (July 2024). "Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7, pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Visited on 12/23/2024).

Mehrabi, Ninareh et al. (July 2022). "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6, pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. (Visited on 01/07/2025).

Verma, Sahil and Julia Rubin (May 2018). "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Visited on 11/16/2024).