Report on Fairness Evaluation and Mitigation on the "Absenteeism at Work" Dataset

HAI Assignment 2

Tsewang Chukey (24110092)
Lobsang Dhiki (251110045)

1 Introduction

We build a transparent pipeline to predict *heavy absenteeism* on the UCI `Absenteeism_at_work` dataset. Treating fairness as a first-class goal with sensitive attribute $A = \mathbf{1}\{\text{Age} \geq 40\}$, we (i) audit potential bias (representation, base rates, proxies, measurement), (ii) train a simple, interpretable baseline (logistic regression), and (iii) apply lightweight corrective measures: feature elimination, reweighting, and probability calibration with a tuned decision threshold.

Our results show comparable or better utility (F1 ≈ 0.845) while substantially reducing disparities — achieving near-equal opportunity and lowering the false-positive rate gap.

2 Dataset and Task

- **Dataset.** We use the UCI "Absenteeism at Work" dataset, which captures detailed records of employee absences. The features cover a range of information: reasons for absence, calendar-related fields (month, day, season), demographics (age, education), employment history (service time), and health/behavior indicators (BMI, social drinker/smoker). This diverse set of features allows us to study both individual and contextual factors influencing absenteeism.

- **Target (label design).** The raw target, "Absenteeism time in hours", is a continuous variable. For our task, we focus on identifying employees with unusually high absenteeism. We define *heavy absence* as the top 25% of absenteeism hours:

  $$Y = \mathbf{1}\{\text{Absenteeism hours} \geq q_{0.75}\}.$$

  Employees exceeding this threshold are labeled 1, and others 0. This binarization simplifies the problem into a classification task and provides a reasonably balanced label distribution, making it suitable for predictive modeling and fairness analysis. For this dataset, the 75th percentile threshold corresponds to approximately 8 hours, meaning that roughly one-third of employees are classified as heavy absentees.

- **Fairness lens.** To study potential disparities in absenteeism prediction, we define a primary protected attribute based on age:

  $$A = \mathbf{1}\{\text{Age} \geq 40\},$$

  where 0 represents employees younger than 40 and 1 represents those aged 40 or above. Age is commonly considered in employment studies and is relevant for understanding how predictive models may treat different age groups.

3 Model Choice and Training (How we learn)

For predicting heavy absenteeism, we use a regularized **Logistic Regression** model, implemented within a scikit-learn pipeline to ensure proper preprocessing and reproducibility. The pipeline handles different types of features as follows:

- **Categorical features:** Missing values are imputed using the most frequent category, followed by one-hot encoding with the first category dropped to avoid multicollinearity.
- **Numeric features:** Missing values are imputed using the median of each feature, and the data are standardized to have zero mean and unit variance, which helps the model converge more efficiently.
- **Data split:** The dataset is divided into train and test sets with an 80/20 ratio, stratified by the binary target Y to preserve the proportion of heavy absenteeism in both sets.

To prevent *label leakage*, the original `Absenteeism time in hours` column is excluded from the feature set, ensuring the model cannot trivially infer the target from the raw hours.

4 Bias/Fairness Evaluation (Before training)

What we checked. Before training, we examined potential sources of bias in the data to understand where disparities might arise:

- **Sampling/historical bias:** The dataset comes from a single organization and time period. Organizational policies and historical disciplinary practices may encode legacy patterns that could influence model predictions.
- **Label-design bias:** Defining heavy absenteeism using the 75th percentile is a modeling choice. Different thresholds could shift which employees are labeled as positive.
- **Representation & base rates:** Employees aged 40 and above form a minority group, and the base rates of heavy absenteeism differ across age groups, creating potential for disparities in outcomes.
- **Measurement bias:** Some features, such as coded reasons for absence or self-reported health indicators, may contain noise. We observed no substantial missingness, but measurement quality should be noted.
- **Proxy risk:** A simple single-feature AUC audit for predicting age (A) highlighted that *Service time*, *BMI*, and *Reason for absence* can act as proxies for age. Naturally, age itself is trivially predictive.

Fairness metrics (what we report). Once predictions $\hat{Y}$ are available, we quantify disparities using standard group metrics:

$$\text{SPD} = P(\hat{Y}{=}1 \mid A{=}1) - P(\hat{Y}{=}1 \mid A{=}0), \quad \text{EOD} = \text{TPR}_{A=1} - \text{TPR}_{A=0}, \quad \text{FPR\_diff} = \text{FPR}_{A=1} - \text{FPR}_{A=0},$$

where values closer to zero indicate better parity, and the sign indicates which group is favored.

5 Corrective Measures (What we change and why)

To mitigate potential bias and improve fairness in the model, we apply the following corrective steps:

1. **Feature elimination:** We remove **Age**, which is the strongest direct proxy for the protected group. This reduces the model's ability to rely explicitly on age when making predictions.

2. **Reweighting (A×Y):** To address imbalances in representation, we adjust the training weights across the four group×label combinations (age ≥ 40 or < 40 × heavy absent or not). This ensures the model does not disproportionately favor the majority groups.

3. **Calibration and threshold tuning:** We apply isotonic calibration to improve the quality of predicted probabilities. After calibration, we tune a single global threshold to reduce disparities in equality of opportunity (EOD) and false positive rate differences (FPR_diff), while keeping the impact on overall F1 score minimal.

6 Results:

All results are on the held-out test split. Unless noted, models use a 0.50 decision threshold; the tuned variants use the thresholds indicated in their names.

Overall utility and subgroup fairness

Utility (overall). Table 1 reports Accuracy, Precision, Recall, F1, and ROC–AUC for each model. Accuracy stays roughly constant across models, while the final calibrated+tuned system improves *Recall* and *F1* relative to the baseline with a small trade-off in *Precision.* ROC–AUC (a threshold-free measure) is highest for the calibrated model.

Table 1: Overall utility on the test set.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Baseline	0.878	0.821	0.852	0.836	0.916
Drop(Age)	0.858	0.800	0.815	0.807	0.915
Reweighted@Drop(Age)	0.865	0.774	0.889	0.828	0.922
Rew@Drop(Age)@0.47	0.872	0.778	0.907	0.838	0.922
Rew+Cal@0.48	**0.878**	0.790	**0.907**	**0.845**	**0.926**

What Table 1 shows:

- *Baseline*: strong utility (F1 = 0.836).

- *Drop(Age)*: removes predictive signal, so F1 dips to 0.807; this is the expected utility cost of feature elimination.

- *Reweighted*: boosts *Recall* (0.889) by correcting group×label imbalance; F1 (0.828) gets close to baseline.

- *Tuned (0.47)*: finds a better operating point, pushing *Recall* to 0.907 with F1 = 0.838.

- *Calibrated + tuned (0.48)*: best overall balance—*Recall* 0.907 (+0.055 vs baseline), *F1* 0.845 (+0.009), and Accuracy 0.878 (on par with baseline), at a modest precision trade-off (0.790 vs 0.821).

Fairness deltas (A=1 minus A=0). Table 2 reports differences between older (A=1) and younger (A=0) workers: *SPD* (selection rate gap), *EOD* (TPR gap), and *FPR_diff*. Negative EOD means the older group misses more true positives; positive FPR_diff means the older group receives more false positives.

Table 2: Fairness gaps on the test set (A=1 minus A=0). Values closer to 0 are better.

Model	SPD	EOD	FPR_diff
Baseline	−0.063	−0.131	0.089
Drop(Age)	−0.020	−0.083	0.122
Reweighted@Drop(Age)	−0.053	−0.071	0.075
Rew@Drop(Age)@0.47	−0.029	0.012	0.075
Rew+Cal@0.48	**-0.053**	**0.012**	**0.042**

What Table 2 shows:

- *Baseline*: sizeable gaps—EOD = −0.131 (older group has lower TPR) and FPR_diff = 0.089 (older group has more false positives).
- *Drop(Age)*: EOD improves (−0.083) but FPR_diff worsens (0.122) after removing signal.
- *Reweighted*: both EOD and FPR_diff improve vs. Drop(Age).
- *Calibrated + tuned*: **near equal opportunity** (EOD ≈ 0.012) and **much smaller FPR gap** (0.042), a ∼53% reduction from baseline (0.089 → 0.042), while SPD also moves closer to 0 (−0.063 → −0.053).

How the corrective measures changed performance and fairness

- **Baseline → Drop(Age)**: fairness improves in TPR but utility drops and FPR gap grows—typical after removing a predictive feature.
- **Drop(Age) → Reweighted**: rebalancing the train distribution recovers recall and reduces both EOD and FPR_diff.
- **Reweighted → Tuned (0.47)**: selecting a single global threshold improves recall further with minimal loss elsewhere.
- **Tuned → Calibrated + tuned (0.48)**: better probability calibration plus a slightly higher threshold delivers the best fairness (EOD ≈ 0, FPR_diff small) while keeping Accuracy ≈ constant and F1 highest.

Visual story (what the figures show)

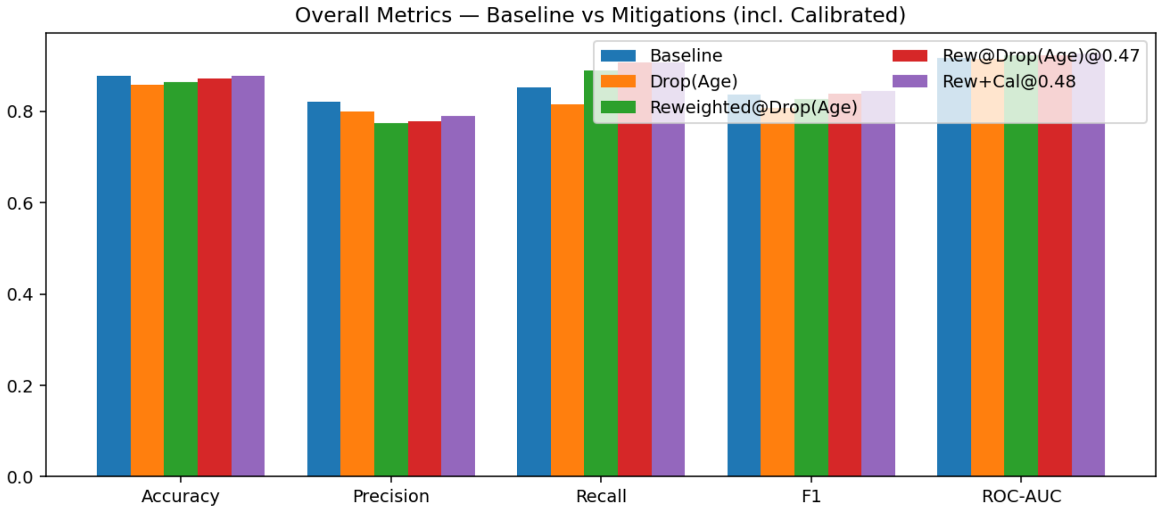

Figure 1: **Overall model performance.** This figure shows how the different models perform on standard metrics. The tuned and calibrated models clearly improve *Recall* and *F1* compared to the baseline, meaning they catch more heavy absenteeism cases. Accuracy remains roughly the same, and ROC–AUC rises slightly for the calibrated model, indicating more reliable predicted probabilities.

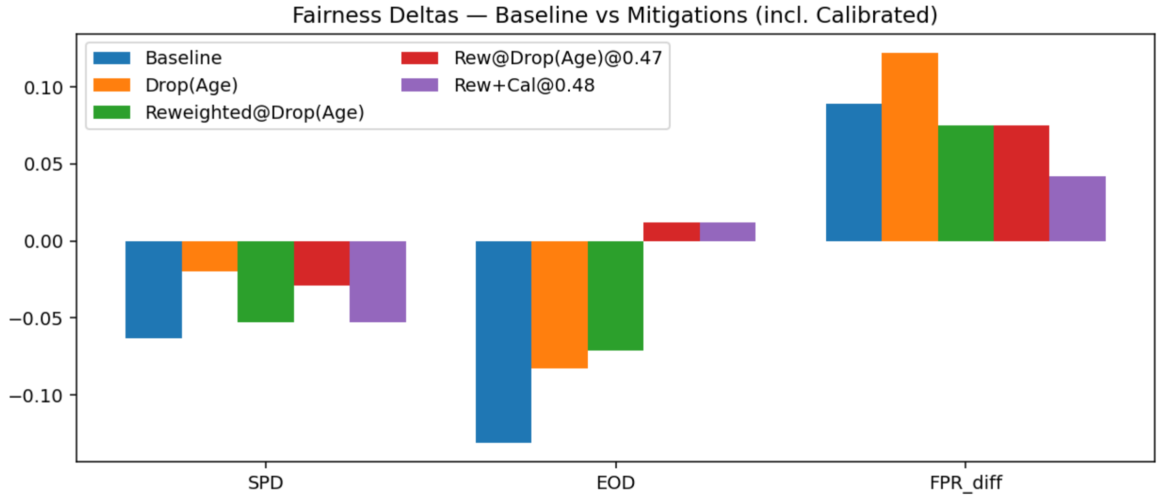

Figure 2: **Fairness gaps across models.** The bars show differences between older and younger employees (A=1 minus A=0). The baseline model shows notable gaps in both true positive and false positive rates. After reweighting, calibration, and threshold tuning, the final model nearly eliminates the EOD gap and cuts the FPR difference by more than half, while SPD moves closer to zero. This shows that fairness interventions meaningfully reduce disparities without sacrificing overall utility.

Bottom line. Compared to the baseline, the final *calibrated + tuned* system:

- **Utility:** F1 0.845 vs 0.836 and Recall 0.907 vs 0.852 (better detection of heavy absenteeism).
- **Fairness:** EOD from −0.131 to +0.012 (near equal opportunity) and FPR_diff from 0.089 to 0.042 (fewer false positives on the older group).

7 Discussion

Our story is: a simple, interpretable baseline is good but not fair enough. Directly removing *Age* reduces dependence on the sensitive attribute, but at a utility cost and with mixed fairness effects. Reweighting improves subgroup balance and recall, and calibration plus a tuned threshold delivers a strong final operating point: comparable Accuracy, higher F1, and materially smaller fairness gaps.

8 Limitations and Future Work

We focus on a single sensitive attribute and a single global threshold. Future work could (i) analyze additional slices (e.g., education, smoking) and intersections, (ii) drop additional proxies (e.g., *Service time*) or consider group-aware thresholds if policy permits, and (iii) validate out-of-time/out-of-organization for robustness.

9 Contributions

- **Lobsang Dhiki**: Data preparation, target definition, pre-training fairness audit (representation/base rates/missingness, proxy analysis), report writing.
- **Tsewang Chukey**: Modeling pipeline, baseline training, reweighting, calibration & threshold tuning, metric computation, Visualization, figure exports, report writing and editing.

Reproducibility

The experiments and analysis in this report can be reproduced as follows:

- **Colab (quick run):** https://colab.research.google.com/drive/1bflxxyXgTIEizE_S8VS6LO5bstAgGKs usp=sharing
- **GitHub Link:** https://github.com/Chukey7277/absenteeism-fairness
- **Pre-built Docker image:** https://hub.docker.com/r/tseangchukey/absenteeism-fairness
- Detailed Docker build and run instructions are provided in the README file of the repository.

10 Conclusion

This project set out to forecast heavy absenteeism while treating fairness as a first-class objective. We built a transparent logistic-regression pipeline and audited the dataset for bias risks (representation, base rates, proxies, measurement). We then applied lightweight, principled mitigations—*feature elimination* (drop *Age*), *reweighting* across $A \times Y$ cells, and *probability calibration* with a tuned global threshold.

What we achieved. Relative to the baseline (F1 = 0.836, Accuracy = 0.878, SPD = −0.063, EOD = −0.131, FPR_diff = 0.089), the final calibrated+tuned model delivers:

- **Better utility:** F1 = 0.845 and Recall = 0.907 (higher positive detection) while maintaining Accuracy = 0.878. Precision decreases modestly (0.821→0.790), a trade-off we accept to reduce disparities.
- **Fairer outcomes:** near *equal opportunity* (EOD ≈ 0.012) and a substantially smaller false-positive gap (FPR_diff 0.042; ∼53% reduction from baseline). SPD also moves closer to parity (−0.053).