CounterFair: Counterfactual Burden Optimization for Bias Detection and Actionability-oriented Fairness

APPENDIX

A. Datasets

The task in the synthetic Athlete dataset is to predict whether an athlete will earn a medal in the olympics. It consists of the following attributes:

- Age: continuous, immutable. Decreases chances linearly from age 25 up to 50 (-1% decrease/year) and remains at 5% at 50 and on.
- Gender: binary, immutable.
- Sport: categorical, mutable, any value. (Football: 0%, Running: +5%, Swimming: +10%, Shooting: +20%).
- Training per week: ordinal, mutable, any direction: (3: -15%, 4: -10%, 5: +15%, 6: +30%).
- daily sleep hours: continuous, mutable, any direction. (+1% increase/hour of daily sleep).
- Diet: Categorical, mutable, any value: (Vegan: +20%, Vegetarian: +15%, Pescetarian: -5%, Omnivore: -10%).

The properties of mutability and directionality for each dataset are shown in Table I.

B. Subgroups identified

The following figures detail the different subgroups identified for the Adult, Athlete, Dutch, German and Student datasets.

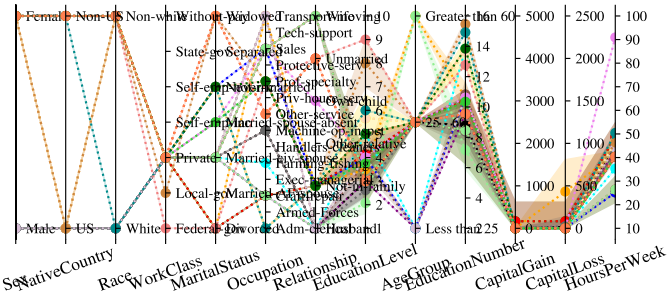

Fig. 1: Adult dataset subgroup details with $\alpha = 0.1$

Dataset	Binary / Categorical	Ordinal / Continuous
Adult 488842 ins. 14 feat.	Sex (✗,-) NativeCountry (✗,-) Race (✗,-) WorkClass (✓,⇳) MaritalStatus (✓,⇳) Occupation (✓,⇳) Relationship (✓,⇳)	EducationLevel (✓,⇑) AgeGroup (✗,-) EducationNumber (✓,⇑) CapitalGain (✓,⇳) CapitalLoss (✓,⇳) HoursPerWeek (✓,⇳)
Athlete 1000 ins. 6 feat.	Sex (✗,-) Diet (✓,⇳) Sport (✓,⇳) TrainingTime (✓,⇳)	Age (✗,-) SleepHours (✓,⇳)
Compas 7214 ins. 52 feat.	Sex (✗,-) ChargeDegree (✓,⇳) Race (✗,-)	PriorsCount (✓,⇳) AgeGroup (✓,⇑)
Dutch 60420 ins. 12 feat.	Sex (✗,-) HouseholdPosition (✓,⇳) HouseholdSize (✓,⇳) Country (✗,-) EconomicStatus (✓,⇳) CurEcoActivity (✓,⇳) MaritalStatus (✓,⇳)	EducationLevel (✓,⇑) Age (✓,⇑)
German 1000 ins. 20 feat.	Sex (✗,-) Single (✓,⇳) Unemployed (✓,⇳) PurposeOfLoan (✓,⇳) InstallmentRate (✓,⇳) Housing (✓,⇳)	Age (✓,⇑) Credit (✓,⇳) LoanDuration (✓,⇳)
Student 395 ins. 33 feat.	Sex (✗,-) School (✓,⇳) AgeGroup (✗,-) FamilySize (✓,⇳) ParentStatus (✓,⇳) SchoolSupport (✓,⇳) FamilySupport (✓,⇳) ExtraPaid (✓,⇳) ExtraActivities (✓,⇳) Nursery (✓,⇳) HigherEdu (✓,⇳) Internet (✓,⇳) Romantic (✓,⇳) MotherJob (✓,⇳) FatherJob (✓,⇳) SchoolReason (✓,⇳)	MotherEducation (✓,⇑) FatherEducation (✓,⇑) TravelTime(✓,⇳) ClassFailures (✓,⇳) GoOut (✓,⇳)

TABLE I: Mutability and directionality for each feature. ✓ means the feature is mutable (may be modified), while ✗ means the feature is immutable. ⇳ means the feature may increase or decrease, while ⇑ means the feature may rise only.

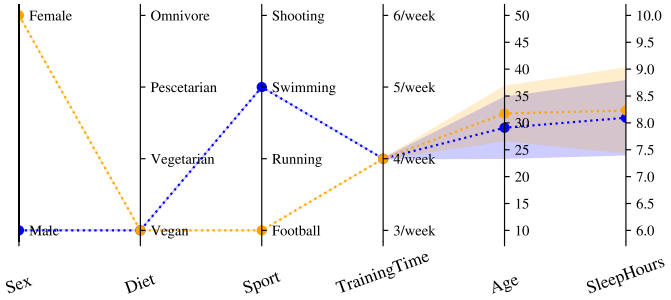

Fig. 2: Athlete dataset subgroup details with $\alpha = 0.1$

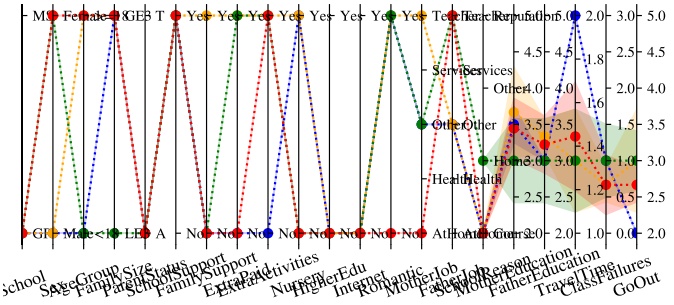

Fig. 3: Student dataset subgroup details with $\alpha = 0.1$

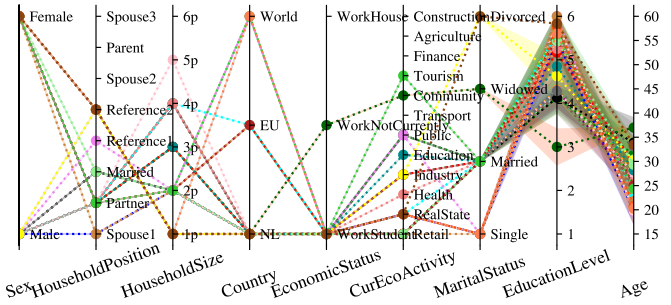

Fig. 4: Dutch dataset subgroup details with $\alpha = 0.1$

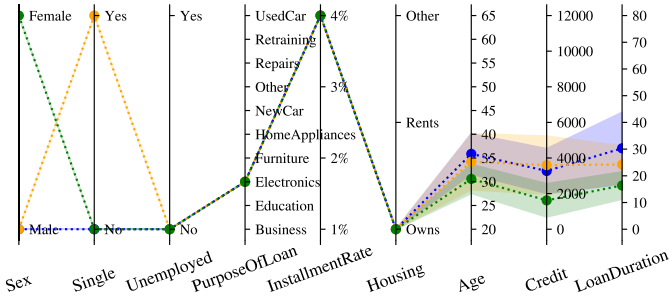

Fig. 5: German dataset subgroup details with $\alpha = 0.1$