Faceted Disparate Impact: A Legal-Consistent Bias Evaluation in Machine Learning Supplementary Material

Anonymous Author(s)

ABSTRACT

We provide supplementary proofs and figures here. Sections 1-4 provide details on the combinatorics and distribution of $\mathcal{B}$ and Section 5 gives more empirical insights.

ACM Reference Format:
Anonymous Author(s). 2018. Faceted Disparate Impact: A Legal-Consistent Bias Evaluation in Machine Learning Supplementary Material. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 PRELIMINARIES

This section goes over the preliminaries for our counting and distribution analyses. We begin by redefining the confusion matrix as a row vector for convenience. We abstract away the titles of TP, FN, FP, and TN since all have equal occurrences in the set of all possible confusion matrices.

Definition 1.1. The confusion matrix, CM, is a quadruple of non-negative integers.

$$CM \in \mathbb{N}_0^4 \tag{1}$$

Definition 1.2. $\mathcal{M}(n)$ is the set of all possible confusion matrices (CMs) for a given amount of samples, $n > 0$.

$$\mathcal{M}(n) = \{\text{CM} : \forall \text{CM} \in \mathbb{N}_0^4 \text{ s.t. } \sum_{c \in \text{CM}} c = n.\} \tag{2}$$

We give an example with the set $\mathcal{M}(1)$ in (3).

$$\mathcal{M}(1) = \{[1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]\} \tag{3}$$

We move from $\mathcal{M}(1)$ to $\mathcal{M}(2)$ and begin to see counting patterns. We denote the count of the value x for any given cell and n samples as $C(x;n)$. For example, $\mathcal{M}(1)$ shows that $C(1;1) = 1$ and $C(0;1) = 3$. As we continue this walkthrough, we identify triangular numbers and generalize using said numbers. Note that 1 and 3 are the first two triangular numbers. We follow Knuth's notation for the w^{th} triangular number in (4) [2].

$$w? = \sum_{j=1}^{w} j \tag{4}$$

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

Next, we show that the triangular sequence continues for $n = 2$ by enumerating in (5) and counting in (6). Then, we generalize the relation of x to n in Theorem 2.1.

$$\mathcal{M}(2) = \begin{cases} [2,0,0,0], [0,2,0,0], [0,0,2,0], [0,0,0,2], \\ [1,0,0,1], [1,0,1,0], [1,1,0,0], [0,1,0,1], \\ [0,1,1,0], [0,0,1,1]. \end{cases} \tag{5}$$

$$\big(C(2;2), C(1;2), C(0;2)\big) = \big(1?, 2?, 3?\big) = \big(1, 3, 6\big) \tag{6}$$

2 COUNTING

Theorem 2.1. *Given the number of samples, n, the count of value $x \in [0,n]$, denoted as $C(x;n)$, appearing in $\mathcal{M}(n)$ is found by*

$$C(x;n) = 0.5(n-x+1)(n-x+2) = (n-x+1)? \tag{7}$$

Proof. First, we prove that C follows the triangular sequence. Since all cells are interdependent and required to sum to n, we use the stars and bars method for n^* stars and $k-1$ bars (8).

$$\binom{n^*+k-1}{k-1} \tag{8}$$

If cell c has value x, then the other three cells must comprise of three non-negative integers summing to $n-x$. In the stars and bars terminology, there are $n-x$ stars and $k=3$ other buckets/cells to choose from.

$$C(x;n) = \binom{n-x+3-1}{3-1} = \binom{n-x+2}{2} \tag{9}$$

Rewriting $C(x;n)$ from the binomial coefficient, we create (10).

$$C(x;n) = \frac{(n-x+2)!}{2!(n-x)!} = 0.5(n-x+1)(n-x+2) \tag{10}$$

Now we prove that $C(x;n)$ is a triangular number:

$$0.5(n-x+1)(n-x+2) = (n-x+1)? \tag{11}$$

It is well established that the w^{th} triangular number is found by

$$w? = 0.5w(w+1). \tag{12}$$

So by substituting $w = n-x+1$, we see that (11) follows the formula for $w?$. □

Corollary 2.2. *The number of possible combinations for any unique CM for a given n is the cardinality of $\mathcal{M}(n)$, denoted as $\mathcal{N}(n)$. $\mathcal{N}(n)$ can be found in a similar fashion to $C(x;n)$ with $n^* = n$ and $k = 4$. $k = 4$ since there are four cells to choose from.*

$$\mathcal{N}(n) = |\mathcal{M}(n)| = \binom{n+4-1}{4-1} = \frac{(n+1)(n+2)(n+3)}{6} \tag{13}$$

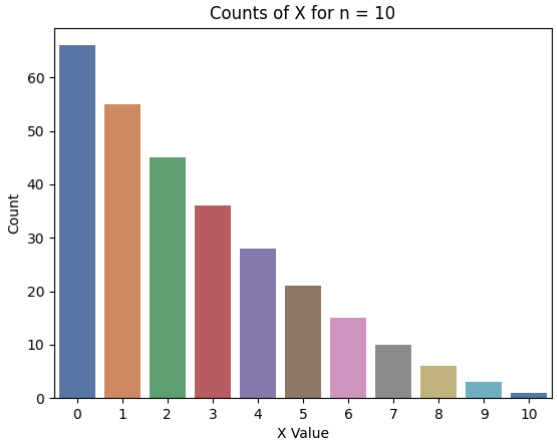

Figure 1: The number of times the value x appears in $\mathcal{M}$.

Corollary 2.3. *The sum of the counts of all values for a given cell is the total number of possible combinations.*

$$\mathcal{N}(n) = \sum_{\alpha=0}^{n} C(\alpha; n) \tag{14}$$

3 A QUASI-TRIANGULAR DISTRIBUTION

We move to prove that $\mathcal{B}$ mimics a triangular distribution, but will not ever be triangular.

Lemma 3.1. *The count of value x increases by $n - x + 2$ for each increment of n. We illustrate this in Figure 1.*

Proof. From Theorem 2.1, we prove that $C(x; n) = (n - x + 1)?$. As such,

$$C(x; n+1) = (n - x + 2)? \tag{15}$$

$$C(x; n+1) - C(x; n) = n - x + 2. \tag{16}$$

□

Theorem 3.2. *$\mathcal{B}$'s distribution is not triangular, although it mimics a triangular distribution as n approaches ∞.*

Proof. A triangular distribution is defined by a peak (mode), a minimum a, a maximum b, and a linear increase from a to the mode and from b to the mode. $\mathcal{B}$'s minimum and maximum are trivially -1 and 1. We will now prove that $\mathcal{B}$'s mode is zero.

The distribution of $\mathcal{B}$ is calculated by collecting all scores, including duplicates. We name this distribution $d_{\mathcal{B}}$.

$$d_{\mathcal{B}} = \{\mathcal{B}(FP, FN), \forall CM \in \mathcal{M} \tag{17}$$

We define the distribution to highlight that every valid confusion matrix is considered. Furthermore, $\mathcal{B}$ is the difference between two values with the same distribution, subject to the constraint $FP + FN \leq n$. As such, we find that equivalent values are the most common pairs. The difference between equivalent values is zero, hence the mode is zero.

If $d_{\mathcal{B}}$ was triangular, the final step would be to prove its piecewise linear property with the endpoints of extrema and mode. For $d_{\mathcal{B}}$, this is to prove that each total (sum of all counts yielding a score) increments linearly with respect to the score. We found this difficult since some values (x) for a finite n have their multiplicity decrease by one from the prior value's ($x - 1$) multiplicity when the trend should be overall increasing. This discrepancy becomes exceedingly minuscule as we scale n as evident.

But, we prove that $\mathcal{B}$ is not a triangular distribution by contradiction. Should $\mathcal{B}$ be triangular, then its standard deviation would be found by $1/\sqrt{6}$ since it is symmetric with bounds $[-1, 1]$ [3]. However, we show in Theorem 4.2 that $d_{\mathcal{B}}$'s standard deviation converges to $1/\sqrt{10}$. □

4 STANDARD DEVIATION AND MEAN OF $\mathcal{B}$

We find the mean $\bar{\mathcal{B}}$, variance σ^2, and standard deviation σ for the marginal benefit $\mathcal{B}$. Additionally, we provide their limits/convergences. We begin with $\mathcal{B}$'s mean in Lemma 4.1 as it is a necessary preliminary for the variance and standard deviation.

Lemma 4.1. *$\mathcal{B}$ has a mean of zero.*

$$\bar{\mathcal{B}} = 0 \tag{18}$$

Proof. $\mathcal{B}$ is equivalent to $\mathcal{B}(FP, FN; n)$, a function of two cells in CM, parameterized by the count of all cell values, n. With the input space $\mathcal{M}(n)$, for any possible elements of FP, there exists the same element of FN at the same multiplicity. For simplicity, let

$$FP = \{a_1, a_2, \ldots, a_i, \ldots, a_{\mathcal{N}(n)}\} \tag{19}$$

$$FN = \{a_1, a_2, \ldots, a_i, \ldots, a_{\mathcal{N}(n)}\} \tag{20}$$

Then we use the Associative Law of Summation.

$$\bar{\mathcal{B}}|n = \frac{1}{\mathcal{N}(n)} \sum_{i=1}^{\mathcal{N}(n)} \frac{a_i - a_i}{n} = 0 \tag{21}$$

□

We now move to Theorem 4.2 to show how to calculate the standard deviation of $\mathcal{B}$ given any n in constant time. After, we give the limit of $\sigma(\mathcal{B}; n)$ in Corollary 4.3. Next, we give the variance and its limit in Corollaries 4.4 and 4.5.

Theorem 4.2. *The standard deviation of $\mathcal{B}|n$ is given by*

$$\sigma(\mathcal{B}; n) = \sqrt{\frac{n+4}{10n}} \tag{22}$$

Proof. We first cite the formula for calculating the standard deviation for a set S with cardinality $\mathcal{N}$ and mean $\bar{s}$.

$$\sigma = \sqrt{\frac{\sum_{s_i \in S}(s - \bar{s})^2}{\mathcal{N}}} \tag{23}$$

In $\mathcal{B}$, each score $s \in S$ is calculated by iterating over all confusion matrices (the set $\mathcal{M}$) and mapping through function $\mathcal{B}$. To find all the confusion matrices in $\mathcal{M}$, we take three loops (summations). The fourth cell (TN) is equal to the remainder of n minus the three cells so it is not looped over.

Furthermore, from Lemma 4.1, the mean for $\mathcal{B}$ is $\bar{\mathcal{B}} = 0$. So, FDIU's standard deviation is given by

$$\sigma(\mathcal{B};n) = \sqrt{\frac{1}{\mathcal{N}(n)}\sum_{TP=0}^{n}\left(\sum_{FP=0}^{n-TP}\left(\sum_{FN=0}^{n-TP-FP}\left(\mathcal{B}(FP,FN;n)-\bar{\mathcal{B}}\right)^2\right)\right)}$$
$$= \sqrt{\frac{1}{\mathcal{N}(n)}\sum_{TP=0}^{n}\left(\sum_{FP=0}^{n-TP}\left(\sum_{FN=0}^{n-TP-FP}\left(\frac{FP-FN}{n}\right)^2\right)\right)} \tag{24}$$

The summations inside the radical simplify to

$$\sum_{TP=0}^{n}\left(\sum_{FP=0}^{n-TP}\left(\sum_{FN=0}^{n-TP-FP}\left(\frac{FP-FN}{n}\right)^2\right)\right) = \frac{(n+1)(n+2)(n+3)(n+4)}{60n} \tag{25}$$

Using the definition of $\mathcal{N}$ from (13),

$$\sigma(\mathcal{B};n) = \sqrt{\frac{1}{(n+1)(n+2)(n+3)/6}\cdot\frac{(n+1)(n+2)(n+3)(n+4)}{60n}} = \sqrt{\frac{n+4}{10n}} \tag{26}$$

□

Corollary 4.3. *The standard deviation of $\mathcal{B}$ converges to $1/\sqrt{10} \approx 0.316$.*

$$\lim_{n\to\infty}\sigma(\mathcal{B};n) = \lim_{n\to\infty}\sqrt{\frac{n+4}{10n}} = \frac{1}{\sqrt{10}} \tag{27}$$

Corollary 4.4. *The variance of $\mathcal{B}$ is $(n+4)/10$.*

$$\sigma^2(\mathcal{B};n) = \frac{n+4}{10} \tag{28}$$

Corollary 4.5. *The variance of $\mathcal{B}$ converges to $1/10$.*

$$\lim_{n\to\infty}\sigma^2(\mathcal{B};n) = \frac{1}{10} \tag{29}$$

We use SymPy's [4] computer algebra system (CAS) to verify Theorem 4.2 and Corollaries 4.3, 4.4, and 4.5.

5 EMPIRICAL STUDIES

We present four case studies from two well-regarded baselines: COMPAS (risk of recidivism) and Folktable's Adult Employment (is person employed) datasets.

In Figure 2, we see that FDI affirms preliminary findings of DI. This is an important revelation, as DI is susceptible to arguments and cases of higher prediction rates being correct. In return, some argue that higher rates would be due to systematic bias. However, FDI sidesteps the need to prove incorrect ground labels and shows that even if one assumes correct ground labels, there's algorithmic bias against certain ethnicities. We believe this revelation will encourage more people to take corrective measures.

In Folktable's cases, we investigate Georgia's census database from 2014 to 2017 (a total of 393,236 observations) and predict if an individual is employed. Using the random forest classifier (RF), Figure 3 reveals that FDI corroborates with DI in this case study, as it was in COMPAS. However, in the Naïve Bayes (NB) case study with Folktables, we find that FDI strongly disagrees with DI when context necessitates it.

In the NB case study, FDI strongly disagrees with DI's findings regarding Pacific Islanders. DI indicates that Pacific Islanders have a bias for them compared to 7/8 other races while FDI shows that, with context, Pacific Islanders have a bias for them over only 2/8 other races. The most prominent distinction occurs when comparing Pacific Islanders with Whites: DI suggests a large bias for Pacific Islanders while FDI shows a slight bias against them (Figure 4).

We further investigate FDI's nuances by drawing 60 random samples to mimic a smaller dataset. In Figure 5, we see TE and DI's problematic undefined cases. Furthermore, the "Two or more races" holds a constant 0 in DI, while FDI gives more information. FDI's nuance still shows substantial bias against the mixed-race class but indicates that they are more susceptible to bias when compared with whites as opposed to blacks.

Our practical use cases show that FDI gives unique insights into understanding the manifested bias.

Notes: due to DI's asymmetry, the heatmap's color map's gradient centers toward blue instead of gray like TE and FDI do. Furthermore, TE's gradient is flipped as this definition has larger values indicating a bias toward Race i, unlike the others. Finally, since we assume the positive label is preferable, we flip the COMPAS labels so the positive is "not a recidivist".

ETHICS STATEMENT

This research is founded on the commitment to deepen the understanding of biases in machine learning through the critical lenses of legal philosophy and ethics. We emphasize that users of FDI should gain a thorough understanding of its meaning before taking corrective actions. Additionally, we recommend ongoing compliance checks with relevant local and international laws to ensure that the use of FDI adheres to evolving legal standards and contributes positively to the broader goal of fairness in machine learning applications. Through these practices, we aim to foster a responsible and legally consistent approach to the mitigation of bias in AI systems.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[2] Donald E. Knuth. 1968. *The Art of Computer Programming, Volume I: Fundamental Algorithms.* Addison-Wesley.

[3] Samuel Kotz and J. Dorp. 2004. Beyond beta. Other continuous families of distributions with bounded support and applications. (01 2004). https://doi.org/10.1142/9789812701282

[4] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. 2017. SymPy: symbolic computing in Python. *PeerJ Computer Science* 3 (2017), e103. https://doi.org/10.7717/peerj-cs.103

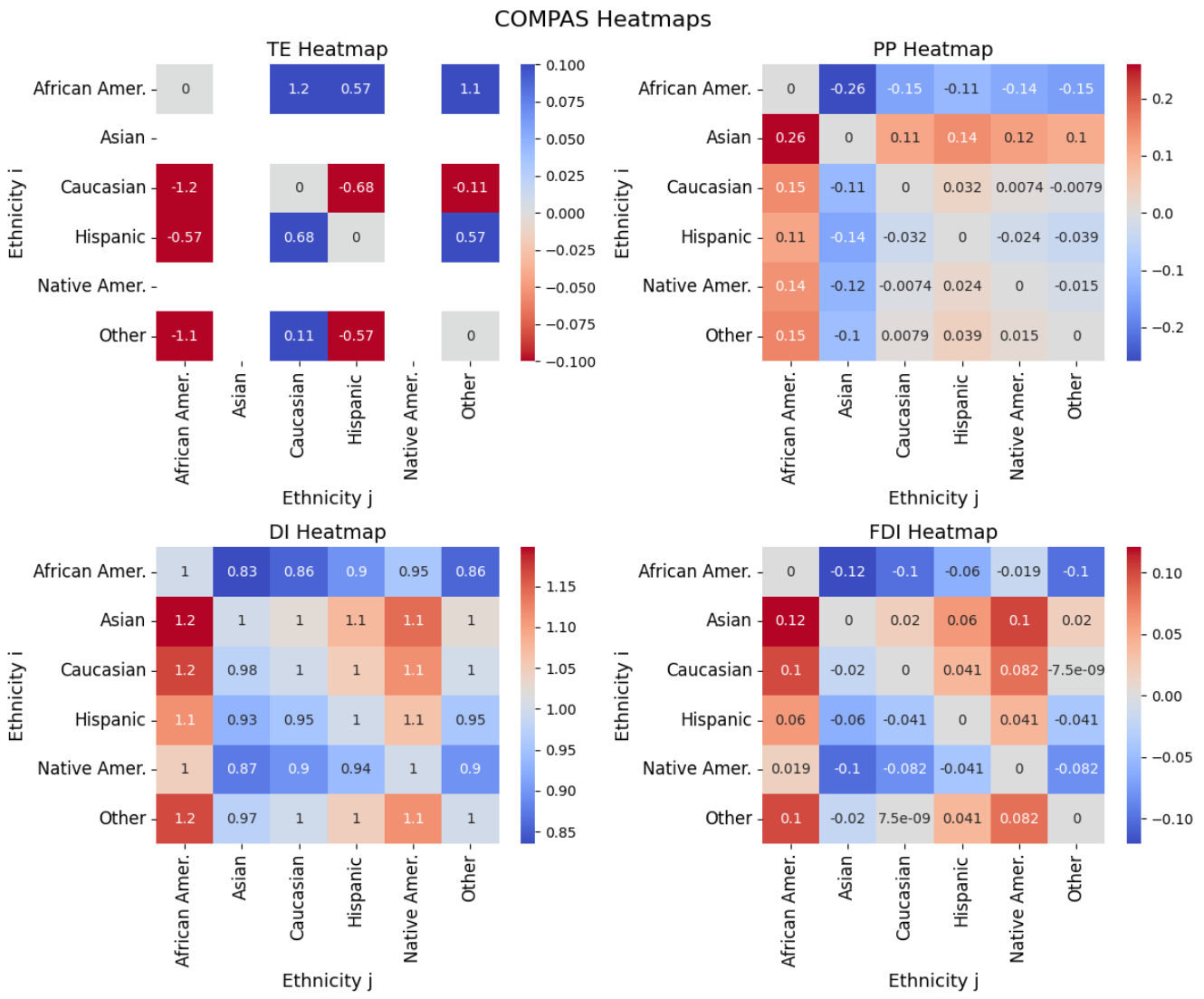

Figure 2: Assessing COMPAS with results published by Propublica[1].

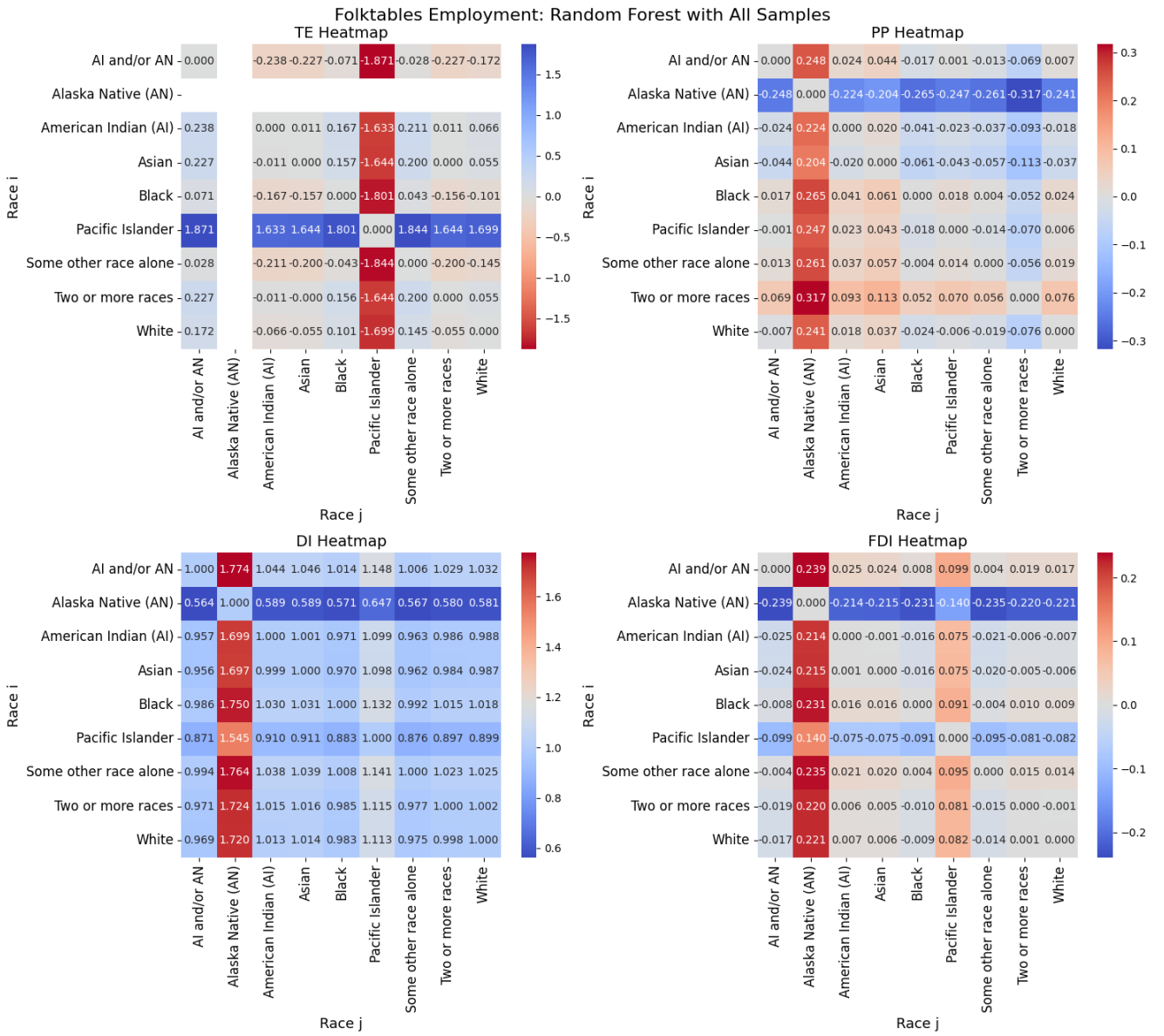

Figure 3: Assessing Folktables with a random forest classifier (all samples).

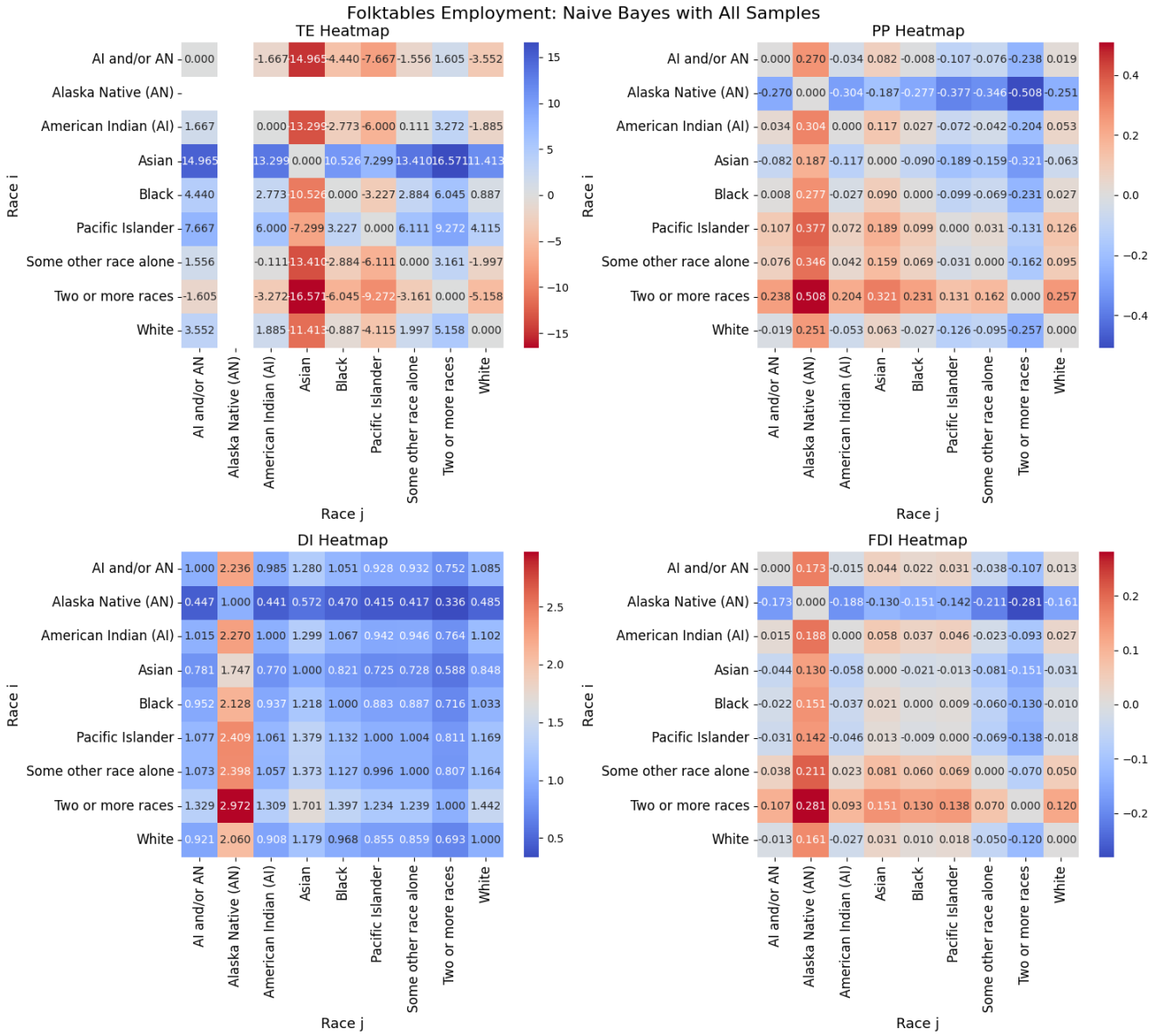

Figure 4: Assessing Folktables with Naïve Bayes.

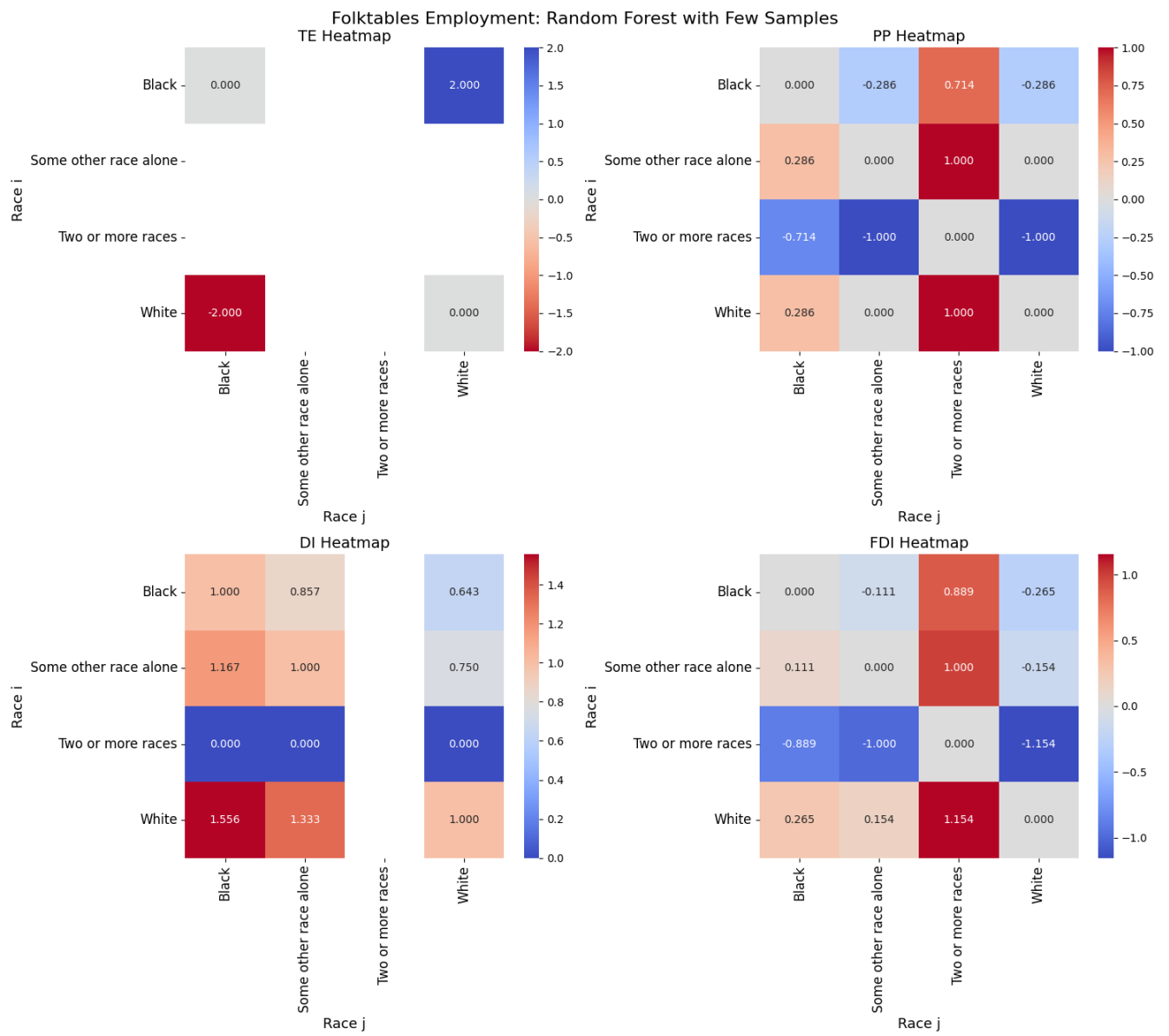

Figure 5: Assessing Folktables with a random forest classifier (60 samples).