Evaluasi AI Bias and Fairness dalam Akuisisi Agen Penjualan Perbankan (Agen BRIlink-Bank Rakyat Indonesia)

1st Berliana Shafa Wardani
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
berlianashafa@students.telkomuniversity.ac.id

2nd Siti Sa'adah
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
sitisaadah@telkomuniversity.ac.id

3rd Dade Nurjanah
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
dadenurjanah@telkomuniversity.ac.id

***Abstrak* — Membangun loyalitas dengan para nasabah dengan melakukan pengambilan kepemilikian perusahaan atau aset (akuisisi) untuk menjadi pihak yang diajak bekerja sama (agen) bank disebut dengan akuisisi agen. Fitur-fitur penting nasabah dipertimbangkan dalam proses akuisisi. Penelitian ini dilakukan dengan dataset BRIlink yang merupakan penerapan akuisisi agen penjualan perbankan di Bank Rakyat Indonesia (BRI). Dengan banyaknya data nasabah BRI dapat menimbulkan keberagaman data yang memungkinkan menyebabkan hasil akuisisi agen tidak merata. Dengan ini, diperlukan algoritma pendeteksi dan mitigasi bias untuk mencapai fairness. AI fairness 360 (AIF 360) merupakan sebuah toolkit yang menyediakan algoritma deteksi dan mitigasi bias. Algoritma mitigasi bias pada AIF 360 dibagi menjadi tiga proses, yaitu: reweighing dan learning fair representation pada tahap pre-processing, prejudice remover dan adversarial debasing pada tahap in-processing, serta calibrated equalized odds dan reject option classification pada tahap post-processing. Luaran penelitian ini berupa hasil perbandingan perhitungan deteksi bias dengan disparate impact (DI) dan statistical parity difference (SPD) sebelum dan sesudah mitigasi. Algoritma reweighing menghasilkan rata-rata DI 0,8% dan SPD 0,102% yang menunjukkan berhasilnya mitigasi, tetapi nilai AUC pada reweighing berkurang. Berbeda dengan reweighing, adversarial debiasing dan reject option classification dapat memitigasi bias sembari mempertahankan nilai AUC. Dilakukannya penelitian ini dapat membantu akuisisi agen BRIlink secara lebih adil.**

***Kata kunci*— akuisisi agen, bias, fairness, mitigasi, BRIlink.**

I. PENDAHULUAN

A. Latar Belakang

Bank Rakyat Indonesia (BRI) membangun loyalitas nasabah salah satunya dengan melakukan akuisisi atau pengambilan kepemilikan perusahaan berupa aset untuk menjadi agen (pihak yang diajak bekerja sama). Agen BRIlink yang tersebar di seluruh Indonesia merupakan penerapan akuisisi agen di BRI. Dengan jumlah nasabah BRI yang besar, hal ini dapat menimbulkan peluang terjadinya akuisisi agen BRIlink yang tidak adil. Ketidakadilan ini dapat menyebabkan bias sehingga perlu dilakukan pembelajaran dalam suatu data. Bias dapat muncul disebabkan oleh fitur-fitur, sebagaimana [1] menunjukkan bahwa jenis kelamin dapat menyebabkan bias pada data. Bias yang muncul pada sebuah dataset, akan menyebabkan individu atau grup dirugikan [2].

Terdapat beberapa fitur demografi dalam data BRIlink yang dapat menimbulkan bias, seperti jenis kelamin, usia, lokasi, pekerjaan, dan ras [3-5]. Disebut fitur demografi, karena fitur tersebut memiliki banyak nilai yang beragam sehingga dapat mengindikasi bias. Misalnya pada penelitian [6] ras digolongkan sebagai fitur demografi karena, ras dapat menentukan seberapa berat hukuman yang diberikan berdasarkan histori kriminal yang ada. Jika histori 2 kriminal condong pada orang berkulit hitam maka orang berkulit hitam akan mendapatkan hukuman yang lebih berat, hal ini menunjukkan suatu diskriminasi. Merujuk pada dataset BRIlink, agen yang diakuisisi lebih sedikit dibandingkan dengan agen yang tidak diakuisisi, hal ini menunjukkan bahwa fairness belum tercapai. Oscar, Deho et.al (2021) mengimplementasikan algoritma mitigasi pada dataset learning analytics, dan menunjukkan bahwa dataset menghasilkan prediksi yang lebih baik ketika data fair. Fair adalah keadaan dimana tidak adanya diskriminasi pada data. Oleh karena itu mitigasi bias perlu dilakukan pada dataset BRIlink. Untuk melakukan mitigasi bias terdapat suatu toolkit bernama AI Fairness 360 (AIF 360) yang merupakan toolkit terbaru dan lengkap karena AIF 360 juga menyediakan algoritma untuk deteksi bias [7]. Deteksi bias pada dataset BRIlink dilakukan dengan metode disparate impact (DI) dan statistical parity difference (SPD) karena metode ini sederhana untuk dikomputasikan. Setelah dilakukan pendeteksian bias, dilakukan mitigasi bias melalui pre-processing dengan metode reweighing dan learning fair representation, in-processing dengan metode prejudice remover dan adversarial debiasing, dan post-processing dengan metode calibrated equalized odds dan reject option classification [8].

Hasil deteksi bias dengan DI dan SPD sebelum dan sesudah melalui proses mitigasi akan dibandingkan untuk melihat metode yang paling efektif dalam melakukan mitigasi pada datset BRIlink. Penelitian ini dapat mengatasi bias pada data dan menghasilkan fairness prediksi untuk akuisisi agen BRIlink. Penelitian ini tentunya akan berbeda

dengan penelitian-penelitian terdahulu [7], [9]-[12] terkait bias and fairness, karena pada penelitian ini menggunakan dataset BRIlink dan kombinasi metode AIF 360 yang berbeda dengan penelitian sebelumnya.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, rumusan masalah penelitian ini adalah bagaimana mendeteksi bias yang dilakukan dengan metode DI dan SPD serta bagaimana melakukan mitigasi bias dengan metode reweighing, learning fair representation, prejudice remover, adversarial debiasing, reject option classification, dan calibrated equalized odds sehingga dapat menghasilkan data yang lebih fair.

C. Tujuan

Adapun tujuan dari penelitian untuk tugas akhir ini adalah melakukan deteksi bias dengan DI dan SPD serta menerapkan metode mitigasi bias yang ada pada tahap pre-processing, in-processing, dan post-processing dari AIF 360 untuk menghasilkan prediksi yang lebih fair dalam akuisisi agen. Hasil deteksi bias sebelum dan sesudah mitigasi akan dibandingkan untuk menemukan metode mitigasi bias yang paling efektif untuk mengatasi bias.

II. KAJIAN TEORI

A. Studi Terkait

Pada penelitian ini akan dilakukan deteksi dan mitiasi bias. Untuk mitigasi bias, terdapat protected attribute sebagai acuan fitur yang diduga menyebabkan bias. Untuk melihat seberapa fair data tersebut, dapat dilakukan perhitungan fairness atau fairness metrics. Fairness metrics yang digunakan pada penelitian ini adalah DI [9] dan SPD . Dalam penelitian Kozodoi N, et al. (2021), proses mitigasi bias dilakukan dengan tiga tahapan, yaitu pre processing, in-processing, dan post-processing. Pengukuran performa dihitung dengan AUC. Pada pre processing, reweighing mencapai keadilan terbaik tetapi menurunkan profitabilitas sebesar 23%. Pada in processing, prejudice remover menunjukkan kinerja terbaik dengan kenaikan AUC sebesar 0.37% dibandingkan dengan adversarial debiasing. Pada post processing didominasi oleh reject option classification dengan kenaikan independence 74.8% dan separation 74.55%.

B. BRIlink

BRIlink merupakan salah satu penerapan dari akuisisi agen penjualan perbankan yang dilakukan oleh BRI. BRIlink biasa ditemukan di kios-kios kecil seperti pada gambar 1 dengan tujuan untuk mempermudah para nasabah BRI untuk melakukan transaksi. Penelitian ini dilakukan untuk membantu BRI melakukan akuisisi agen secara adil.

GAMBAR 1
Agen BRIlink
Sumber : https://foto.bisnis.com/

C. Bias and Fairness

Istilah “Bias” dikenalkan oleh Mitchell (1980) yang berarti basis untuk memilih suatu generalisasi (hipotesis) atas individu atau grup lain tanpa memperhatikan konsistensi yang ketat dengan pelatihan yang diamati [13]. Oleh karena itu, bias harus ditangani agar dapat mencapai fairness seperti yang ditunjukkan pada gambar 2. Fairness merupakan keadaan dimana suatu data atau keputusan dianggap adil dan tidak terjadi diskriminasi baik antar individu atau kelompok.

GAMBAR 2
Ilustrasi bias and fairness

D. AI Fairness 360 Toolkit

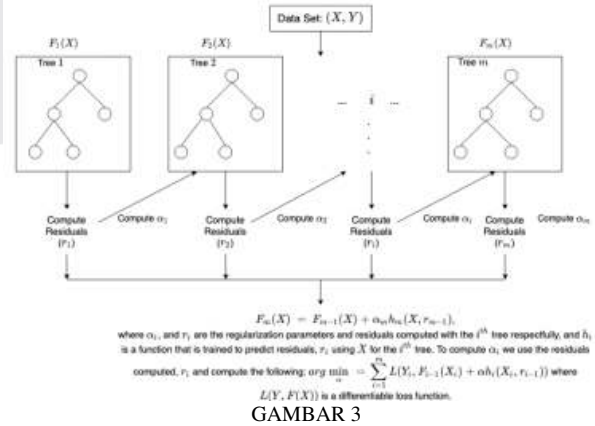
AI Fairness 360 (AIF 360) dikembangkan dan menjadi salah satu toolkit opensource yang dapat digunakan untuk mendeteksi, memahami, dan memitigasi bias pada algoritma [7]. Toolkit ini merupakan penggabungan toolkit bias and fairness sebelumnya, sehingga AIF 360 ini lebih baik dan lengkap. Pada jurnal [14] dilakukan analisis serta perbandingan model yang disediakan oleh AIF 360 dan hasilnya dapat membantu menghasilkan prediksi data yang fair.

E. Classifier pada tahap Pre-Processing dan Post-Processing

Untuk dataset latih akan dilakukan training model dengan 3 metode, yaitu:

1. XGBoost

XGBoost merupakan singkatan dari eXtreme Gradient Boosting. Algoritma ini dapat dilakukan untuk fungsi regresi, klasifikasi, maupun ranking. Package dari algoritma XGBoost ini dilengkapi beberapa fitur, seperti input type, speed, sparsity, customization, dan performance [15]. XGBoost merupakan teknik scalable machine learning yang menggunakan tree boosting untuk menghindari overfitting [16]. Cara kerja XGBoost dilampirkan pada gambar 3.



GAMBAR 3
cara kerja XGBoost (sagemaker XGBoost documentation)

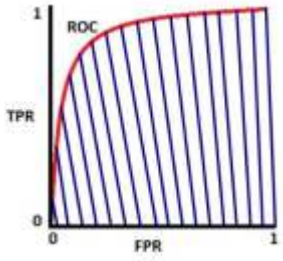
2. LightGBM

LightGBM adalah gradient boosting framework berdasarkan decision tree untuk meningkatkan efisiensi model dan mengurangi penggunaan memori. Algoritma ini didesain seefisien mungkin, dengan beberapa kelebihan, seperti : kecepatan pelatihan lebih cepat dan efisiensi lebih tinggi, penggunaan memori yang lebih rendah, akurasi yang lebih baik, dukungan pembelajaran paralel, terdistribusi, dan GPU, serta mampu menangani data berskala besar.

3. Random Forest

Algoritma random forest merupakan algoritma yang menggabungkan keluaran dari beberapa decision tree untuk mencapai satu hasil . Random forest termasuk akurat, tidak memerlukan feature scaling, categorical feature encoding, dan membutuhkan sedikit tuning parameter. Random forest sangat baik digunakan dalam klasifikasi ataupun regresi, deteksi outlier, pengelompokan, dan menafsirkan kumpulan data.

F. ROC dan AUC



GAMBAR 3
ROC dan AUC

Dalam mengukur performasi model dilakukan perhitungan ROC dan AUC. ROC biasa digunakan untuk memvisualisasikan performa dari klasifikasi biner. Kurva ROC merupakan suatu plot true positive rate (y) dan false positive rate (x) dalam setiap klasifikasi. Dalam kurva tersebut terdapat area dibawah kurva atau sering disebut dengan AUC (Area Under Curve) yang ditunjukkan dengan daerah arsiran biru pada gambar 3. AUC sendiri memiliki batas atau constraint untuk mengevaluasi model. Model yang memiliki AUC = 1 memiliki model yang baik, dan jika model memiliki AUC = 0, model dikatakan buruk. Untuk menghitung AUC, dapat digunakan rumus (1).

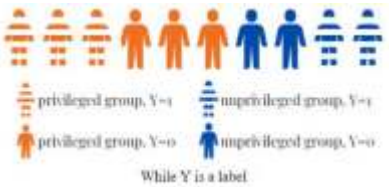
$$AUC = ROC - (\int_0^1 TPR(FPR)dFPR) \quad (1)$$

G. Bias Detection

Bias detection pada penelitian ini menerapkan dua metode, yaitu:

1. DI DI

(DI) merupakan salah satu pengukuran untuk melakukan evaluasi terhadap fairness. Konsep pada DI adalah membandingkan proporsi antara unprivileged dan privileged individu atau kelompok dari yang berlabel positif. Unprivileged adalah kelompok yang tidak diuntungkan dan privileged adalah kelompok yang diuntungkan dari protected attribute [9].



GAMBAR 4
Privileged dan unprivileged group

$$Dispare\ Impact = \Pr(Y = 1|D = unprivileged)\ \Pr(Y = 1|\ D = privileged) \quad (2)$$

Ilustrasi contoh sederhana dicantumkan pada gambar 4. Kalkulasi DI pada (2) memiliki rentang [0, ∞). Hasil kurang dari 0,8 menandakan bahwa kondisi privileged lebih diuntungkan daripada unprivileged, disebut dengan positive bias. Hasil lebih dari 1 dikatakan bahwa unprivileged lebih diuntungkan daripada privileged sendiri, hal ini mengacu pada kondisi negative bias.

2. SPD

SPD (SPD) adalah metrik untuk mengevaluasi keadilan yang menghilangkan proporsi individu penerima luaran positif untuk dua kelompok, yaitu: kelompok yang tidak memiliki hak istimewa (unprivileged) dan kelompok yang memiliki hak istimewa (privileged). Perhitungan SPD ditunjukkan pada (3).

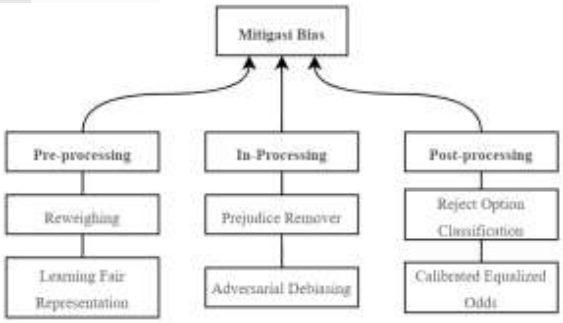
$$SPD = \Pr(Y = 1|D = unprivileged) - \Pr(Y = 1|\ D = privileged) \quad (3)$$

Pr pada (3) merupakan probabilitas dari unprivileged maupun privileged group. Hasil dari perhitungan rumus SPD dapat dikatakan fair jika hasil semakin mendekati 0.

H. Mitigasi Bias

Mitigasi bias adalah sebuah proses untuk menghilangkan bias yang tidak dinginkan dalam data. Pada mitigasi bias di AIF 360 dibagi menjadi 3 tahap, yaitu pre-processing, in-processing, dan post-processing seperti 5 pada Gambar 5.

Pre-processing merupakan tahap mitigasi bias sebelum model dilatih, sedangkan post-processing merupakan tahap mitigasi bias setelah model dilatih. Pada in-processing AIF 360, proses pelatihan model terjadi bersamaan dengan proses mitigasi bias.



GAMBAR 5
Flowchart mitigasi bias

I. Pre-processing

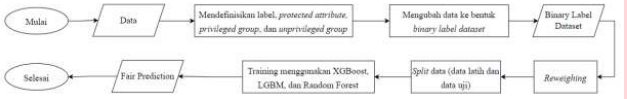
Dalam penelitian ini akan digunakan dua algoritma mitigasi bias pada pre-processing, yaitu:

1. Reweighing

Reweighing adalah teknik untuk memberi bobot yang berbeda di setiap kombinasi (kelompok, label) untuk memastikan keadilan sebelum klasifikasi [17]. Pada reweighing akan diasumsikan bahwa diskriminasi atau bias akan dihilangkan hingga 0 sambil mempertahankan probabilitas kelas positif. Untuk menghitung weight, dapat digunakan rumus sebagai berikut:

$$w(x) = Pexp\ (s{=}x(s) \wedge class{=}x(class))\ Pobs\ (s{=}x(s) \wedge class{=}x(class)) \quad (4)$$

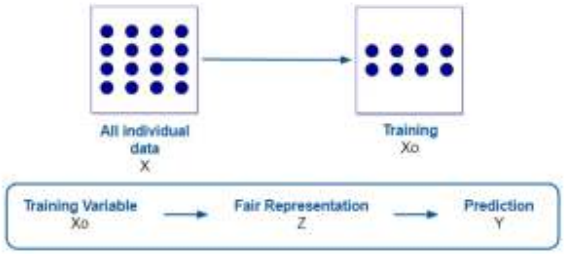
Pada rumus (4), Pexp menunjukkan probabilitas ekspetasi dan Pobs menunjukkan probabilitas observasi. Simbol s menujukkan protected attribute, atau atribut yang dilindungi, seperti gender, umur, dan fitur demografi lain. Sedangkan class merupakan label atau fitur luaran dari dataset tersebut. Gambar 6, merupakan alur proses terjadinya reweighing.



GAMBAR 6
Proses Reweighing pada Pre-processing.

2. Learning Fair Representations

Learning fair representations menemukan representasi laten yang mengkodekan data tetapi mengabaikan informasi tentang atribut yang dilindungi (protected attribute). Terdapat dua tujuan utama dari learning fair representation, yaitu group fairness dan individual fairness. Group fairness memastikan bahwa proporsi keseluruhan anggota dalam protected attribute yang menerima klasifikasi positif atau negatif identik dengan proporsi populasi secara keseluruhan. Individual fairness merupakan kondisi bahwa dua individu mana pun yang serupa harus diklasifikasikan bersama [18].
Pada proses learning fair representation data akan diambil sebagian sebagai data latih. Pada gambar 7, data latih (Xo) akan melalui proses fair representation dengan sehingga memiliki luaran Z. Selanjutnya Z akan melalui proses prediction sehingga terdapat luaran prediksi (Y) yang diekspetasikan Y merupakan hasil prediksi yang fair.



GAMBAR 7
Data dibagi menjadi data latih kemudian melakukan proses yang menghasilkan prediksi Y

Y adalah variabel biner (0/1) yang mewakili hasil klasifikasi individual sedangkan Z adalah variabel yang mewakili group fairness. Setelah itu data latih akan di pelajari sistem dengan rumus (5).

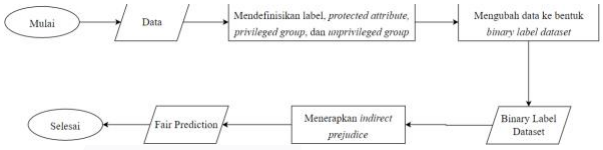
$$L = Az\ .\ Lz + Ax\ .\ Lx + Ay\ .\ Ly \quad (5)$$

Az,Az,Ay pada (5) merupakan parameter yang mengatur tradeoff dari sistem yang diinginkan. Hasil deteksi bias akan fair ketika hasil loss pada Lz, Lx, dan Ly semakin kecil. Proses terjadinya learning fair representation ditunjukan pada Gambar 8.

GAMBAR 8
Proses Learning Fair Representation pada Pre-processing.

3. In-processing

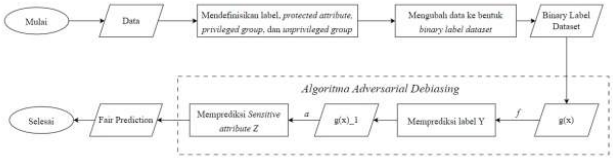
Adapun dua algoritma in-processing yang akan digunakan untuk penelitian ini, yaitu: a. Prejudice Remover Arti dari prejudice adalah prasangka. Pada penelitian ini diterapkan indirect prejudice. Indirect prejudice memberi hasil prediksi Y yang bergantung pada protected attribute. Dipilihnya indirect prejudice, karena dalam indirect prejudice menerapkan red lining effect (Bahwa mengabaikan fitur sensitif atau protected attribute tidak efektif). Pada prejudice remover fokus nuntuk melakukan klasifikasi dan pembentukan regularisasi dengan metode logistic regression [19]. Alur proses prejudice remover berjalan dapat dilihat pada Gambar 9.



GAMBAR 9
Flowchart proses mitigasi bias dengan prejudice remover

J. Adversarial Debiasing

Adversarial debiasing bergantung pada adversarial training untuk menghilangkan bias dari representasi laten yang dipelajari oleh model. Misalkan Z merupakan protected attribute yang ingin dicegah agar tidak terjadi diskriminasi, mis. usia atau ras. Menghapus Z tidak cukup, karena terkadang Z berkorelasi dengan fitur lain. Tujuannya adalah untuk mencegah model mempelajari representasi input yang bergantung pada Z. Untuk mencapai tujuan tersebut maka model dilatih yang secara bersamaan memprediksi label Y dan mencegah jointly trained adversary untuk memprediksi Z.



GAMBAR 10
Proses adversarial debiasing

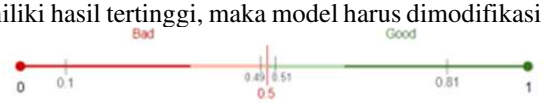
Idenya adalah jika model asli menghasilkan representasi X yang menyandikan informasi tentang Z (misalnya ras), model adversarial dapat dengan mudah memulihkan dan memprediksi Z menggunakan representasi. Oleh karena itu pada Gambar 10 data akan melalui dua proses f dan a. f merupakan fungsi prediksi dimana Y = f(g(x)). Prediksi hasil tergantung pada input data g(x), sedangkan a merupakan fungsi adversarial dimana Z = a(g(x)) (protected attribute diprediksi oleh fungsi adversarial).

1. Post-processing

Berikut merupakan algoritma post-processing yang akan digunakan pada penelitian ini:

a. Reject Option Classification

Pada reject option classification terdapat asumsi, bahwa diskriminasi atau bias terbanyak terjadi jika model memilki hasil prediksi mendekati dengan batas keputusan dari classification threshold. Sehingga, jika prediksi model memiliki hasil tertinggi, maka model harus dimodifikasi [20].



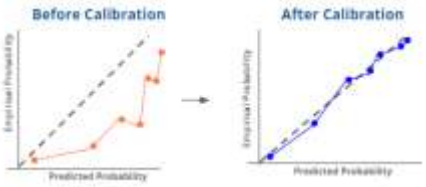
GAMBAR 11
Contoh hasil klasifikasi

Dilihat dari gambar 11, dengan batas klasifikasi 0,5, jika prediksi model adalah 0,81 atau 0,1, model memiliki hasil prediksi yang jelas (termasuk bad/good) tetapi untuk 0,51 atau 0,49, model tidak pasti terkait hasil prediksi yang dimilikinya. Dengan mengolah wilayah yang memiliki hasil prediksi rendah dari pengklasifikasi untuk pengurangan diskriminasi dan menolak prediksinya, diharapkan dapat mengurangi bias dalam prediksi model. Hasil prediksi yang di-reject akan masuk ke Critical region (rejected instances) yang dianggap ambigu dan dipengaruhi oleh bias. Alur proses reject option classification (ROC) dapat dilihat pada Gambar 12.

GAMBAR 12
Proses Reject Option Classification pada Post-processing

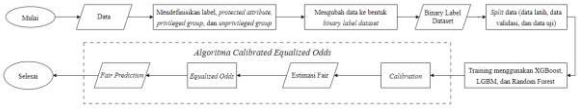
b. Calibrated Equalized Odds

Pada Calibrated Equalized Odds akan mengoptimalkan luaran skor hasil klasifikasi yang dikalibrasi, karena dalam aplikasi praktis, uncalibrated probabilitas dapat mengarahkan ke pemahaman yang salah [21].



GAMBAR 13
Probabilitas model sebelum dan sesudah dikalibrasi

Pada Gambar 13, diasumsikan bahwa garis abu-abu adalah garis dasar probabilitas yang ideal. Di sebelah kiri terdapat model yang belum dikalibrasi dan di sebelah kanan terdapat model yang sudah dikalibrasi. Harapannya, model yang dikalibrasi dapat mendekati hasil probabilitas kalibrasi yang ideal. Equalized odds merupakan keadaan dimana True Positive Rate (TPR) dan False Positive Rate (FPR) sama untuk protected attribute. Flowchart pada Gambar 14 menunjukkan proses dari metode calibrated equalized odds.

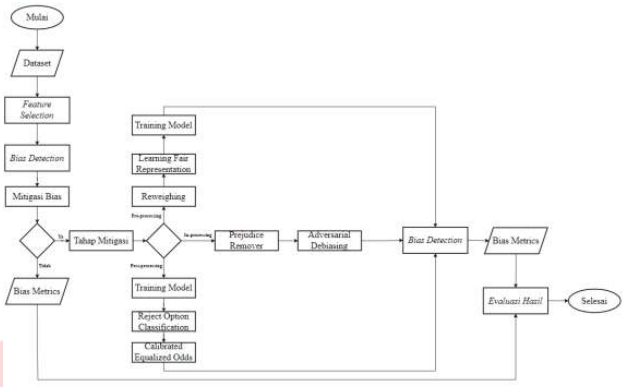


GAMBAR 14
Flowchart calibrated equalized odds.

III. METODE

A. Perancangan Sistem

Dalam melakukan deteksi bias menggunakan AIF 360, akan dilakukan beberapa tahapan seperti pada block diagram dibawah ini:

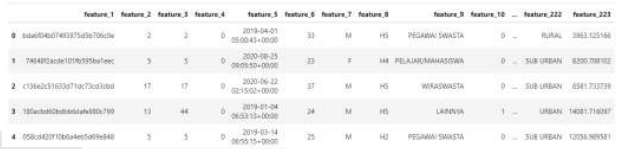


GAMBAR 15
Block diagram perancangan sistem

Pada Gambar 15 proses diawali dengan pengumpulan dataset setelah itu dilakukan pemilihan fitur (feature selection) untuk menentukan fitur yang akan digunakan dan fitur yang termasuk protected attribute. Setelah protected attribute ditentukan, dapat dilakukan deteksi bias untuk mengetahui bias awal sebelum mitigasi. Selanjutnya adalah proses mitigasi bias dengan menerapkan pre-processing, in-processing, dan post-processing. Deteksi bias kembali dilakukan setelah proses mitigasi untuk melihat keberhasilan mitigasi bias.

B. Pengumpulan Dataset

Data berasal dari BRI, yaitu dataBRIlink pada Gambar 16. Dalam dataset ini terdiri dari 30.000 data nasabah BRI yang berasal dari seluruh Indonesia dengan 230 fitur berdasarkan kurun waktu transaksi selama satu tahun. 9.



	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	...	feature_222	feature_223
0	bda6f04b07493875d5b706c0e	2	2	0	2019-04-01 05:00:43+00:00	33	M	HS	PEGAWAI SWASTA	0	...	RURAL	3963.125166
1	74048f2acde1016fb595ba1eec	5	5	0	2020-08-25 09:05:50+00:00	23	F	HI	PELAJAR/MAHASISWA	0	...	SUB URBAN	8200.788102
2	c1364e2c51633d71dc73cd3cbd	17	17	0	2020-06-22 02:15:02+00:00	37	M	HS	WIRASWASTA	0	...	SUB URBAN	8581.733739
3	100acbd60bdbb6daefe880c709	13	44	0	2019-01-04 06:53:13+00:00	24	M	HS	LAINNYA	1	...	URBAN	14081.716267
4	058cd42f010b0a4e5d69e848	5	5	0	2019-03-14 06:55:15+00:00	25	M	HI	PEGAWAI SWASTA	0	...	SUB URBAN	12058.989581

GAMBAR 16
Dataset BRIlink.

C. Seleksi Fitur

Beberapa fitur di dataset BRIlink merupakan fitur demografi, seperti : jenis kelamin, usia, pendidikan, pekerjaan, dan lokasi dapat dikategorikan dalam protected attribute. Karena fitur-fitur pada protected attribute memiliki banyak nilai yang berbeda, akan dilakukan pengelompokan ke dalam dua kategori saja, yaitu grup privilleged dan grup unprivileged dengan melakukan kalkulasi potential percentage pada (8).

$$Potential\ percentage = (Y{=}1\ |Unprivileged\ Individual) \\ Total\ Individual \text{ x } 100 \quad (8)$$

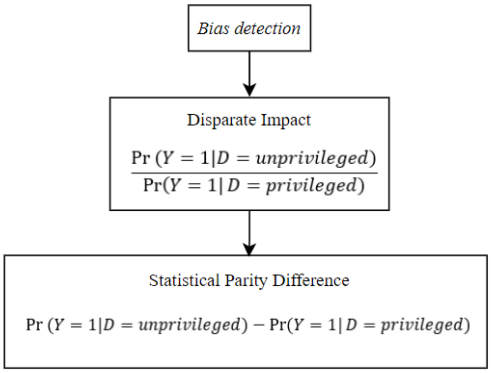
Pada fitur yang memiliki banyak nilai seperti fitur occupation, dilakukan perhitungan potential percentage seperti pada Tabel 1. Hasil dari setiap potential percentage pada Tabel 1 dihitung rata-rata potential percentage dan ditemukan hasil 53.34%. Dapat diasumsikan bahwa fitur

occupation dengan persentase ≥ 53,34% akan termasuk dalam privilleged dan yang memiliki persentase.

TABEL 1
Potential Percentage pada fitur occupation

Occupation	*Potential Percentage*
Wiraswasta	48.60%
Lainnya	49.88%
Pedagang	47.62%
Pegawai Swasta	55.76%
Ibu Rumah Tangga	46.40%
Pelajar/Mahasiswa	55.76%
Tidak Bekerja	43.89%
Pegawai BUMN	82.08%
Guru	52.88%
Pegawai Negeri Sipil	54.18%

D. Bias Detection



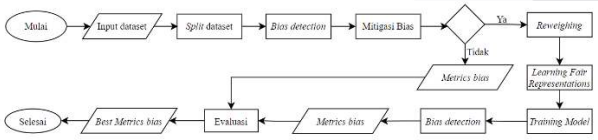
GAMBAR 1
Flowchart Bias Detection

Untuk melakukan deteksi bias, digunakan dua metode deteksi bias pada **GAMBAR 1**, yaitu *DI* dan *SPD*.

E. Mitigasi Bias

Mitigasi bias adalah proses mengurangi bias yang terdapat dalam data sehingga dapat menghasilkan prediksi yang *fair*. Mitigasi bias pada AIF 360 memiliki tiga tahap proses mitigasi, yaitu :

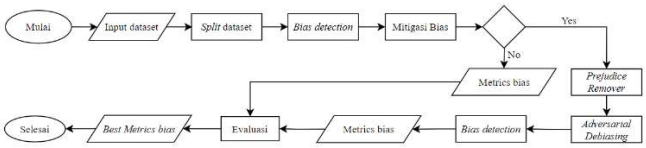
1. Pre-processing



GAMBAR 2
Flowchart Pre-Processing pada mitigasi bias

Proses dilakukannya *pre-processing* terdapat pada GAMBAR 2. Selanjutnya akan masuk pada tahap *pre-processing* dengan menerapkan metode *reweighing* dan *learning fair representations.* Pada tahap *pre-processing*, *training model* dilakukan setelah mitigasi bias.

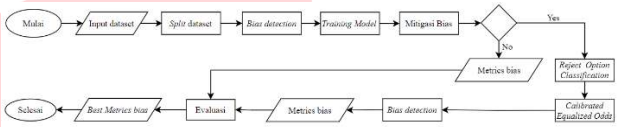
2. In-Processing



GAMBAR 3
Flowchart in-processing pada mitigasi bias

Dalam tahapan *in-processing* pada GAMBAR 3, *training model* dilakukan saat data sedang di mitigasi. Metode yang dilakukan dalam *in-processing* ini adalah *prejudice remover* dan *adversarial debiasing.*

3. Post-processing



GAMBAR 4
Flowchart post-processing pada mitigasi bias

Tahap *post-processing* pada GAMBAR 4 melakukan *training model* sebelum data melakukan mitigasi. Metode yang dilakukan dalam *post-processing* ini adalah *calibrated equalized odds* dan *reject option classifications.*

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengujian

Eksperimen mitigasi bias pada dataset BRIlink dilakukan sebanyak 3 tahapan, yaitu *pada pre-processing, in-processing*, dan *post-processing*. Hasil pengukuran bias berupa perhitungan dari DI dan SPD, serta dilakukan perhitungan AUC untuk melihat performasi model. Pada *baseline* dilakukan pengukuran di setiap *protected attribute* dengan menggunakan *classifier* XGBoost, LightGBM, dan random forest yang ada pada TABEL 1.

TABEL 1
Baseline untuk setiap fitur dan *classifiers*

Features	XGBoost			LightGBM			Random Forest		
	AUC	DI	SPD	AUC	DI	SPD	AUC	DI	SPD
Age	0.639	1.047	0.024	0.647	0.989	-0.006	0.625	0.945	-0.029
Location	0.639	1.002	0.001	0.647	1.004	0.002	0.622	1.017	0.009
Gender	0.639	0.986	-0.007	0.647	0.995	-0.003	0.625	0.965	-0.019
Occupation	**0.639**	1.037	0.019	0.647	**1.056**	0.029	0.622	1.038	0.020
Education	0.639	0.981	-0.010	0.647	1.014	0.007	0.623	1.027	0.014

TABEL 2
Rata-rata deteksi bias pada *baseline*

Features	DI	SPD
Age	0.99%	0.020%
Location	1,008%	0,004%
Gender	0.982%	0.010%

Occupation	1.043%	0.023%
Education	1.007%	0.010%

Pada TABEL 1, fitur *occupation* memiliki nilai DI 1.037 untuk XGBoost, 1.056 untuk lightGBM, dan 1.038 untuk random forest. Nilai DI lebih dari satu menandakan data tersebut merupakan bias negatif. Pada TABEL 2 disajikan rata-rata dari setiap metode deteksi bias pada XGBoost, lightGBM, dan random forest. Untuk nilai SPD pada fitur *occupation* memiliki rata-rata 0.023% dimana nilai ini lebih besar dan paling jauh dari 0 daripada rata-rata nilai yang dimiliki fitur lain. Oleh karena itu, fitur *occupation* merupakan *protected attribute* yang paling bias, sehingga fitur *occupation* akan diterapkan menjadi *protected attribute* pada eksperimen selanjutnya.

TABEL 3
Hasil mitigasi bias pada *pre-processing*

Pre-processing	XGBoost			LightGBM			Random Forest		
	AUC	DI	SPD	AUC	DI	SPD	AUC	DI	SPD
Original Data	0.639	1.037	0.019	0.647	1.056	0.029	0.622	1.038	0.020
Reweighing	**0.498**	**0.802**	**-0.102**	**0.497**	**0.800**	**-0.105**	**0.490**	**0.804**	**-0.098**
Learning Fair Representation	0.498	0.689	-0.005	0.500	1.369	0.006	0.500	0.799	-0.001

Berdasarkan TABEL 3, algoritma *reweighing* sama baiknya dalam mengurangi bias untuk metode DI. Hasil DI pada XGBoost adalah 0.802, pada lightGBM adalah 0.800, dan pada random forest adalah 0.804. Hal ini menunjukkan mitigasi bias dengan *reweighing* berhasil, karena hasil DI mendekati 0.8 yang dapat dikatakan *fair*. Pada SPD, bias mengalami peningkatan, tetapi masih mendekati 0. AUC *score* mengalami penurunan, dan model yang memiliki AUC paling tinggi adalah XGBoost.

Algoritma *learning fair representation* tidak menunjukkan adanya mitigasi bias dalam metode DI. Hal ini ditandai dengan nilai DI yang awalnya 1.037 menjadi 0.689 pada XGBoost yang berada di bawah 0.8 (batas *fair* untuk DI). Untuk SPD pada *learning fair representation* terjadi sedikit pengurangan bias. AUC score yang ada pada *learning fair representation* mengalami penurunan sebesar 1-1,5%.

TABEL 4
Hasil mitigasi bias pada *in-processing* dengan metode *prejudice remover*

In-Processing	AUC	DI	SPD
Baseline (Original Data dengan *Logistic Regression)*	0.586	1.001	0.001
Prejudice Remover	0.586	0.668	-0.189

TABEL 5
Hasil mitigasi bias pada *in-processing* dengan metode *adversarial debiasing*

In-Processing	AUC	DI	SPD
*Baseline (*Tanpa *Debiasing)*	0.629	0.770	-0.149
Adversarial Debiasing	**0.632**	**0.963**	**-0.018**

TABEL 4 dan TABEL 5 akan dibandingkan untuk melihat metode *in-processing* yang lebih baik. Mitigasi bias tidak berhasil dalam algoritma *prejudice remover* untuk DI maupun metode SPD. *Debiasing* merupakan proses pengurangan bias. Algoritma *adversarial debiasing* memiliki parameter *debiasing* yang dapat diatur menjadi *true/false*. Algoritma *adversarial debiasing* berhasil memitigasi bias dengan baik yang dihitung oleh DI maupun SPD. DI yang awalnya memiliki nilai 0.770 berubah menjadi 0.963 setelah mitigasi bias, menandakan model mengalami pengurangan bias. Pada *SPD*, nilai berkurang dari -0.149 menjadi -0.018 setelah proses mitigasi bias. Nilai SPD tersebut semakin mendekati dengan 0 yang menandakan semakin *fair*. *Adversarial debiasing* juga memiliki performa model yang baik (AUC meningkat).

TABEL 6
Hasil mitigasi bias pada *post-processing*

Post-processing	XGBoost			LightGBM			Random Forest		
	AUC	DI	SPD	AUC	DI	SPD	AUC	DI	SPD
Original Data	0.639	1.037	0.019	0.647	1.056	0.029	0.622	1.038	0.020
Reject Option Classification	**0.637**	**0.931**	**-0.302**	**0.643**	**0.899**	**-0.053**	**0.623**	**0.942**	**-0.029**
Calibrated Equalized Odds	0.638	0.651	-0.190	0.643	0.654	-0.194	0.625	0.633	-0.200

Reject option classification berhasil memitigasi bias. Nilai DI pada lightGBM memiliki penurunan bias secara signifikan dari 1.056 (*baseline*) menjadi 0.899 yang dapat dikatakan mendekati batas *fair* 0.8. Meskipun demikian, hasil SPD meningkat terutama pada XGBoost sebesar 0.283%. Ada sedikit peningkatan 0.001% untuk Skor AUC pada random forest.

Algoritma *Calibrated Equalized Odds* gagal menangani bias baik yang dihitung dengan DI maupun SPD. Pada rata-rata nilai DI, model menjadi semakin bias dengan nilai 0.6, sedangkan untuk SPD juga mengalami kenaikan untuk ketiga classifiers sebanyak 0.200.

B. Analisis Hasil Pengujian

Berdasarkan hasil eksperimen sebelumnya, didapatkan bahwa metode yang berhasil melakukan mitigasi bias adalah *reweighing, adversarial debiasing,* dan *reject option classifier* yang ada pada TABEL 7.

TABEL 7
Hasil analisis pengujian

Algoritma Mitigasi		Classiffier	Berhasil Melakukan Mitigasi Bias? (Ya/Tidak)
Pre-Processing	***Reweighing***	**XGBoost**	**Ya**
		LightGBM	**Ya**
		Random Forest	**Ya**

	Learning Fair Representation	XGBoost	Tidak
		LightGBM	Tidak
		Random Forest	Tidak
In-Processing	*Prejudice Remover*	-	Tidak
	Adversarial Debiasing	**-**	**Ya**
Post-Processing	***Reject Option Classification***	**XGBoost**	**Ya**
		LightGBM	**Ya**
		Random Forest	**Ya**
	Calibrated Equalized Odds	XGBoost	Tidak
		LightGBM	Tidak
		Random Forest	Tidak

Pada tahap *pre-processing*, algoritma *reweighing* efektif untuk mengurangi bias tetapi mengurangi skor AUC. Pada XGBoost menurunkan 0,235% nilai DI, untuk lightGBM menurunkan 0,201% nilai DI, dan untuk random forest menurunkan 0, 256% nilai DI. Hal ini membawa model ke batas yang adil. SPD juga meningkat, tetapi masih mendekati 0. Untuk metode *learning fair representation* ini gagal dalam *learning*, karena hasil *loss* Lx, Ly, dan Lz tidak semakin kecil, hal ini menandakan bahwa tidak tercapainya kondisi *fair*.

Prejudice remover, prasangka yang diterapkan dalam data gagal menghasilkan perkiraan yang adil dan menyebabkan hasil mitigasi bias yang buruk. Sedangkan adversarial debiasing mengurangi bias cukup efektif, DI meningkat dari 77% menjadi 96%, serta SPD tetap mendekati 0. Meningkatkan skor AUC dari 0,629 menjadi 0,632.

Pada *post-processing*, algoritma *reject option classification* dan *calibrated equalized odds* cukup baik dalam mempertahankan nilai AUC. *Reject option classification* berhasil memitigasi bias pada DI meskipun meningkatkan hasil SPD. Pada *reject option classification* mengalami sedikit pengurangan nilai AUC untuk xgboost dan lightGBM, tetapi nilai AUC meningkat untuk random forest. Sedangkan pada *calibrated equalized odds* menghasilkan mitigasi bias yang buruk.

V. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, saat melakukan mitigasi, terjadi *tradeoff* antara performasi model dengan model yang *fair*, hal ini dibuktikan dengan menurunnya AUC *score* pada model. Pada *pre-processing*, *reweighing* menunjukkan performa yang lebih baik untuk mitigasi bias daripada *learning fair representation* dengan hasil pengukuran DI sebesar 0.802% dan SPD sebesar -0.102% (XGBoost), DI sebesar 0.8% dan SPD sebesar -0.105% (lightGBM), dan DI sebesar 0.804 % dan SPD sebesar -0.098% (random forest). Saat *fairness* berhasil tercapai pada metode *reweighing*, AUC score mengalami penurunan sebesar 0,2% di setiap *classifier*. Pada *in-processing*, *Adversarial debiasing* sukses melakukan mitigasi bias ditandai dengan hasil DI yang awalnya 0.770% (dibawah 0.8%) menjadi 0.963% dan SPD sebesar -0.018%. Metode *adversarial debiasing* secara bersamaan berhasil mempertahankan performa model yang ditunjukkan dengan adanya kenaikan AUC *score* dari 0.629% menjadi 0.632%. Pada *post-processing, Reject Option Classification* menunjukkan performa yang lebih baik daripada *calibrated equalized odds* dalam mitigasi bias ditunjukkan dengan hasil pengukuran DI sebesar 0.931% dan SPD sebesar -0.302% pada XGBoost, DI sebesar 0.899% dan SPD sebesar -0.053% pada lightGBM, dan DI sebesar 0.942% dan SPD sebesar -0.098% pada random forest. *Reject Option Classification* hanya sedikit mengurangi AUC *score* sebesar -0.029% pada *classifier* XGBoost dan LightGBM tetapi mengalami kenaikan 0.001% pada *classifier* random forest.

Dengan tercapainya *fairness* pada dataset brilink dapat membantu menentukan agen mana yang dapat diakuisisi dengan mempertimbangkan fitur yang ada dan dapat memperoleh agen brilink dengan lebih merata serta mengaburkan fitur demografis yang ada. Untuk penelitian kedepannya, dapat mengembangkan algoritma mitigasi yang juga dapat menangani kumpulan data yang tidak seimbang sehingga meningkatkan keadilan sambil mempertahankan performa model itu sendiri

REFERENSI

[1] J. Buolamwini, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *,” 2018.

[2] J. M. Zhang and M. Harman, “‘Ignorance and Prejudice’ in software fairness,” in *Proceedings - International Conference on Software Engineering*, May 2021, pp. 1436–1447. https://doi.org/10.1109/ICSE43902.2021.00129.

[3] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, “Fairness under unawareness: Assessing disparity when protected class is unobserved,” in *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Jan. 2019, pp. 339–348. https://doi.org/10.1145/3287560.3287594.

[4] K. Martinus and B. Reilly, “To boundary or not: The structural bias of ‘fair representation’ in rural areas,” *J Rural Stud*, vol. 79, pp. 136–144, Oct. 2020, https://doi.org/10.1016/j.jrurstud.2020.08.039.

[5] L. Doornkamp, L. D. van der Pol, S. Groeneveld, J. Mesman, J. J. Endendijk, and M. G. Groeneveld, “Understanding gender bias in teachers’ grading: The role of gender stereotypical beliefs,” *Teach Teach Educ*, vol. 118, Oct. 2022, https://doi.org/10.1016/j.tate.2022.103826.

[6] T. Burch, “Skin Color and the Criminal Justice System: Beyond Black-White Disparities in Sentencing,” Journal of Empirical Legal Studies, vol. 12, no. 3, pp. 395–420, Sep. 2015, doi: 10.1111/jels.12077.

[7] R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.

[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.09635

[9] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing DI,” Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.3756

[10] S. Raza, “A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission,” *Healthcare Analytics*, vol. 2,

p. 100100, Nov. 2022, https://doi.org/10.1016/j.health.2022.100100.

[11] P. Mosteiro, J. Kuiper, J. Masthoff, F. Scheepers, and M. Spruit, "Bias Discovery in Machine Learning Models for Mental Health," *Information (Switzerland)*, vol. 13, no. 5, May 2022, https://doi.org/10.3390/info13050237.

[12] P. Cerrato, J. Halamka, and M. Pencina, "A proposal for developing a platform that evaluates algorithmic equity and accuracy," *BMJ Health and Care Informatics*, vol. 29, no. 1. BMJ Publishing Group, Apr. 11, 2022. https://doi.org/10.1136/bmjhci-2021-100423.

[13] T. M. Mitchell, "The Need for Biases in Learning Generalizations by The Need for Biases in Learning Generalizations," 1980.

[14] D. O. Blessed and L. Liu, "How do the Existing Fairness Metrics and Unfairness Mitigation Algorithms contribute to Ethical Learning Analytics? Identification of miRNA sponge network and modules in human cancers View project Estimating heterogeneous treatment effects by balancing heterogeneity and fitness View project", doi: 10.13140/RG.2.2.20988.67204.

[15] T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," 2022.

[16] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Machine Learning with Applications*, vol. 6, p. 100154, Dec. 2021, doi: 10.1016/j.mlwa.2021.100154.

[17] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl Inf Syst*, vol. 33, no. 1, pp. 1–33, 2012, doi: 10.1007/s10115-011-0463-8.

[18] R. Zemel, Y. (Ledell,) Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," 2013.

[19] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "LNAI 7524 - Fairness-Aware Classifier with Prejudice Remover Regularizer." [Online]. Available: http://www.kamishima.net

[20] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2012, pp. 924–929. doi: 10.1109/ICDM.2012.45.

[21] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration."