Seminar Keinsinyuran 2024
ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

Deteksi Bias dalam Model Machine Learning untuk Prediksi Penyakit Cardiovascular Menggunakan DALEX

Setio Basuki[1], Ali Mokhtar[2]
Universitas Muhammadiyah Malang
[1,2] Program Profesi Insinyur, Universitas Muhammadiyah Malang, Jl. Raya Tlogomas 246 Malang

Kontak Person:
Setio Basuki
Jl. Raya Tlogomas 246 Malang
E-mail: setio_basuki@umm.ac.id

Abstrak

Penelitian ini bertujuan untuk deteksi potensi bias pada model machine learning (ML) untuk dataset medis penyakit cardiovascular. Penelitian ini didasarkan pada fakta dimana ML telah banyak diadopsi untuk berbagai macam skenario prediksi, termasuk pada bidang medis. Mengingat bahwa bidang medis merupakan bidang yang critical karena berkaitan dengan manusia, maka penerapan ML perlu akurat dan fair. Tahapan penelitian ini dimulai dari identifikasi data medis penyakit cardiovascular yang paling popular di Kaggle, identifikasi protected attribute seperti jenis kelamin, usia, etnis, ras, dsb., pembangunan model klasifikasi ML menggunakan Random Forest, identifikasi potensi bias pada setiap dataset medis. Deteksi bias dilakukan berbasis Model Agnostic Language for Exploration and Explanation (DALEX) dengan terlebih dahulu mengidentifikasi protected attribute seperti jenis kelamin, usia, ras, etnis, dsb. Hasil pengujian mengungkap bahwa walaupun menghasilkan akurasi klasifikasi yang tinggi, bahkan mencapai 98% seperti pada dataset MIMIC-III dan Heart Diseases, masih ditemukan atribut yang berpotensi bias. Bias ini muncul umumnya pada atribut yang terkait dengan jenis kelamin dan usia. Dari 11 dataset cardiovascular, terdapat dua dataset yang tidak berpotensi bias yaitu Heart Diseased dan Cardiovascular Risk Factor. Hasil penelitian ini menguatkan bahwa penggunaan ML dalam bidang medis, khususnya penyakit cardiovascular, perlu mempertimbangkan aspek fairness pada model ML agar hasil prediksi yang dihasilkan dapat dipertanggung jawabkan.

Kata kunci: *bias, penyakit cardiovascular, dalex, fairness, protected attribute.*

1. PENDAHULUAN

Perkembangan metode Machine Learning (ML) yang semakin canggih menjadi penyebab luasnya adopsi teknologi ini di berbagai bidang. Salah satu bidang yang menunjukan tren adopsi ML yang signifikan adalah bidang medis [1], [2], dimana ML dimanfaatkan secara luas mulai dari membantu proses diagnosis penyakit sampai dengan intervensi metode perawatan pasien yang tepat. Secara global, investasi teknologi ML, atau secara umum Artificial Intelligence (AI), pada bidang kesehatan diproyeksikan akan meningkat secara signifikan hingga 164 miliar dollar pada tahun 2030, dengan laju pertumbuhan per-tahun sebesar 49,1% dari perkiraan 14,92 miliar dollar sejak tahun 2024 [3]. Dalam dunia kesehatan, penerapan ML harus dilakukan dengan sangat hati-hati untuk memastikan bahwa hasil yang diperoleh bukan hanya akurat, namun juga harus adil tanpa diskriminasi. Dalam proses pembangunan model ML, kualitas dataset memagang peran yang krusial. Secara lebih spesifik, dataset yang digunakan haruslah bebas dari unsur bias terhadap protected attribute (atribut yang melekat pada personal yang tidak bisa dipilih) seperti jenis kelamin, usia, etnis, ras, dsb. Keberadaan bias pada atribut ini dapat menyebabkan adanya diskriminasi dan ketidakadilan pada hasil prediksi model ML yang dibangun, bahkan membahayakan pasien [4], [5]. Dengan demikian, keberadaan suatu metode yang dapat mendeteksi adanya potensi bias pada protected attribute menjadi hal yang wajib. Sistem cerdas di bidang medis yang bebas bias tidak hanya memastikan akurasi dan keadilan prediksi, tetapi juga meningkatkan kepercayaan berbagai pihak, termasuk tenaga medis, pengambil kebijakan, dan pasien, terhadap integritas dalam pengambilan keputusan [6], [7].

Implementasi Machine Learning (ML) dalam prediksi penyakit kardiovaskular telah menjadi fokus utama, mengingat penyakit ini merupakan salah satu penyebab utama kematian di seluruh dunia, dan kematian jantung mendadak atau sudden cardiac death (SCD) menyumbang 4-5 juta kematian secara global dan bertanggung jawab atas lebih dari 50% dari seluruh kematian terkait penyakit kardiovaskular [8]. Berbagai penelitian terkini telah menunjukkan potensi ML dalam meningkatkan

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

akurasi prediksi risiko penyakit jantung [9], [10], [11], dengan algoritma yang mampu menganalisis data pasien dan faktor risiko dengan lebih efisien. Namun demikian, penelitian terkait topik ini masih memerlukan jaminan atas keadilan (fairness) yang terepresentasi pada dataset medis yang digunakan. Mengingat kualitas sistem cerdas berbasis ML ditentukan oleh kualitas dataset medis yang digunakan, dibutuhkan teknik-teknik untuk menjamin integritas data tersebut. Hal ini menegaskan perlu adanya studi lanjutan lebih lanjut yang berfokus pada deteksi bias dalam dataset-dataset kardiovaskular yang populer, agar model yang dihasilkan tidak hanya akurat tetapi juga adil dan dapat diandalkan untuk semua kelompok populasi.

Penelitian ini bertujuan untuk mendeteksi potensi bias pada dataset penyakit cardiovaskular untuk membangun model ML. Untuk tujuan tersebut, penelitian ini mengumpulkan 11 dataset paling populer dari Kaggle[1], yang semuanya memiliki atribut terlindungi seperti jenis kelamin, ras, dan etnis, yang dapat mempengaruhi keadilan hasil prediksi. Tahapan penelitian ini mencakup beberapa tahapan, yaitu: pertama, pengambilan dataset cardiovascular yang relevan, yang diikuti dengan data preprocessing-transformasi untuk memastikan kualitas dan konsistensi. Selanjutnya, kami akan membangun model ML berbasis algoritma Random Forest [12]. Selanjutnya, proses deteksi bias menggunakan Model Agnostic Language for Exploration and Explanation (DALEX) [13] dalam dilakukan untuk setiap dataset. Tahap akhir penelitian ini adalah analisis hasil deteksi bias untuk memberikan wawasan tentang potensi unfairness dalam model prediksi yang dihasilkan. Hasil akhir penelitian ini diharapkan penelitian ini dapat memberikan kontribusi terhadap pemahaman dan mitigasi bias dalam aplikasi ML di bidang kesehatan.

Secara lebih spesifik, beberapa temuan penting yang dihasilkan dari penelitian ini adalah sebagai berikut:

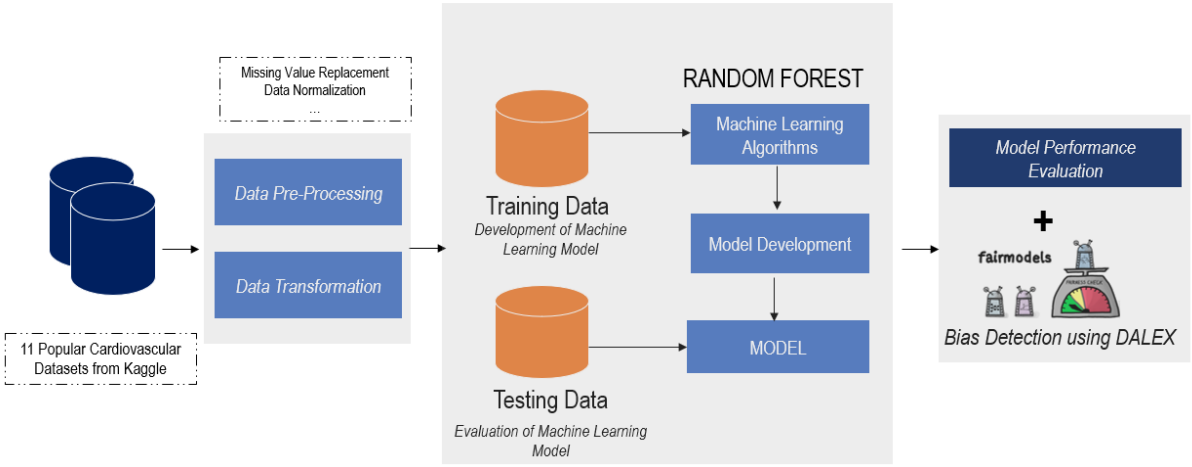
- Algoritma Random Forest menghasilkan performasi klasifikasi yang kompetitif, dimana akurasi tertinggi mencapai 98,86% pada dataset Heart Diseases – 1.
- Dari 11 dataset yang dipilih, hanya dua dataset yang tidak memiliki potensi bias pada protected attributenya, yaitu Heart Diseases -2 dan Cardiovascular Risk Factor.
- Potensi bias terjadi pada atribut jenis kelamin dan usia, serta ras dan keyakinan keagamaan pada sebagian kecil dataset.
- Indikator fairness yang paling sering terjadi pada 9 dataset yang memiliki potensi bias adalah Specificity atau STP.
- Potensi bias terjadi pada atribut jenis kelamin dan usia, serta ras dan keyakinan keagamaan pada sebagian kecil dataset.

2. METODE PENELITIAN

Pada bagian ini paparkan tahapan deteksi potensi bias pada dataset medis terkait dengan penyakit cardiovascular. Pada tahap pertama dilakukan identifikasi dataset yang sesuai dengan kriteria pada repository Kaggle. Selanjutnya, dilakukan proses data preprocessing dan transformasi atas daftar dataset terpilih. Di tahap ketiga dilakukan proses pembangunan model klasifikasi ML berbasis algoritma Random Forest. Terakhir, dilakukan deteksi potensi bias pada setiap dataset. Tahapan penelitian ini disajikan pada Gambar 1.

[1] https://www.kaggle.com/

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775



Gambar 1 Alur kerja sistem deteksi bias pada model machine learning untuk prediksi penyakit Cardiovascular

2.1 Akuisisi Data Cardiovascular

Penelitian ini menggunakan dataset bidang cardiovascular yang paling popular di Kaggle. Dataset tersebut diidentifikasi dengan menggunakan kata kunci "cardiovascular" dengan mempertimbangkan popularitasnya. Selanjutnya, dilakukan evaluasi terkait dengan ketersediaan protected attribute pada beberapa kandidat dataset. Protected attribute merupakan atribut personal yang tidak boleh mendapatkan diskriminasi secara hukum seperti jenis kelamin, usia, status disabilitas, ras, etnis, dsb. Dari proses ini dihasilkan 11 dataset tentang penyakit cardiovascular yang ditunjukan pada Table 1.

Tabel 1 Daftar 11 Dataset Medis Populer terkait Penyakit Cardiovascular pada Kaggle

No.	Nama Dataset	Jumlah Data	Jumlah Atribut
1	Cardiovascular Disease – 1	70.000	13
2	Cardiovascular Disease – 2	68.205	17
3	Cardiovascular Disease – 3	1.000	14
4	Heart Disease Comprehensive	1.190	12
5	Heart Disease – 1	1.319	9
6	Heart Disease – 2	918	12
7	Cardiovascular Disease – 4	2.000	13
8	Cardiovascular Risk Factor Data	3.390	17
9	Cardiovascular Disease Risk Prediction	30.8855	19
10	CVD Risk	42.5195	13
11	MIMIC-III	23.468	39

2.2 Model Agnostic Language for Exploration and Explanation (DALEX)

DALEX, singkatan dari Model Agnostic Language for Exploration and Explanation adalah kerangka kerja analisis dan interpretasi model machine learning yang agnostik terhadap jenis model. Dengan melihat hasil dari indicator fairness seperti TPR, ACC, FPR, STP, dan PPV (untuk kasus klasifikasi) kita dapat melihat apakah terdapat bias dalam sebuah model. Dalam analisis fairness menggunakan Dalex, terdapat epsilon guna untuk memberi batasan nilai dari setiap indikator dapat dikatakan bias atau tidak, rentang **0.8 - 1.25 epsilon** diterapkan untuk menilai apakah suatu model prediktif menunjukkan bias terhadap kelompok tertentu. Salah satu dasar utama untuk nilai epsilon 0.8 adalah aturan hukum yang dikenal sebagai 80% Rule atau four-fifths rule, yang digunakan secara luas dalam undang-undang diskriminasi di Amerika Serikat [14]. Berdasarkan aturan ini, suatu metode atau

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

proses pemilihan dianggap adil jika kelompok yang kurang diuntungkan masih memiliki minimal 80% dari peluang kelompok yang lebih diuntungkan dalam mencapai hasil positif, yang merupakan tolok ukur umum untuk menentukan ketimpangan secara praktis. Konsep Disparate Impact juga berkaitan dengan penetapan nilai epsilon, dalam konteks machine learning dan analisis fairness, Disparate Impact didefinisikan sebagai perbandingan rasio antara probabilitas hasil positif pada kelompok yang kurang diuntungkan dan kelompok yang lebih diuntungkan. Nilai epsilon ini juga berkaitan erat dengan konsep Statistical Parity dan Demographic Parity, di mana fairness didefinisikan sebagai kesamaan peluang untuk mendapatkan hasil positif antar kelompok [14].

Skor dari setiap metrik didefinisikan pada formula (1) berikut ini [15]:

$$\forall_{i\in\{a,b,\ldots,z\}^{\varepsilon}} < \frac{metric_i}{metric_{privileged}} < \frac{1}{\varepsilon}. \quad (1)$$

Dalam menilai tingkat fairness dari model klasifikasi, DALEX menggunakan beberapa indicator sebagai berikut:

- *TPR (True Positive Rate)*
  Dikenal juga sebagai sensitivity atau recall, TPR mengukur kemampuan model untuk mendeteksi kelas positif dengan benar. Nilai TPR yang tinggi mengindikasikan bahwa model secara efektif mengidentifikasi kasus positif tanpa bias.
- *ACC (Accuracy)*
  Akurasi adalah tingkat keseluruhan kebenaran model, yang dihitung sebagai rasio antara instance yang diprediksi dengan benar (baik positif maupun negatif) terhadap total instance.
- *PPV (Positive Predictive Value)*
  Proporsi prediksi positif yang benar di antara semua prediksi positif, yang menunjukkan keandalan prediksi positif.
- *FPR (False Positive Rate)*
  Rasio instance positif yang salah diprediksi terhadap semua instance negatif aktual. Nilai FPR yang rendah menunjukkan lebih sedikit false positives di seluruh kelompok.
- *STP (Specificity)*
  Mengukur TPR yang dikondisikan pada kelompok tertentu, membantu mengidentifikasi bias dalam kinerja model di berbagai demografi.

3. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil pengujian dari sistem deteksi potensi bias pada model ML untuk prediksi penyakit carsiovascular. Hasil pengujian tersebut terbagi menjadi dua sub-bagian yaitu kinerja model ML untuk klasifikasi penyakit dan deteksi potensi bias pada protected attribute di setiap dataset. Selain itu, pada bagian dipaparkan analisis terhadap hasil deteksi potensi bias.

3.1 Kinerja Model Machine Learning untuk Klasifikasi

Pada bagian ini disajikan hasil klasifikasi 11 dataset cardiovascular. Secara umum, algoritma Random Forest menghasilkan performasi klasifikasi yang kompetitif. Hal ini ditunjukan dengan lima dataset menghasilkan akurasi lebih dari 90%, bahkan heart diseases - 1 dan MIMIC-III mencapai 98%. Selain itu, dua dataset lain yang menghasilkan akurasi diatas 85% seperti pada heart diseases – 2 dan CVD risk data. Namun demikian, performa prediksi menghasilkan performa yang cukup pada tiga dataset yang lain dengan akurasi diatas 70%. Detail perbandingan kinerja algoritma klasifikasi untuk setiap dataset disajikan pada Tabel 2.

Tabel 2 Perbandingan akurasi klasifikasi dataset cardiovascular

No.	Nama Dataset	Accuracy (%)
1	Cardiovascular Disease – 1	70,55
2	Cardiovascular Disease – 2	70,99

3	Cardiovascular Disease – 3	98
4	Heart Disease Comprehensive	94,11
5	Heart Disease – 1	98,86
6	Heart Disease – 2	85,86
7	Cardiovascular Disease – 4	73
8	Cardiovascular Risk Factor Data	92,47
9	Cardiovascular Disease Risk Prediction	93
10	CVD Risk	88,06
11	MIMIC-III	98,18

3.2 Deteksi Fairness Setiap Dataset

Jika pada bagian sebelumnya disajikan performasi model ML untuk klasifikasi, di bagian ini disajikan hasil deteksi bias spesifik untuk setiap dataset. Detail nilai indicator fairness dipaparkan dalam bentuk Tabel 3 – 13. Nilai NaN pada semua metrik kecuali akurasi (ACC) menunjukkan bahwa data untuk subgrup ini sangat tidak seimbang, atau mungkin tidak ada data yang cukup untuk menghitung nilai indikator terkait. Hal ini dapat terjadi karena dua sebab, yaitu (1) Data tidak seimbang untuk subgrup ini sehingga tidak terdapat kasus positif atau negatif yang tersedia untuk perhitungan; dan (2) Data kurang bervariasi untuk menganalisis indikator fairness tertentu.

Dataset Name : Cardiovascular Disease - 1
Protected Attribute : Age, Gender

Tabel 3 Indikator fairness Cardiovascular Disease - 1

Subgrup	TPR	ACC	PPV	FPR	STP
female_over50	1.180223	0.882586	0.978962	**2.672956**	**1.779412**
female_under50	0.992026	1.01847	0.99439	0.90566	0.929412
male_over50	1.185008	0.89314	0.978962	**2.553459**	**1.747059**

Dataset Name : Cardiovascular Disease - 2
Protected Attribute : Age, Gender

Tabel 4 Indikator fairness Cardiovascular Disease - 2

Subgrup	TPR	ACC	PPV	FPR	STP
female_over50	**1.256803**	0.897878	0.987288	**2.753425**	**1.891026**
female_under50	1.076531	1.035809	1.007062	0.952055	1.003205
male_over50	1.241497	0.897878	0.991525	**2.664384**	**1.855769**

Dataset Name : Cardiovascular Disease - 3
Protected Attribute : Age, Gender

Tabel 5 Indikator fairness Cardiovascular Disease - 3

Subgrup	TPR	ACC	PPV	FPR	STP
female_over50	**1.256803**	0.897878	0.987288	**2.753425**	**1.891026**
female_under50	1.076531	1.035809	1.007062	0.952055	1.003205
male_over50	1.241497	0.897878	0.991525	**2.664384**	**1.855769**

Dataset Name : Heart Disease Comprehensive
Protected Attribute : Age, Gender

Tabel 6 Indikator fairness Heart Disease Comprehensive

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	**1.6**	1.067227	0.8	**NaN**	**1.664671**

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

female_young	**NaN**	1.20048	**NaN**	**NaN**	**NaN**
male_old	**1.97**	1.145258	0.949	**NaN**	**4.239521**

Dataset Name : Heart Disease – 1
Protected Attribute : Age, Gender

Tabel 7 Indikator fairness Heart Disease – 1

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	0.975	0.975	0.975	**NaN**	1.070664
female_young	1	1	1	**NaN**	**0.428266**
male_old	1	0.994	0.991	**NaN**	**1.488223**

Dataset Name : Heart Disease – 2
Protected Attribute : Age, Gender

Tabel 8 Indikator fairness Heart Disease – 2

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	**NaN**	0.921	**NaN**	**NaN**	**NaN**
female_young	**NaN**	1	**NaN**	**NaN**	**NaN**
male_old	**NaN**	0.837	**NaN**	**NaN**	**NaN**

Dataset Name : Cardiovascular Disease – 4
Protected Attribute : Age, Gender

Tabel 9 Indikator fairness Cardiovascular Disease – 4

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	1.241379	0.950667	1.145427	**2.046784**	**1.807927**
female_young	1.095238	1.085333	1	**0.754386**	0.85061
male_old	**1.338259**	0.893333	0.985007	**2.923977**	**2.042683**

Dataset Name : Cardiovascular Risk Factor
Protected Attribute : Age, Gender

Tabel 10 Indikator fairness Cardiovascular Risk Factor

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	**NaN**	0.93	**NaN**	**NaN**	**NaN**
female_young	**NaN**	1	**NaN**	**NaN**	**NaN**
male_old	**NaN**	0.916	**NaN**	**NaN**	**NaN**

Dataset Name : Cardiovascular Disease Risk Prediction
Protected Attribute : Age, Gender

Tabel 11 Indikator fairness Cardiovascular Disease Risk Prediction

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	**1.294199**	0.945289	0.938693	**NaN**	**15.294118**
female_young	1.040055	0.997974	0.99196	**NaN**	**1.294118**
male_old	**1.288674**	0.928065	0.927638	**NaN**	**17.264706**

Dataset Name : CVD Risk
Protected Attribute : Age, Gender, Race

Tabel 12 Indikator fairness CVD Risk

Subgrup	TPR	ACC	PPV	FPR	STP
female_old	**1.340909**	0.905363	0.978236	**20.285714**	**7.111111**

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

female_young	1.004261	0.996845	0.983963	1.214286	1.060606
male_old	**1.333807**	0.888538	0.965636	**24.071429**	**7.40404**

Subgrup	TPR	ACC	PPV	FPR	STP
female_amind	1.075388	1.034325	1.058894	**1.295302**	**1.344262**
female_asian/pl	1.047672	1.011442	1.038462	**1.315436**	**1.27459**
female_black/afam	1.063193	0.997712	1.050481	**2.315436**	**1.604588**
female_white	1.013084	1.012586	1.015625	0.939597	1.008197
male_amind	1.087583	1.040046	1.055288	**1.308725**	**1.346311**
male_asian/pl	1.056541	1.017162	1.038462	**1.302013**	**1.272541**
male_black/afam	1.062084	0.993135	1.040865	**2.288591**	**1.579918**

Dataset Name : MIMIC-III
Protected Attribute : Gender, Religion, Ethnicity

Tabel 13 Indikator fairness MIMIC-III

Subgrup	TPR	ACC	PPV	FPR	STP
female_black/african american	1.124859	1.090513	1	**NaN**	**1.49925**
female_white	1.124859	1.090513	1	**NaN**	**1.49925**
male_hispanic/latino - puerto rican	1.124859	1.090513	1	**NaN**	0.526237

Subgrup	TPR	ACC	PPV	FPR	STP
female_chatolic	1.055966	1.022495	1	**NaN**	**2.557545**
female_jewish	1.055966	1.022495	1	**NaN**	**2.557545**
female_not specified	1.055966	1.022495	1	**NaN**	**2.557545**
female_protestant quarter	1.055966	1.022495	1	**NaN**	**2.557545**
female_unobtainable	1.055966	1.022495	1	**NaN**	**2.557545**
male_christian scientist	1.055966	1.022495	1	**NaN**	**2.557545**
male_not specified	1.055966	1.022495	1	**NaN**	**2.557545**
male_protestant quarter	1.055966	1.022495	1	**NaN**	**2.557545**

3.3 Deteksi Fairness Setiap Dataset

Ringkasan dataset yang disertai dengan daftar protected attribute dan indikator fairness yang mensinyalkan adanya potensi bias disajikan pada Table 14 dibawah ini. Dari 11 dataset yang dipilih, hanya dua dataset yang tidak memiliki potensi bias pada protected attributenya, yaitu heart diseases -2 dan cardiovascular risk factor. Selebihnya, ditemukan potensi bias pada protectec attributenya dengan jumlah indicator fairness yang bervariasi. Sebagai contoh, potensi bias pada dataset heart disease – 1 dan MIMIC-III hanya terdeteksi pada satu indicator yaitu STP, dataset Cardiovascular Disease – 1, Cardiovascular Disease – 3, Heart Disease Comprehensive, Cardiovascular Disease Risk Prediction terdeteksi pada dua fairness indicator. Jumlah indikator fairness terbanyak (tiga indikator) dideteksi pada dataset Cardiovascular Disease – 2, Cardiovascular Dataset – 4, dan CVD Risk.

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

Pola lain yang dapat dideteksi terkait dengan atribut yang menimbulkan bias adalah jenis kelamin dan usia yang ditemukan pada sembilan dataset. Namun demikian, pada dua dataset yaitu CVD Risk dan MIMIC-III dimana ditemukan atribut berpotensi bias berupa ras dan keyakinan dalam beragama. Penting diketahui bahwa atribut-atribut yang berpotensi bias tidak dapat dibandingkan antar dataset, karena kemunculan atribut tersebut tidak merata di setiap dataset.

Tabel 14 Ringkasan dataset, daftar atribut yang berpotensi bias, dan indikator fairness yang terdeteksi bias.

No	Nama Dataset	Potential Bias	Indikator Fairness
1	Cardiovascular Disease – 1	female_over50, male_over50	FPR, STP
2	Cardiovascular Disease – 2	female_over50, male_over50	TPR, FPR, STP
3	Cardiovascular Disease – 3	female_old, female_young, male_old	FPR, STP
4	Heart Disease Comprehensive	female_old, male_old	TPR, STP
5	Heart Disease – 1	male_old, female_young	STP
6	Heart Disease – 2	Tidak berpotensi bias	-
7	Cardiovascular Dataset – 4	female_over50, male_over50	TPR, FPR, STP
8	Cardiovascular Risk Factor	Tidak berpotensi bias	-
9	Cardiovascular Disease Risk Prediction	female_old, female_young, male_old	TPR, STP
10	CVD Risk	female_old, male_old, female_AmInd, female_Asian/Pl, female_Black/AfAm, male_AmInd, male_Asian/Pl, male_Black/AfAm	TPR, FPR, STP
11	MIMIC-III	female_chatolic, female_jewish, female_not specified, female_protestant quarter, female_unobtainable, male_christian scientist, male_not specified, male_protestant quarter, female_black/african american, female_white	STP

4. KESIMPULAN

Penelitian ini telah mengimplementasikan metode deteksi potensi bias pada model machine learning (ML) untuk klasifikasi penyakit cardiovascular. Pada umumnya, bias disebabkan oleh keberadaan protected attribute seperti jenis kelamin, usia, status disabilitas, ras, etnis, dsb. Daftar atribut tersebut melekat pada personal yang bersifat bawaan yang tidak bisa dipilih oleh setiap individu, dan rentang terhadap diskriminasi. Dalam konteks pembangunan sistem prediksi berbasis ML, potensi bias pada atribut tersebut perlu dideteksi, guna menjamin model prediksi dapat dipertanggung jawabkan.

Penelitian ini menggunakan library DALEX dalam deteksi potensi bias pada model ML yang dibangun dengan algoritma Random Forest. Hasil eksperiment menunjukan beberapa temuan penting yaitu: (i) meskipun performasi model ML yang dihasilkan sangat kompetitif pada beberapa dataset, namun masih berpotensi terjadinya bias pada pengambilan keputusan, (ii) dari 11 dataset cardiovascular, hanya dua dataset saja yang tidak terindikasi adanya bias pada protected attributenya, (iii) potensi bias terjadi pada atribut jenis kelamin dan usia, serta ras dan keyakinan keagamaan pada sebagian kecil dataset, dan (iv) Indikator fairness yang paling sering terjadi pada sembilan dataset yang memiliki potensi bias adalah Specificity atau STP.

REFERENSI

[1] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, 2022, doi: https://doi.org/10.1016/j.ijin.2022.05.002.

[2] C. Chakraborty, M. Bhattacharya, S. Pal, and S.-S. Lee, "From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare," *Curr Res Biotechnol*, vol. 7, p. 100164, 2024, doi: https://doi.org/10.1016/j.crbiot.2023.100164.

[3] Marketsandmarkets, "Artificial Intelligence (AI) in Healthcare Market: Growth, Size, Share, and Trends." Accessed: Dec. 24, 2024. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html

[4] S. Raza, A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya, "Exploring Bias and Prediction Metrics to Characterise the Fairness of Machine Learning for Equity-Centered Public Health Decision-Making: A Narrative Review," *IEEE Access*, vol. 12, pp. 180815–180829, 2024, doi: 10.1109/ACCESS.2024.3509353.

[5] V. Azimi and M. A. Zaydman, "Optimizing Equity: Working towards Fair Machine Learning Algorithms in Laboratory Medicine," *J Appl Lab Med*, vol. 8, no. 1, pp. 113–128, Jan. 2023, doi: 10.1093/jalm/jfac085.

[6] I. D. Mienye, G. Obaido, I. D. Emmanuel, and A. A. Ajani, "A Survey of Bias and Fairness in Healthcare AI," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 2024, pp. 642–650. doi: 10.1109/ICHI61247.2024.00103.

[7] M. A. K. Akhtar, M. Kumar, and A. Nayyar, "Socially Responsible Applications of Explainable AI," in *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities*, M. A. K. Akhtar, M. Kumar, and A. Nayyar, Eds., Cham: Springer Nature Switzerland, 2024, pp. 261–350. doi: 10.1007/978-3-031-66489-2_9.

[8] J. Pires Da Silva, P. A. Padilla, and A. M. Garcia, "Editorial: The intersection of gene regulation and metabolism in cardiovascular disease," *Front Genet*, vol. 14, 2023, doi: 10.3389/fgene.2023.1253690.

[9] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[10] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction," *Sensors*, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031161.

[11] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, Apr. 2023, doi: 10.3390/pr11041210.

[12] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, and P. Biecek, "dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python," 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1473.html.

[14] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *ACM Comput. Surv.*, vol. 56, no. 7, Apr. 2024, doi: 10.1145/3616865.

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

[15] J. W. Wi´sniewski and P. Biecek, “fairmodels: a Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models,” *R J*, vol. 14, no. 1, Mar. 2022.

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

LAMPIRAN

Berikut ini merupakan URL sumber dari 11 dataset terkait cardiovascular yang dilengkapi dengan daftar atribut/fitur.

Tabel 15. Daftar 11 dataset cardiovascular dan URL sumbernya.

No	Nama Dataset	URL Dataset	Daftar Atribut
1	Cardiovascular Disease – 1	https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset	Age, Height, Weight, Gender, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose, Smoking, Alcohol intake, Physical activity, Presence or absence of cardiovascular disease
2	Cardiovascular Disease – 2	https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease	ID, age, age_years, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke alco, active, cardio, bmi, bp_category, bp_category_encoded
3	Cardiovascular Disease – 3	https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset	Patient Identification Number, Age, Gender, Resting blood pressure, Serum cholesterol, Fasting blood sugar, Chest pain type, Resting electrocardiogram results, Maximum heart rate achieved, Exercise induced angina, Oldpeak, Slope of the peak exercise ST segment, Number of major vessels, Classification (target)
4	Heart Disease Comprehensive	https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final	age, sex, chest pain type, resting bp s, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise angina, oldpeak, ST slope, target
5	Heart Disease – 1	https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset	age, gender, impluse, pressurehight, pressurelow, glucose, kcm, troponin, class
6	Heart Disease – 2	https://www.kaggle.com/datasets/ronanazarias/heart-desease-dataset	Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope
7	Cardiovascular Disease – 4	https://www.kaggle.com/datasets/harshthakkar05092002/prediction-of-cardiovascular-disease	id, Age, Gender, Height, Weight, SBP, DBP, Cholesterol, Glucose, Smoking, Alcohol_intake, Active, Cardiovascular
8	Cardiovascular Risk Factor Data	https://www.kaggle.com/datasets/mamta1999/cardiovascular-risk-data	id, age, education, sex, is_smoking, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, TenYearCHD
9	Cardiovascular Disease Risk Prediction	https://www.kaggle.com/datasets/harshwardhanfartale/cardiovascular-disease-risk-prediction-dataset	General_Health, Checkup, Exercise, Heart_Disease, Skin_Cancer, Other_Cancer, Depression, Diabetes, Arthritis, Sex,

ISSN (Cetak) 2798-0405
eISSN (Online) 2797-1775

			Age_Category, Height_(cm), Weight_(kg), BMI, Smoking_History, Alcohol_Consumption, Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consu
10	CVD Risk	https://github.com/laderast/cvdRiskData/tree/master/data-raw	patientID, age, htn, treat, smoking, race, t2d, gender, numAge, bmi, tchol ,sbp, cvd
11	MIMIC-III	https://www.kaggle.com/datasets/chidozieuzoegwu/cvd-vital-signs	subject_id, icustay_id, heart_rate, blood_pressure, oxygen_saturation, respiratory_rate, temperature, Label
		https://www.kaggle.com/datasets/asjad99/mimiciii?resource=download	row_id, subject_id, hadm_id, admittime, dischtime, deathtime, admission_type, admission_location, discharge_location, insurance, language, religion, marital_status, ethnicity, edregtime, edouttime, diagnosis, hospital_expire_flag, has_chartevents_data