機械学習におけるモデルベースのバイアス軽減に向けて

アルファ・ヨハニス ヨーク大学 イギリス・ヨーク
alfa.yohannis@york.ac.uk

ディミトリス・コロヴォス ヨーク大学 ヨーク（英国）
dimitris.kolovos@york.ac.uk

ABSTRACT

機械学習によって生成されたモデルは、特に差別的な環境で生成されたデータでトレーニングやテストを行った場合、バイアスがないとは保証されない。バイアスは、主にデータに性別、人種、年齢などのセンシティブな属性が含まれている場合、非倫理的である可能性がある。いくつかのアプローチは、バイアスメトリクスや軽減アルゴリズムを提供することで、このようなバイアスを軽減することに貢献している。課題は、ユーザが一般的／統計的なプログラミング言語でコードを実装しなければならないことであり、プログラミングや機械学習における公平性の経験が少ないユーザにとっては厳しいものとなる。我々は、ソフトウェア開発の労力を削減し、バイアスの測定と緩和を容易にするモデルベースのアプローチであるFairMLを発表する。我々の評価では、FairMLは、ベースラインの行動を分類し、ローンのためのプロファイルを分類する。機械学習は

コードによって生成されたものと同等の測定値を生成するために、より少ないコード行数で済むことを示している。

CCSのコンセプト

• **計算の理論** → *学習のモデル*; - **応用計算** → **法学、社会科学、行動科学**; - **ソフトウェアとその工学** → **ソースコード生成**; **ドメイン固有言語**.

キーワード

モデル駆動工学, 生成計画法, バイアス軽減, バイアス測定, 機械学習

ACMリファレンスフォーマット:

アルファ・ヨハニス、ディミトリス・コロボス2022.機械学習におけるモデルベースのバイアス緩和に向けて。*ACM/IEEE 25th International Conference on Model Driven Engineering Languages and Systems (MODELS '22), October 23-28, 2022, Montreal, QC, Canada.*ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3550355.3552401

1　はじめに

認証のための顔認識やインテリジェントアシスタント（アレクサやシリなど）と話す際の音声処理といった個人的な日常活動から、犯罪の予測といった繊細で倫理的なタスクの実行に至るまで、機械学習の利用は今日、広く浸透している。

MODELS '22、2022年10月23～28日、カナダ、QC州モントリオール

ACM ISBN 978-1-4503-9466-6/22/10.. . $15.00
https://doi.org/10.1145/3550355.3552401

機械学習は効率性をもたらすが、機械学習によって生成されたモデルは、特に差別的な環境で生成されたデータで訓練されテストされた場合、バイアスがないとは保証されない。これは、主に性別、人種、年齢などのセンシティブな属性に触れ、不公平さを増幅させる場合、容認できないことがある。

2016年、再犯を予測するアルゴリズムであるCOMPASSは、白人に対しては高い偽陰性率を、黒人に対しては高い偽陽性率を生み出すことが判明した[5]。また、市販の顔認識サービスの中には、肌の黒い女性に対する精度が著しく低いものもあることが判明した[11]。さらに、ある求人プラットフォームは、同じような性質を持つにもかかわらず、適格な女性候補者を適格な男性候補者よりもはるかに低くランク付けすることが判明した[33]。これらは、機械学習におけるバイアスがいかに不公平を助長するかを示すいくつかの事例である。

いくつかのアプローチは、バイアスメトリクスとデビアスアルゴリズムを提供することで、このようなバイアスの軽減に貢献してきた（詳細はセクション2.2と2.3で）。様々なツールキットがこれらのメトリクスやアルゴリズムを実装している。しかし、それらは異なる方法論的アプローチと機能を備えており、特定のシナリオに最適なツールキットを決定する前に、ユーザはそれらを深く理解する必要がある[34]。

データサイエンティストは通常、与えられた目標、データセット、ドメインに最適なモデルを見つけるために、アルゴリズム、パラメータ、その他の要因の組み合わせの数を絞り込むために直感を働かせている[37]。その後、多くの実験と試行錯誤[12]を繰り返しながら、絞り込まれたすべての組み合わせを調べ、生成されたモデルをテストして、どのモデルが最良かを特定しなければならない。さらに、機械学習ライブラリの有無にかかわらず、データサイエンティストは、一般的／統計的プログラミング言語（Python、Rなど）でゼロから検索プロセスを構築しなければならない。

モデル駆動型ソフトウェア開発（MDSE）は、自動化可能な実装の技術的な詳細を隠蔽す ることによって、ソフトウェア開発の負担を軽減するものである[10]。したがって、ユーザーは、より単純なモデリング言語を使用することで、本質的な側面に集中することができ、ターゲットとなる実装は自動的に生成されるため、生産性が向上する[47]。MDSEを使用することで、データサイエンティストは、より高い抽象化レベルで、与えられたケースに最適なバイアス緩和方法を検索することができるため、メリットがある。また、探索プロセスの実装を一般的/統計的プログラミング言語でコーディングする必要もない。
を自動生成し、後で微調整する。

本稿では、機械学習におけるバイアスの測定と緩和をモデル化し自動化するMDSEアプローチを実装したツールFairML[1]を紹介する。

[1] FairMLのプロトタイプはhttps://doi.org/10.5281/zenodo.7007839、httpsで見ることができる:
//github.com/York-and-Maastricht-Data-Science-Group/fairml

(1) FairMLはバイアスの測定と緩和を抽象化することで、ユーザーが一般的／統計的プログラミング言語でコーディングすることなく、人間に優しい宣言言語であるYAML（YAML Ain't Markup Language）[19]でバイアス緩和モデルを設定できるようにする。
(2) FairMLはある程度の表現力をサポートしており、ユーザーは様々な種類のバイアスメトリクス、バイアスを軽減するアルゴリズム、データセット、分類器、およびそれらのパラメータを実験し、バイアスを軽減しつつも許容できる精度を維持する最適な組み合わせを見つけることができます。
(3) サポートするツールはPythonとJupyter Notebookファイルを自動的に生成し、ユーザーは与えられたデータセットのバイアスを測定し、緩和するために実行することができる。生成されたファイルはすべて、微調整やさらなる開発のために変更および拡張が可能です。

本稿の構成は以下の通りである。まずセクション2で機械学習における公平性について述べる。このセクションでは、いくつかの関連用語の定義、実際の例、バイアスメトリクス、バイアスの緩和について説明する。セクション3では、我々の開発の概要と、バイアス緩和を自動化するために機械学習分野でデビアス技術を実装するアプローチを紹介する。またこのセクションでは、分類におけるバイアスを緩和するソリューションを表現する際に、ユーザが既存のツールキットをどのように利用するかを紹介する。また、FairMLがより読みやすく簡潔な方法で解決策を表現し、手作業で作成された解決策と同様の正しさで、バイアスの測定と緩和コードを自動的に生成できることを示す。セクション4では、既存のバイアス緩和ツールキットのコードをベンチマークとして、表現力、正しさ、生成時間、実行時間についてFairMLを評価する。セクション4.1では評価結果を示し、議論する。また、セクション5では、FairMLの研究開発中に学んだいくつかの教訓を振り返る。セクション6では関連する研究について、セクション7では本研究の結論と今後の課題について述べる。

2　機械学習におけるバイアス

このセクションでは、機械学習における公平性について簡単に説明し、いくつかの関連用語の定義、実際の例、バイアスメトリクス、バイアスの軽減について取り上げる。

2.1　定義と例

公平性とは、「個人または集団が本来持っているまたは後天的に獲得した特 徴に基づいて、その個人または集団に対して偏見やえこひいきがないこと」 [36] と定義されている。公平性の欠如は、データの収集と処理、研究デザイン、分析、解釈における欠陥のために、実際の状態から系統的な誤りや歪曲であるバイアスによって引き起こされる可能性がある［38］。不公正は、バイアスが特権的な集団を特権的でない集団に対して有利な立場に置くときに起こる [7]。バイアスはまた、区別が意図的または非意図的なステレオタイプや、敏感な属性（人種、年齢、

しかし、集団の公平性は、必ずしも個人にとっての公平性を示さない。さらに、有名なベンダーが提供する一般に利用可能な商用顔認識オンライン・サービスは、肌の色が濃い女性に対してはるかに低い精度を達成することに苦しんでいることが判明している[11]。機械学習におけるバイアスの例は他にもあるが、これら3 つの例は、実際には機械学習が必ずしも公平ではなく、その不公平さが特定のグループに不利益をもたらす可能性があることを示している。

2.2 バイアス指標

機械学習におけるバイアスを検出し測定するために、いくつかのメトリクスが開発されている。各測定基準にはバイアスを計算する方法があり、したがって特定の状況においては他のものよりも望ましい。IBM AI Fairness 360 [26, 35] とAequitas[2]は、図1に要約された選択のためのガイダンスを提供している。したがって、バイアス測定が集団と個人の公平性を包含する測定基準を必要とする場合、Theil Index [7, 16] とGeneralised Entropy Index [44] が望ましい。これらは、グループと個人の利益配分における不平等を測定する統一的な指標として使用することができる[26, 35]。スコアが低いほど強い公平性を反映し、高いほどその逆を示す。完全な公平性は 0 で示される [25]。

ユークリッド距離（Euclidean Distance ）、マンハッタン距離（Manhattan Distance ）、マハラノビス距離（Mahalanobis Dis-tance ）は、個人の公平性を測定する指標である。これらの測定基準は、元のデータセットと偏向されたデータセットで、同じ個人の距離を測定するために使用される[7]。前処理でバイアス緩和を適用する際にデータ変換を行うと、グループの公平性は達成できるが、個人に不公平が生じる可能性があるため、一貫性を確保するために必要である。13] では、グループの公平性と個人の公平性の両方を達成できるように、バイアス緩和の歪みを制限するための制約として、これらのメトリクスが使用されている。

グループの公平性には、主に2 つの世界観がある：平等な公平性と部分的な公平性 [26, 35]。平等な公平性とは、十分に代表されていないグループが、予測結果に関して他のグループと同様の機会を持つべきであることを意味し、通常、構造的な差別があると認識されているシステムに適用される。例えば、SAT は構造的差別を含んでいると疑われている [26, 35] 。統計的パリティ差[18, 26, 35] と格差影響[20, 26, 35]は、公平性を測定するための一般的な指標である。

統計的パリティ差は、特権グループと非特権グループの間で有利なラベルが貼られる確率の差として計算される（式1）[4, 7, 18]。

$$SPD = Pr(\square = \text{unprivate}) - P^{\alpha}(\square = 1|D = \text{priv}) \quad (1)$$

このメトリックの理想的な値は0 である。負 の値は、ラベリングが特権グループに有利であることを意 味し、正の値はその反対を意味する[4, 7]。格差の影響[4, 7, 20] は、特権のないグループと特権のあるグループの間で、有利にラベリングされる確率の比率として公平性を計算する（式2）。

$$\square\square = \square\square \frac{Pr(\square = 1|D = \text{特権なし})}{P(\square = 1|D = \text{特権})} \quad (2)$$

このメトリックの理想的な値は1.0である。値 <1は、ラベリングが特権階級に有利であることを示し、値 >1は非特権階級に不利であることを示す[4, 7]。

性など） [15, 36]。

例えば、2006年には、あるアルゴリズムが再犯率予測では、黒人の偽陽性率が白人よりもはるかに高い[5]。別の例として、ある求人プラットフォームでは、男性の求職者の方がより適格でないことが判明している。より有能な女性候補者よりも高い[33]。同論文は次のように述べている。女性グループと男性グループの間の公平性が達成されていること、

[2] http://aequitas.dssg.io/static/images/metrictree.png

バイアス・メトリック・ツリー

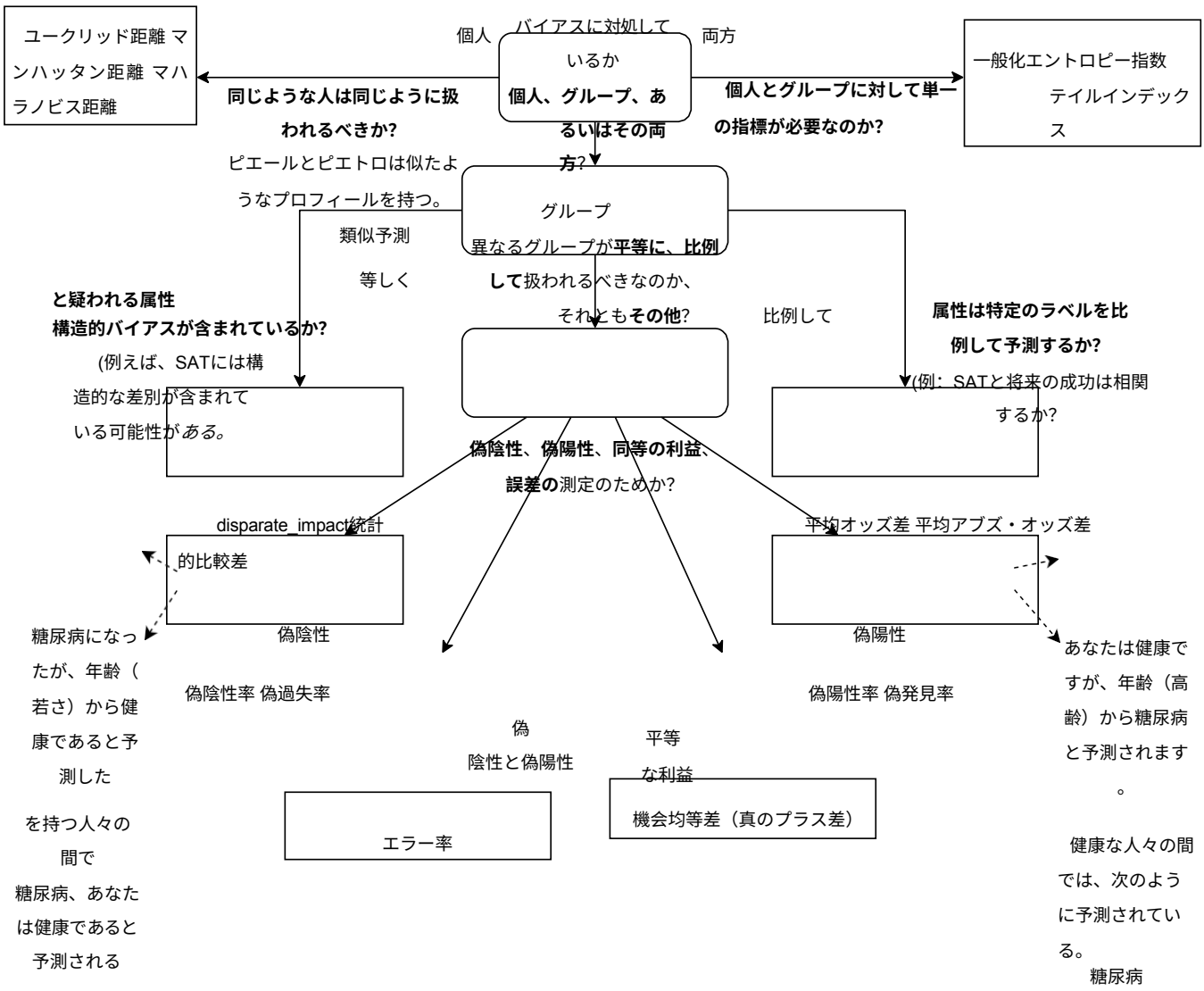


図1：特定の状況に対して最も適切なバイアス測定基準を自動的に選択するための決定木。

対照的に、WYSIWYG（What you see is what you get）とも呼ばれる比例的公平性は、特定の特徴が特定のラベルと相関していると認識する。例えば、SATの得点は将来の所得と相関する[26, 35]。平均オッズ差は以下を測定するための指標である。

比例公平性。この指標は、偽陽性率（□ □ □ = □/□ , 偽陽性 / 全陰性）と真陽性率（□= □ □/□ , 真陽性 / 全陰性）の平均差として偏りを計算する。
すべての陽性）は、恵まれないグループと恵まれたグループの間にある（式
3) [4, 7].

$$A = 1* ((_{F=\text{特権なし}} - \square_{=\text{特権あり}})+)$$

真陽性率とは、あるグループの全陽性結果の総数（□ = □/□ ）に対する真陽性の比率である。理想値は0 であり、負の値は嗜好性が高いことを示す。

一方、正はその逆である [4, 7] 。この分野の他の測定基準には、False Negative Rate Ratio and Difference 、False Omission Rate Ratio and Difference 、Error Rate 、False Positive Rate Ratio and Difference 、False Discovery Rate Ratio and Difference がある [26, 35]。

$$(□_{=特権なし} - T_{=特権あり})) \quad (3)$$

メトリックの理想的な値は0である。
特権階級は非特権階級より有利であり、正の値はその逆を示す[4, 7]。

また、均等公正と比例公正の中間に位置する測定基準もある。例えば、Equal Opportunity Difference [4, 7]は、バイアスを特権のないグループと特権のあるグループの間の真の陽性率の差として計算するメトリックである（式4）。

$$EO\ D = PT - □\ T = R - □ D = RD = D \quad (4)$$

2.3 バイアスの緩和

ここでは、マのバイアスを減らすために開発されてきたアルゴリズムは、

機械学習は、機械学習パイプラインで適用される段階に基づいて分類することができる、

インプロセス、ポストプロセス。26, 35]の研究は、特定の状況に対して最も適切なデビアスアルゴリズムを選択するためのガイダンスを提供している。このガイダンスは図2に要約されている。

参考文献[35]では、できるだけ早い段階で、特に前処理の段階で、不公正さを軽減することを推奨している。なぜなら、その段階で軽減されるバイアスは、データセットに内在しているからである。残念ながら、データセットの改変が禁止されている場合もある。そのため、データセットの改変が禁止されている場合は、Reweighing [27]を利用することができる。

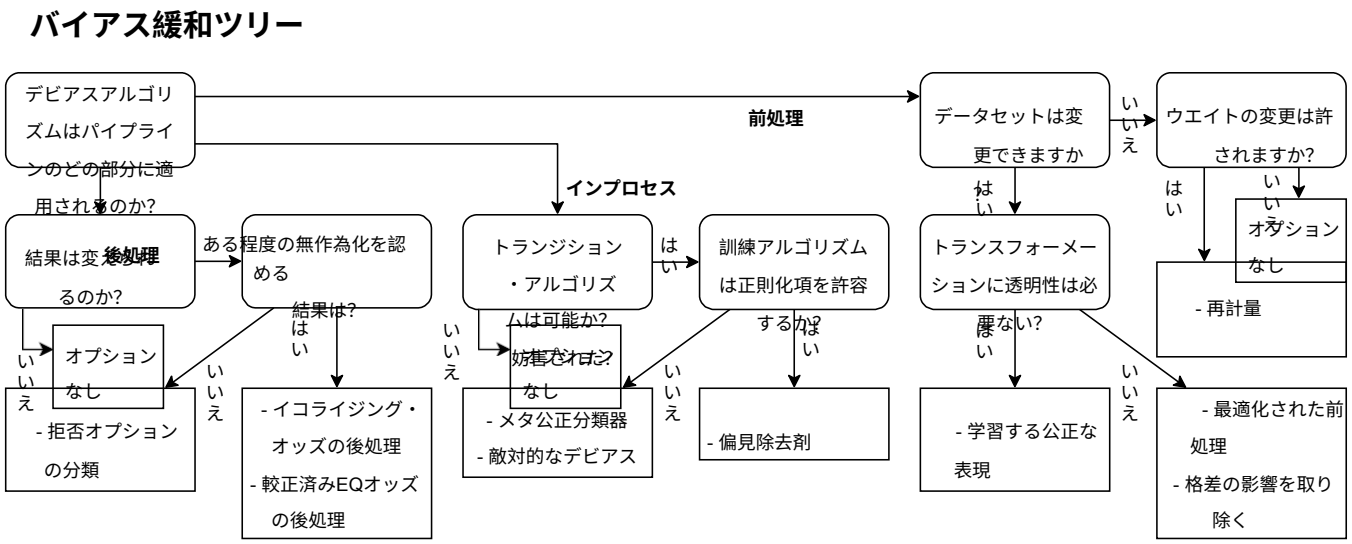


図2：特定の状況に最も適したデビアスアルゴリズムを選択するための決定木。

というのも、これは特徴量の値を変更するのではなく、各ラベル／予測の重みのみを変更することで公平性を得るものだからである。データセットを変更することが許される場合、最適化前処理[13]、Disparate Impact Remover [20]、Learning Fair Representation (LFR)[50]が適切な選択肢となり得るが、LFRはデータを潜在空間にエンコードするため、トランスペアレンシーを提供しない。最適化前処理（Optimized Preprocessing）は、データセットのラベルと特徴を修正する確率的変換を学習するが、個々の歪み、データの忠実度、グループの公平性に関する目的と制約がある[13, 35]。Disparate Impact Removerは、グループ内の順位付けを維持しながら、グループの公平性を高めるために特徴量の値を変更する[20, 35]。LFRは、データを潜在空間に符号化することによってデータセットを前処理し、保護された属性の値を難読化する[35, 50]。

バイアスの軽減は、学習アルゴリズムに干渉することによって、処理中にも適用できる。処理中に適用されるデバイアシング・アルゴリズムは、主にバイアスにペナルティを与えることで機能し、通常、分類器のバイアスを制限または正則化する公平性制約または目的を使用する[35]。例えば、Prejudice Remover [29]は、学習目的に識別を意識した正則化項を追加する。しかし、これは正則化項をサポートする学習アルゴリズムに限定される[26, 35]。この制限のため、ユーザーは、Adversarial Debiasing [26, 35]のような、より一般的な学習アルゴリズムを可能にする他のインプロセッシングアルゴリズムを考慮することができる。Adversarial Debiasing [51]は、予測から保護された属性を決定する敵の能力を低下させながら、その予測精度を最大化するように分類器を学習します。このアプローチにより、予測は敵が利用できるグループ識別情報を持たないようになり、したがって公正な分類器を導く[26]。他の処理中デビエージング・アルゴリズムは、Meta Fair Classification [14]、Gerryfair Classification [30、31]、Exponentiated Classification [26]などがある。勾配削減[2]、グリッドサーチ削減[2, 3]。

後処理によるバイアスの緩和は、予測結果を変更できる場合にのみ実行できる。拒否オプション分類[28]は恵まれないグループに有利な結果をもたらし、決定を中心とした信頼帯域では恵まれたグループには反対の結果をもたらす。

境界の不確実性が最も高い。等化オッズ後処理[23, 39]と較正等化オッズ後処理[39]は，等化オッズ目標を最適化するために，出力ラベルを変更する確率を見つけようとする．前者は線形プログラムを解くことでこれを達成し，後者は較正された分類器スコアに対して最適化を行う．2つの等化オッズ・アルゴリズムには，ランダム化要素があり，その特性のためにどちらも好ましい．そうでない場合は，リジェクト・オプション・アルゴリズムが決定論的であるため，代替オプションとなる．

3　モデルに基づくバイアスの軽減

このセクションでは、機械学習におけるモデルベースのバイアス緩和を実現するためのアプローチを紹介する。

3.1　ツールキットの選択

MDSEは、ターゲット・ソフトウェアのコードを生成するために、コード生成に大きく依存している。生成されたコードは、それぞれのドメインで特定の機能を実装するために、既存のツールやソフトウェア・ライブラリを使用することが多い。機械学習も例外ではなく、機械学習における倫理的バイアスを低減するために設計された既存のツールやライブラリを活用することができる。

Leeら[34]は、アルゴリズム倫理、特に公正な機械学習のためのオープンソースツールキットの状況を分析した。ツールキットは、1）オープンソースであること、2）実務家によって使用される可能性が高いこと、3）公正に関連する方法論を実装していること、という基準で、データ科学者のフォーカス・グループ・ディスカッションを通じて、公正な機械学習のための6つのベスト・ツールキットを選択し、分析に使用した。そのツールキットとは、Aequitas[41]、 Google What-If[48]、 Scikit-fairnet/scikit-lego[42, 43]、Fairlearn[9]である、
とIBM AI Fairness 360 [7]がある。

我々は、これらの6つのツールキットを評価し、我々のモデルベースアプローチのインフラとしての適合性を評価した。選定の主な基準は、1）オープンソースであること、2）バイアスの軽減をサポートしていること、3）プログラム可能でモデルベースアプローチに統合するためのAPI（Application Programming Inter-face）を提供していること、4）さまざまなバイアスの測定とその結果をサポートしていること、であった。

軽減アルゴリズムにより、ユーザーは最適なバイアス軽減戦略を見つけることができる。

すべてのツールキットは、オープンソースプロジェクトであるため、最初の基準を満たしている。Aequitasはバイアスの軽減（2番目の基準）ができないため除外した。を削除した。また、Google What-Ifは、他のアプリケーションと統合するためのAPI[3]を持っておらず（第3の基準）、そのカスタマイズはカスタム予測機能[4]に限られているため、除外した。IBM AI Fairness 360は、他よりも多くの機能を備えている。FairlearnとScikit-fairness/legoは2位と3位である[34]。これらはすべて4番目の基準を満たしている。しかし、IBM AI Fairness 360は他よりも多くの機能を備えているため、モデルベースのソリューションのインフラとして選択しました。

3.2 コンストラクトとワークフロー

機械学習における偏り緩和をモデル化するために、ドメイン固有言語は、ユーザーがドメインにおける本質的な構成要素とワークフローを表現できるようにする必要がある[47]。7]の研究は、IBM AI Fairness 360ツールキットを使用して偏り緩和を実装する際の重要な構成要素と典型的なワークフロー／パイプラインを文書化しており、ここではそれらについて簡単に説明する。

データセットのセットアップ（T01）。バイアスの軽減は、一般的に*データセットの*セットアップから始まる。このタスクでは、ユーザーはデータセットの*ソースを*定義する。また、どの属性が*予測される属性であるかを*決定し、atトリビュートにおける*有利なクラス*、*センシティブな属性*、およびそれらの*特権クラスと非特権クラスを*含む。データセットは、特定のツールキットが提供する*組み込みデータセットから*ロードすることもできる。*訓練データセット*、*検証データセット*、*テスト*データセットもここで設定する。これらのデータセットがすべて異なるデータセットから取得されたものであっても、同じデータセットが異なる比率で*分割されたもの*であっても構わない。また、データセット、*モデル*、*予測に適用*されるバイアスメトリクスと*緩和アルゴリズム*も選択する。

予測の前に元のバイアスを測定する(T02)。このタスクでは、ユーザはデータセットの*元のバイアスを*、精度や平均差などの異なるメトリクスで測定する。この結果は、予測(T04)またはバイアス緩和(T06)後の測定結果と比較され、後にベンチマークとして使用されます。

Train and Predict (T03)。このタスクは、タスクT01で定義された訓練データセット上で特定の*分類器/分類アルゴリズムを*一般的に使用する*モデルを訓練する*。必要であれば、モデルの最適な*パラメータを*得るためにハイパーパラメータを調整するための*検証を行う*。*その後*、ユーザーはタスク T01 で定義されたテストデータセットでモデルを使用して*予測を行う。*予測が必要ない場合は、このタスクはスキップできる。例えば、元のデータセット（前処理）のバイアスのみを測定し、緩和したい場合などである。

予測後の元のバイアスを測定する(T04)。このタスクでは、タスク T03 で予測を実行した後に、元の精度とバイアスを測定します。この結果は、後でバイアス緩和後のバイアスと比較する際のベンチマークとして使用されます。予測またはタスクT03が実行されない場合、このアクションはスキップされます。

バイアスを軽減する（T05）。このタスクでは、ユーザーはデビアスアルゴリズムのタイプ（前処理、イン処理、後処理）を選択する。

[3] https://groups.google.com/g/what-if-tool/c/zw4Rk5kxPIM
[4] https://pair-code.github.io/what-if-tool/get-started/
[5] https://aif360.mybluemix.net/

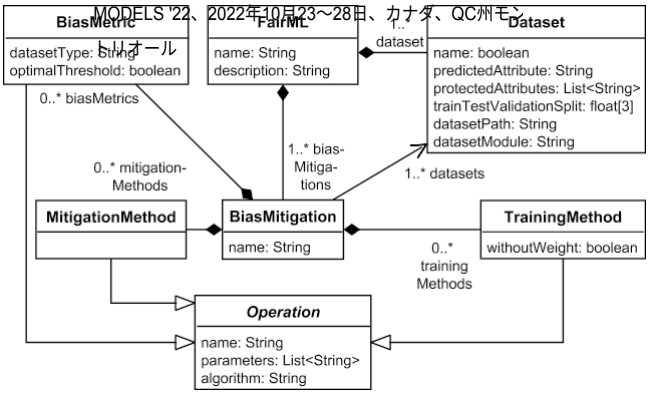


図3：FairMLのメタモデル。

- を適用する。デビアスアルゴリズムに最適なパラメータを得るために、ハイパーパラメータチューニングが必要な場合にも、*検証を*行うことができる。

デビアス後のバイアスを測定する(T06)。このタスクでは、タスクT05でデビアスアルゴリズムを適用した後の精度とバイアスを、テストデータセットを用いて測定する。

結論(T07)。このタスクでは、デバイアス後のバイアスを、タスクT02とT04で得られた元のバイアスと比較します。そして、ユーザはその結果を分析して、目的に合ったバイアスと精度を持つデータセット、分類器、デバイアスアルゴリズムの最適な組み合わせを見つけます。比較結果は、データセット、分類器、デバイアスアルゴリズムの効果を分析するために、数値、表、グラフで表示することができます。

3.3 メタモデル

上記の構成要素は、ユーザーがバイアス緩和実験を定義する際に表現できるように、FairML のメタモデルに組み込まれた。図 3 にメタモデルの簡略版を示す。FairMLクラスは機械学習におけるバイアス緩和のプロジェクトを表す。1つ以上のデータセットとバイアス緩和を含むことができる。

Dataset クラスはデータセットを定義します．このクラスでは、データセットの名前、予測属性、保護属性、train、test、validationの分割比率、CSVファイルのパスを指定します。

BiasMitigation クラスは、FairML プロジェクトで実行されるバイアス緩和を指定します。*name* 属性はバイアス緩和の識別子として使用される。このクラスは他の 3 つの重要なクラスから構成されます： TrainingMethod、 MitigationMethod、 Bias-Metricです。これらは実行され、パラメータを持ち、結果を返すので、すべてOperation抽象クラスから派生しています。*name*属性と*parameters*属性は、それぞれ操作の識別子とパラメータを定義するために使用されます。Operationクラスのalgorithm属性は、それぞれの派生クラスが実行するアルゴリズムを定義します。TrainingMethodクラスは学習と予測のための分類器を定義し、MitigationMethodクラスはバイアスを軽減するためのデビアスアルゴリズムを指定します。BiasMetricクラスは、公平性を測定するために使用するメトリクスを決定します。

3.4 FairMLの例

リスト1は、図3のメタモデルに準拠し、YAMLで表現されたFairMLモデルを示しており、IBM AI Fairness 360のデモ例のFairMLバージョンである[6]。この例では、*成人*[7]データセットを予測モデルの訓練とテストに使用している。データセットのバイアスを軽減する（前処理によるバイアス軽減）代わりに、この例では機械学習パイプラインの処理段階でバイアスを軽減するためにExponentiated Gradient Reduction技術を適用している。最後に、処理中のバイアスを軽減したモデルとしなかったモデルの精度と公平性の比較で締めくくります。オリジナルの例には、コメント行と空行を除いて93行のPythonコードがあり、リスト1の約3倍の長さがあります。

モデルには名前と説明があり、それぞれDemoとDebiasing using EGRです（3〜4行目）。モデルには1つのデータセット（7行目）と1つのバイアス緩和（15行目）があります。データセットはAdult と名付けられ、load_preproc_data_adult.csvというcsvファイルから読み込まれる。また、保護属性として性別と人種を設定し、予測属性として所得バイナリを設定し、元のデータセットを7:3:0の比率で訓練、テスト、検証データセットに分割する（7〜12行目）。

15〜32行目は、Demoモデルで実行されるバイアス緩和の定義です。バイアス緩和はExponentiated Gradient Reductionと名付けられ、7〜12行目で定義したAdultデータセットをデータセットとして使用します。

バイアスの緩和は，予測値がデバイアスをかけずにロジスティック回帰分類器のみを使用した場合（19-21行目）と，予測値がExponentiated Gradient Reductionを使用してデバイアスをかけた場合（23-25行目）の精度と公平性（平均差，平均オッズ差）のトレードオフを分析するために，3つの異なるメトリクスを使用します（27-32行目）．前者は，1つのパラメータしか使用しません；'lbfgs' がソルバーとして設定され，後者は3つのパラメータを持ちます；同じ分類器が推定器として使用され，'EqualizedOdds' が制約として使用され，保護された属性は削除されません．

リスト1: YAMLで表現されたDemo Exponentiated Gradient Reductionを使ったバイアスの軽減。

```
1  nsuri : fairml
2  fairml :
3  - name : デモ
4  - 説明 : EGRによるデビアス
5
6  #datasetset
7  - データセット :
8      - name : アダルト
9      - 予測属性 : 所得バイナリ
10     - 保護属性 : 性別 , 人種
11     - train TestValidation 分割 : 7 , 3
12     - datasetPath :
       load_preproc_data_adult.
13
14 #definettebiassmitigation
15 - バイアスの緩和 :
16     - name : 指数化勾配リダクション
17     - dataset : アダルト
18
19     - トレーニング方法 :
```

[6] https://github.com/Trusted-AI/AIF360/blob/master/examples/demo_exponentiated_gradient_reduction.ipynb.
[7] https://archive.ics.uci.edu/ml/datasets/Adult

```
    - アルゴリズム：ロジスティック回帰
    - パラメータ：solver='lbfgs'

  - 軽減方法：
    - アルゴリズム： Exponentiated GradientReduction
    - parameters：estimator= ロジスティック回帰（solver='lbfgs'），
        constraints='等化オッズ'，drop_prot_attr= False

  - バイアス・メトリック：
    - 名前：精度
  - バイアス・メトリック：
    - name：mean_difference
  - バイアス・メトリック：
    - name：平均オッズ差
```

3.5 ウィザード

FairMLはバイアスの軽減を指定する手段を提供しますが、ユーザーが適切に使用するためには、サポートされているすべてのデバイアスアルゴリズムとバイアスメトリクスを理解する必要があります。自然言語で質問するウィザードを使用してFairMLモデルの構築を支援できることは重要です。ウィザードは図1と図2の決定木に従い、それぞれセクション2.2と2.3で説明した。

ウィザードを完了すると、FairMLは自動的にFairMLモデルn Flexmiを生成します（下記参照）。このファイルは、ユーザが後から他のデビアスアルゴリズムやバイアスメトリックを追加したり、既存のものを修正したりしたい場合に変更可能です。ウィザードは、生成されたモデルのPythonファイルやJupyterノートブックファイルも自動的に生成します。

リスト2: FairMLのウィザードで、ユーザーが最適なデビアスアルゴリズムとバイアスメトリックを選択するための質問の一部。

```
...
---- 軽減アルゴリズム
# 1.前処理
前処理でバイアス軽減を適用する（デフォルト：true）：
データセットの重みは変更可能（デフォルト：true）：
バイアスの軽減は潜在空間を許容する（デフォルト：false）：
...
...
---- バイアスメトリック
グループの公平性を測定する（デフォルト：true）：
個々の公平性を測定する（デフォルト：false）：
個人とグループの両方に単一のメトリクスを使用する（デフォルト：false）：
公平性を測る（デフォルト：false）：
...
```

3.6 世代

図4は、FairMLモデルを消費してPythonとJupyter Notebookファイルを生成する変換を示しています。ユーザーは

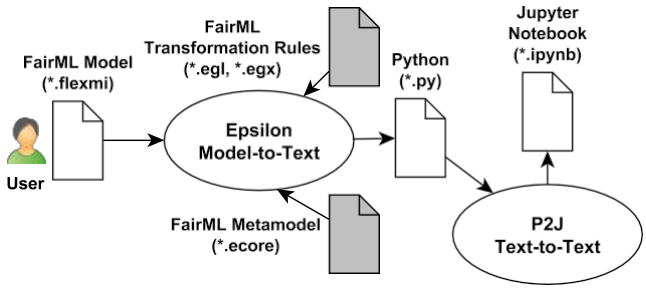


図4: fairMLにおけるモデルからテキストへの変換。

Flexmi[8][32]のバイアス軽減モデル。FlexmiはEMFモデル[45]のための反射的テキスト構文であり、ファジーマッチングを使用してYAML/XMLドキュメントのタグと属性をターゲットメタモデルのEcoreクラス/フィーチャー名[45]にマッピングする。デフォルトのXMLベースのフォーマットをサポートするだけでなく、FlexmiはYAML風味のフォーマットで表現されたモデルを読み込むこともできます。Epsilon Gen- eration Language (EGL) [40]で書かれた変換は、FairMLメタモデルに準拠したモデルを消費し、Pythonファイルを生成します。同じ変換は、どのモデルからテキストへの言語でも実装可能であることに留意されたい。生成されたPythonファイルは、P[2]J9エンジンを使用してJupyterノートブックファイルに変換されます。

3.7 生成ファイル

リスト1で定義されたモデルに基づいて、FairMLはバイアス軽減を実行するコード、測定されたバイアスの説明、選択された分類器とデバイアスアルゴリズムへの参照を含むJupyterノートブックファイルを生成し、ユーザーが適用されたバイアス軽減と測定されたメトリクスを理解できるようにします。また、異なるデータセット、分類器、デバイアスアルゴリズムによって測定されたバイアスメトリクスを色分けして比較した表も含まれており、ユーザーがコンテキストに合った最適な組み合わせを選択するのに役立ちます。

例として、図5は、リスト1によって生成されたバイアス測定値のサマリー表（生成されたノートブックのスクリーンショットの一部）を示しています。予測の組み合わせは3つあります。最初の組み合わせでは、平均差のみを測定します、
元のデータセットにおける統計的パリティ差（Statistical Parity Difference）。ロジスティック回帰を使用して予測を実行した後、2番目の組み合わせは、バイアスを緩和することなく、いくつかのメトリクスを測定する。表を見ると、平均差は-0.206とわずかに悪化し、0から遠ざかり、特権階級が有利になっている（負の符号）。偏りを緩和するアルゴリズムとして指数化勾配削減（トリオール）を適用すると、予測精度は0.79にわずかに低下するが、公平性は大幅に改善される。

データセット、分類器、デバイアスアルゴリズム、バイアスメトリック値の多くの異なる組み合わせがユーザに提示される状況がある。これらの選択肢の中から、どの組み合わせが自分のバイアス軽減戦略に最適かを選択する必要があります。そこで、各メトリクスの最適な測定値を太字にし、色分けすることで、ユーザーを支援します。その

[8] https://www.eclipse.org/epsilon/doc/flexmi
[9] https://pypi.org/project/p2j/

は、白から黄色、緑へと等級付けされ、それぞれがそれぞれの測定基準の最悪、中程度、最良の測定値に対応する。図5では、ユーザーは2番目の組み合わせが最も精度が高いことに容易に気づくことができる。また、3番目のオプションが最もバイアスの少ない組み合わせであるが、それでも精度が高いこともわかる。

4 エバリュエーション

我々はFairMLの表現力、正しさ、簡潔さ、実行時間を評価した。表現力の評価では、「*FairMLは実際の偏り緩和のユースケースをサポートできるか*」という問いに答えることを目的としました。そのために、FairMLの表現力を評価するためのベースラインとして、IBM Fairness AI 360が提供する例を使用しました。ツールキットには、実際の文脈におけるツールキットの使用例を示すいくつかの例（表1）が付属している。我々はツールキットのバージョン3.0に含まれる12の例を使用した。各例で使用されているバイアスメトリクス、デビアスアルゴリズム、分類器、データセットなどのすべての要素を特定し、その複雑さを測定した。例題に含まれる要素が多ければ多いほど、FairMLはより複雑なケースを表現することができる。

元の例で測定されたバイアス指標値（測定値）と生成されたコードで測定された値を比較し、FairMLが生成したのコードを使用する。どちらも、±0.1の許容誤差の範囲内で、同じか似たような値を出すはずである。

簡潔さの評価では、書き言葉の比率を比較した。vs.FairMLの生成コード vs.例のコード。この指標は多くの要因（セクション4.2を参照）により、必ずしも生産性を保証するものではありませんが、ユーザーが書いた行数は大幅に減少しています。測定では、空白行とコメント行はカウントされませんでした。

また、FairML の実行時間も測定した。まず、*生成時間*（FairMLがバイアス検出と緩和のコードを生成するのにかかる時間）を計測しました。次に、生成されたターゲット・コードとサンプル・コードの処理時間を比較しました（*生成された実行*時間と*元の実行時間の比較*）。評価は、Windows 10 64ビットオペレーティングシステム、第11世代Intel(R) Core(TM) i9-11900H @ 2.50GHz 8コアプロセッサ、32.0GB RAM DDR4、Open- JDK Runtime Environment 18.9 (build 11.0.14.1+1)、Python 3.9.7を搭載したマシンで実施しました。

4.1 結果と考察

このセクションでは、FairMLの評価結果[10]を紹介し、議論する。

*4.1.1 表現力と正しさ。*FairMLは、表1の12の例すべてを再現することができた。合計で、シナリオは6つのユニークなデータセット（Adult, German, Compas, MEPSDataset19, MEPSDataset20, MEPSDataset21）、6つの分類器（Logistic Regression, Linear Regression, Linear SVR, Decision Tree, Kernel Ridge, Ran- dom Forest）、11のバイアス軽減アルゴリズム（Adversarial Debiasing、Calibrated Equalising Odds，Disparate Impact Remover，Exponent- iated Gradient Reduction，Gerryfair，Learning Fair Representation，Reweighing，Meta Fair Classification，Optimized Preprocessing，Re-ject Option Classification，Prejudice Remover），および13のバイアス緩和アルゴリズムがある．

[10] 評価データは https://github.com/York-and-Maastricht- Data-Science-Group/fairml/blob/main/data/evaluation.xlsx にある。

ワースト＝白、ベスト＝緑、ミッド＝黄

	軽減	データセット	Classifier	accuracy	mean_difference
	average_odds_difference				
1	オリジナルアダルト(7.0:3.0:0.0)		なし	-0.198048	なし
2	オリジナル・アダルト(7.0:3.0:0.0)		LogisticRegression solver='lbfgs'	**0.804204**	-0.205572	-0.272736
3	ExponentiatedGradientReduction estimator=LogisticRegression(solver='lbfgs')、制約='EqualizedOdds'、drop_prot_attr=False	大人(7.0:3.0:0.0)	ExponentiatedGradientReduction estimator=LogisticRegression(solver='lbfgs')、制約='EqualizedOdds', drop_prot_attr=False	0.787552	**-0.052739**	**0.010994**

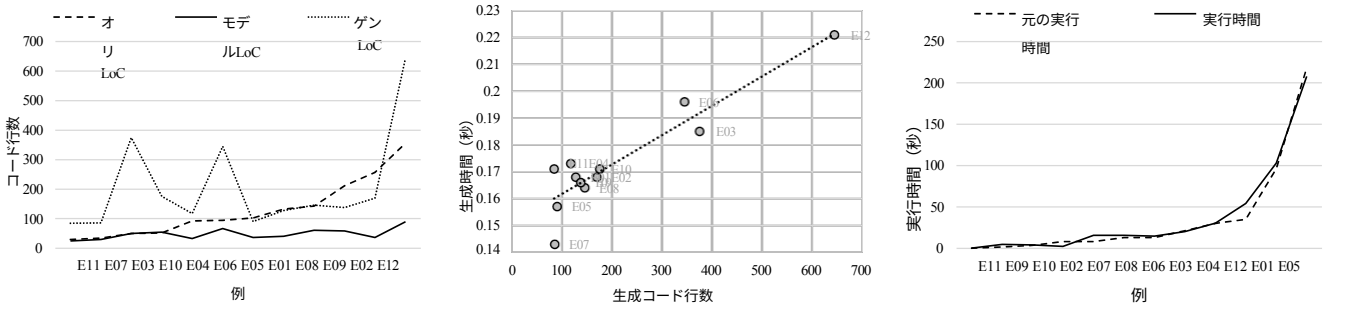
図5: リスト1で定義されたモデルに基づいて生成されたJupyterノートブックファイルに表示される要約表。

表1: IBM AI Fairness 360の例と、各例における固有のデータセット、分類器、デバイアスアルゴリズム、バイアス測定基準、および測定値の数。

コード	例/ファイル名 (*.ipynb)	データセット	分類子	デビアス・アルゴリズム	バイアス指標	確実性
	E01デモ敵対的デビアス	アダルト	否定的・否定	的デビアス	精度、バランス精度格差影響、平均オッズ、統計的平等、機会均等、ザイル指数	28
E02	デモ較正EQ奇数後処理	アダルト、ドイツ語、コンパ	ロジスティック回帰	較正eqオッズ後処理	平均差、偽陽性率、偽陰性率、バランス精度、機会均等	13
	E03Demo 格差の影響除去装置	成人	ロジスティック	回帰比較影響リムーバー	格差の影響	11
E	04デモ指数化された勾配の減少	成人	ロジスティック回帰	指数化グラデーションリダクション	精度、平均差、平均オッズ	8
	E05デモLFR 5	アダルト		ロジスティック	回帰代表権の差、格差の影響	平均精度
	E06Demo メタ分類器	アダルト	N	/AMeta公正分類器	精度、バランス精度格差の影響、偽り発見率	13
	E07デモ・オプティム・プリプロック・アダルト	アダルト	該当なし	最適化された前処理	平均差	2
	E08デモ・リジェクト・オプションの分類	アダルト、ドイツ語、コンパ		ロジスティック回帰		オプションの分類を拒否する

					バランス、精度、格差26インパクト、平均オッズ、統計的平等、機会均等、ザイル指数	
E09	デモ再計量プリプロック	ア ダルト、ドイツ語、コンパ	ロジスティック回帰	再重量法	精度のバランス、ばらつき インパクト, 平均オッズ,.統計的平等、機会均等、ザイル指数	16
E	10デモ・ショート・ジェリーフェア・テスト		成人線形回帰、リ ニ ア	ッジ		
E11	チュートリアル・クレジット・スコアリング	ドイツ語	SVR, 決定木, 該当なしカーネルリ	再計量	平均差	2
E12	チュートリアル・メディカル支出	MEPS データセット19	ランダムフォレスト、ロジスティック回帰	再計量、偏見リムーバー	精度のバランス、格差インパクト、平均オッズ	90

MODELS '22、2022年10月23〜28日、カナダ、QC州モントリオール



(a) オリジナルとモデルのコード行数 **(b) 生成時間とコード行数の比較** **(c) オリジナルコードと生成コードの実行時間**

図6：簡潔性、生成時間、実行時間に関するFairMLの評価。

(精度、バランス精度、平均差、統計的パリティ差、格差影響、機会均等差、Theil指数、ガンマ格差、平均オッズ差、平均絶対誤差、偽陽性率、偽陰性率、偽発見率）。さらに、FairMLは外部CSVファイルからのデータ読み込みもサポートしています。これらにより、データセット、分類器、デバイアスアルゴリズム、バイアスメトリクスの様々な組み合わせでバイアス軽減戦略を表現することができます。

正しさについては、機械学習過程におけるランダム性のため、生成されたコードによって生成されたすべての値が、例題におけるそれぞれの値と正確に等しいとは限らない。しかし
の許容誤差±0.1の範囲内にある。
それぞれのものである。

*4.1.2 簡潔さ。*図6aはオリジナルの例、FairMLモデル、生成されたコード間のコード行数（LoC）の比較を示しています。モデルのLoC（実線）はオリジナルのLoC（破線）より少ない傾向があり、ユーザーが例と同じ結果を得るために書くコードが少ないことを示しています（正しさについてはセクション4.1.1を参照）。また、元の LoC の数が増えるほど効率も高くなる。例 E11（左端）から例 E12（右端）では、元の LoC が 30 から 356 に増えており、ユーザは元の LoC の 83%から 25%のコードを書くだけでよくなっている。

FairMLは、(1)使用されている分類器、バイアス緩和アルゴリズム、バイアス測定基準を説明するためのコード、(2)測定結果を要約表やグラフの形で表示する際の異なるアプローチなど、元の例には含まれていない特徴を含んでいるため、元の例よりも多くのLoC（点線）を生成することが期待される。

*4.1.3 生成時間と実行時間。*図6bはFairMLが生成コードを生成するのにかかる時間（生成時間）を示しています。生成時間に関しては、FairMLは0.2秒未満で元の例のすべての生成コードを生成することができ、例E12に基づき、2,927行/秒（644行/0.22秒）を達成することができます。

図6cは、元の例に対する生成コードの実行時間のパフォーマンスを示している（単位は秒）。一般に、生成されたコード（実線）の実行時間は、元の例（点線）よりもわずかに長い。私たちはこのことを予期しています。

セクション4.1.2で述べられているように、生成されたコード・バージョンはより多くの機能を持ち、より多くのコード行を持つからである。

4.2 妥当性への脅威

我々は、IBM AI Fairness 360 のドキュメントと事例に基づいて FairML を開発し、テストしているが、現場の経験豊富なユーザーからフィードバックを得るためのユーザー評価を行っていない。これには、バイアスの識別と緩和の経験を持つ多数のデータ科学者の参加が必要であるが、これは非常に希少な資源である[34]。とはいえ、我々が再現したドキュメントと例は専門家によって書かれたものであり、彼らが操作するデータは様々なドメインからの実世界のデータである。

5 教訓

この仕事から学んだいくつかの教訓がある：

- バイアスの識別と軽減は、さまざまなタイプのメトリクスを対象とする、よく理解された多くのアルゴリズムによってサポートされている。

  とバイアスがある。そのため、FairMLのような境界の明確なDSLのための強固な基礎を提供することができる。
- バイアスの特定と緩和のプロセスは間違いやすい(アルゴリズムを間違った順序で適用するなど）。それらを表現するDSLでモデルからテキストへの変換を使用すると、生成されたコードに特定の構造が課されます。
- モデルからJupyterノートブックを生成するモデルからテキストへの変換またはモデルからJSONへの変換は技術的に難しい。それでも、Pythonのテキストプログラムを Jupyterノートブックに変換できるP2Jのようなサードパーティエンジンを活用することで、それを回避することができる。

6 関連作品

機械学習におけるバイアスを測定するために、いくつかのツールキットが開発されている。FairMLは[1]によって開発されたツールボックスで、4つの入力ランク付けアルゴリズムとモデル圧縮を活用することで、予測モデルに対する異なる入力の相対的な影響を計算し、予測モデルの公平性を監査する。FairTest [46]は、データセットのバイアスをチェックするために、敏感な属性と予測されたラベルの間の関連性を計算する。また、アルゴリズムが異常な高率のエラーを生成する可能性のある入力空間の領域を特定する手法を提供する。Themis [22]はバ

は、予測システムの決定における差別を測定するためのテストを自動生成することができる。Fairness Measures [49]は、格差影響、平均オッズ比、平均差など、さまざまなバイアス・メトリクスの測定をサポートします。Aequitas [41]は、さまざまな測定基準で公平性を測定する監査ツールキットです。また、特定の状況に対して最も適切な測定基準を選択する際のガイドとなる決定木も提供しています。

他のいくつかのツールキットも、公平性を測定するためだけでなく、バイアスを軽減することができる。ThemisML [6]、Fairness Comparison [21]、Aequitas [41]、Google What-If [48]、Scikit-fairnet/scikit-legoなどがそれである。
[42、43]、Fairlearn [9]、IBM AI Fairness 360 [7]。

Rapidminer [24]、Knime [8]、Orange [17]は、ローコード、モデルベースのアプローチを使用し、コンポーネントベースの手順を組み立てることによって、ユーザーが機械学習パイプラインを視覚的にプログラムできるようにするプラットフォームである。これらのプラットフォームは、データの探索、変換、可視化、機械学習とデータマイニングのためのさまざまなアルゴリズムをサポートしている。さらに、これらはすべて拡張可能であり、ユーザーは新しいモジュールやスクリプトを追加することができる。しかし、私たちが知る限り、バイアスを測定し緩和するための組み込みモジュールはない。

FairMLに最も近い既存の研究は、倫理的な機械学習のために設計されたドメイン固有言語であるArbiter [52]である。これは、機械学習モデルの学習方法を定義するためのSQLライクな宣言型言語で、透明性、公平性、説明責任、再現性という4つの要素で倫理的実践を記述する。透明性
しかし、その実装は特定のメトリックと分類器[11]に限定されている。

7 結論と今後の課題

本論文では、偏り緩和を自動化するためのモデルベースアプローチを実装したツールキットであるFairMLを紹介した。FairMLを用いることで、ユーザは簡潔かつ宣言的なマンナーでバイアス緩和コードを生成し、そこから実行可能なPythonコードを生成することができる。さらに、正しさの観点から、生成されたコードは元の例で測定されたものと同様のバイアス測定値を生成する。

今後の課題として、FairMLによって生成されるコードの行数は、繰り返し行を関数にマージしたり、最初に静的解析を行うことで不要なコードを削除したりすることで、さらに削減することができます。その効果として、不要なコードを削除することで、生成されるコードの実行時間を最適化することができます。

8 謝辞

この研究は、York-Maastricht パートナシップの Responsible Data Science by Designプログラム（https://www.york.ac.uk/maastricht)。マーストリヒトチームの貴重な貢献に感謝する。

参考文献

[1] Julius A Adebayo et al.*FairML: ToolBox for diagnosing bias in predictive modeling*.博士論文。マサチューセッツ工科大学。

[2] Alekh Agarwal 、 Alina Beygelzimer 、 Miroslav Dudik 、 John Langford 、 Hanna Wallach 。 2018.A Reductions Approach to Fair Classification.*Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.).PMLR, 60-69. https://proceedings.mlr.press/v80/agarwal18a.html

[3] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu.2019.公正回帰．定量的定義と削減ベースのアルゴリズム。*Proceedings of*

[11] https://github.com/julian-zucker/arbiter

第36回機械学習国際会議（機械学習研究論文集第97巻）, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). https://proceedings.mlr.press/v97/agarwal19d.html。

[4] AI Fairness 360 (AIF360) 著者。2022.AIフェアネス3 6 0 の ドキュメント：//aif360.readthedocs.io/ja/stable/ Accessed: 2022-01-30。

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner.2016.機械の偏見：将来の犯罪者を予測するために全米で使われているソフトウェアがある。そしてそれは黒人に偏っている。ProPublica (2016).https://www.propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: 2022-01-18.

[6] ニールス・バンティラン2018.Themis-ml： A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation.*Journal of Technology in Human Services* 36, 1 (2018), 15-30. https://doi.org/10.1080/15228835.2017. 1416512 arXiv:https://doi.org/10.1080/15228835.2017.1416512

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varsh- ney, and Yunfeng Zhang.2018. AI Fairness 360：An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943 [cs.AI] https://arxiv.org/abs/1810.01943

[8] ミヒャエル・R・ベルトルト、ニコラ・セブロン、ファビアン・ディル、トーマス・R・ガブリエル、トビアス・ケト
Ter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel.2008.KNIME: The Konstanz Information Miner.*データ解析、機械学習と応用*, Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker (Eds.).Springer Berlin Heidelberg, Berlin, Heidelberg, 319- 326.

[9] サラ・バード、ミロ・ドゥディク、リチャード・エドガー、ブランドン・ホーン、ロマン・ルッツ、ヴァネッサ・ミラン、メフルノシュ・サメキ、ハンナ・ウォラック、キャスリーン・ウォーカー2020.*Fairlearn： AIにおける公平性の評価と改善のためのツールキット*。技術報告書 MSR-TR-2020 32.Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn- a-toolkit-for-assessing-and-improving fairness-in-ai/

[10] M.Brambilla, J. Cabot, and M. Wimmer.2017.*Model-Driven Software Engineering in Practice*.Morgan & Claypool Publishers. https://books.google.co.uk/books?id=dHUuswEACAAJ

[11] ジョイ・ブオラムウィニ、ティムニット・ゲブル。2018.Gender Shades： Intersectional Accu- racy Disparities in Commercial Gender Classification.*Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.).PMLR, 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html.

[12] C. バ ー ン 2017.*Development Workflows for Data Scientists*.O'Reilly Media. https://books.google.co.uk/books?id=84HgwQEACAAJ

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Rama-murthy, and Kush R Varshney.2017.最適化された識別防 止 の ための前処理。*Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol.30.Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/ 9a49a25d845a483fae4be7e341368e36-Paper.pdf

[14] L.Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi.2019.Clas- sification with Fairness Constraints： A Meta-Algorithm with Provable Guarantees.*Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.Association for Computing Machinery, New York、NY, USA, 319-328. https://doi.org/10.1145/3287560.3287586

[15] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell.2019.Fairness Under Unawareness：保護階 級 が 未観測の場合の格差の評価 ． *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.Association for Computing Machin- ery, New York, NY, USA, 339-348. https://doi.org/10.1145/3287560.3287594.

[16] ペドロ・コンセイサォン、ペドロ・フェレイラ2000.1The Young Person's Guide to the Theil Index：直感的解釈の提案と分析的応用の探求 。

[17] Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Š t a j d o h a r , Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan.2013.Orange： Pythonのデータマイニングツールボックス。*Journal of Machine Learning Research* 14 (2013), 2349-2353. http://jmlr.org/papers/v14/demsar13a.html.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel.2012.意識による公平性。*第3回理論計算機科学の革新会議*（マサチューセッツ州ケンブリッジ） *（ITCS '12）予稿集*。Association for Computing Machinery, New York, NY, USA, 214-226：
//doi.org/10.1145/2090236.2090255

[19] クラーク・エバンス、O・ベン＝キキ、I・デット・ネット。2017.YAML Ain't Markup Language (YAML™) Version 1.2. https://yaml.org/spec/1.2.2 Accessed: 2022-01-19.

[20] マイケル・フェルドマン、ソレル・A・フリードラー、ジョン・モーラー、カルロス・シャイデッガー
スレーシュ・ヴェンカタスブラマニアン2015.格差インパクトの認定と除去 。 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*.協会

Computing Machinery, New York, NY, USA, 259-268. https://doi.org/10.1145/2783258.2783311

[21] ソレル・A・フリードラー、カルロス・シャイデッガー、スレシュ・ヴェンカタスブラマニアン、ソナム・チョウダリー、エヴァン・P・ハミルトン、デレク・ロス。2019.A Comparative Study of Fairness-Enhancing Interventions in Machine Learning.*Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.Association for Computing Machinery, New York, NY, USA, 329-338. https://doi.org/10.1145/3287560.3287589.

[22] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou.2017.Fairness Testing：ソフトウェアの差別テスト。*Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering* (Paderborn, Germany) *(ESEC/FSE 2017)*.Association for Computing Machinery, New York, NY, USA, 498-510：//doi.org/10.1145/3106237.3106277

[23] モリッツ・ハルト、エリック・プライス、ナティ・スレブロ。2016.教師あり学習における機会均等。*Neural Information Process- ing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol.29.Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[24] M.Hofmann and R. Klinkenberg.2016.*RapidMiner： Data Mining Use Cases and Business Analytics Applications*.CRC Press. https://books.google.co.id/books?id=Y_wYCwAAQBAJ

[25] IBM AIリサーチ。 LALEのAPIドキュメントへようこそ！ https://lale.readthedocs.io/en/latest/modules/lale.lib.aif360.util.html#lale.lib.aif360.util.theil_index Accessed: 2022-01-30.

[26] IBM Research 信頼されるAI。2022. https://aif360.mybluemix.net/resources#guidance Accessed: 2022-01-30.

[27] ファイサル・カミラン、トゥーン・カルダース。2011.差別のない分類のためのデータ前処理技術。*Knowl.Inf.Syst.*33, 1 (2011), 1-33. https：//doi.org/10.1007/s10115-011-0463-8

[28] Faisal Kamiran, Asim Karim, and Xiangliang Zhang.2012.判別に配慮した分類のための決定理論.In *2012 IEEE 12th International Conference on Data Mining*.924–929. https://doi.org/10.1109/ICDM.2012.45

[29] 神島敏宏、赤穂翔太郎、麻生英樹、佐久間準。2012.偏見除去正則化器を用いた公平性を考慮した分類器.*データベースにおける機械学習と知識発見*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.).Springer Berlin Heidelberg, Berlin, Heidelberg, 35-50.

[30] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu.2018.Preventing Fairness Gerrymandering： Auditing and Learning for Subgroup Fairness.*Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). pmlr、2564-2572。https://proceedings.mlr.press/v80/kearns18a.html

[31] マイケル・カーンズ、セス・ニール、アーロン・ロス、呉志偉。2019.An Em- pirical Study of Rich Subgroup Fairness for Machine Learning.*Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*.Association for Computing Machinery, New York, NY, USA, 100-109. https://doi.org/10.1145/3287560.3287592.

[32] Dimitrios S. Kolovos, Nicholas Matragkas, and Antonio García-Domínguez.2016.Towards Flexible Parsing of Structured Textual Model Representations.In *Pro- ceedings of the 2nd Workshop on Flexible Model Driven Engineering co-located with ACM/IEEE 19th International Conference on Model Driven Engineering Languages & Systems (MoDELS 2016), Saint-Malo, France, October 2, 2016 (CEUR Workshop Pro- ceedings, Vol. 1694)*, Davide Di Ruscio, Juan de Lara, and Alfonso Pierantonio (Eds.). CEUR-WS.org, 22-31. http://ceur-ws.org/Vol-1694/FlexMDE2016_paper_3.pdf

[33] プリーティ・ラーホティ、クリシュナ・P．Gummadi, and Gerhard Weikum.2019. iFair：アルゴリズムの意思決定のための個別に公平なデータ表現の学習。In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*.1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[34] ミシェル・セン・ア・リー、ジャット・シン2021.*オープンソース・フェアネス・ツールキットの現状とギャップ*。Association for Computing Machinery, New York, NY、USA. https://doi.org/10.1145/3411764.3445261

[35] T.T. Mahoney, K.R. Varshney, M. Hind, and an O'Reilly Media Company Safari.2020.*AI Fairness： AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*.https://books.google.co.id/books?id= uSbfzQEACAAJ

[36] ニナレ・メフラビ、フレッド・モースタッター、ニルプスタ・サクセナ、クリスティーナ・ラーマン、アラム・ガルスティヤン2021.機械学習におけるバイアスと公平性に関する調査。*ACM Comput. Surv.*54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[37] A.C.ミュラー、S.グイド。2016.*Introduction to Machine Learning with Python： A Guide for Data Scientists*.O'Reilly Media. https://books.google.co.uk/books?id= vbQlDQAAQBAJ

[38] オックスフォード・リファレンス2022. バイアス https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095504939 Accessed: 2022-01-16.

[39] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger.2017. Fairness and Calibration.*神経情報処理システムの進歩*, I. Guyon, U. V. Luxburg, S. Bengio、H.Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol.30.Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

[40] Louis M. Rose, Richard F. Paige, Dimitrios S. Kolovos, and Fiona A. C. Polack.2008.イプシロン生成言語。*モデル駆動アーキテクチャ - 基礎と応用*, Ina Schieferdecker and Alan Hartman (Eds.).Springer Berlin Heidelberg, Berlin, Heidelberg, 1-16.

[41] ペドロ・サレイロ、ベネディクト・クエスター、ローレン・ヒンクソン、ジェシー・ロンドン、アビー・スティーブンス、アリ・アニスフェルド、キット・T. Rodolfa, and Rayid Ghani.2019.Aequitas： A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs.LG] https://arxiv.org/abs/1811.05577

[42] scikit-fairness.2022. scikit-fairness. https://scikit-fairness.netlify.app/ Accessed: 2022-01-30.

[43] scikit-lego.2022. scikit-lego. https://scikit-lego.readthedocs.io/en/latest/index.html Accessed: 2022-01-30.

[44] Till Speicher、Hoda Heidari、Nina Grgic-Hlaca、Krishna P. Gummadi、Adish Singla、Adrian Weller、Muhammad Bilal Zafar。Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar.2018.A Unified Approach to Quanti-fying Algorithmic Unfairness： Inequality Indices による個人とグループの不公平さの測定。*Proceedings of the 24th ACM SIGKDD International Confer- ence on Knowledge Discovery and Data Mining* (London, United Kingdom) *(KDD '18)*.Association for Computing Machinery, New York, NY, USA, 2239-2248. https://doi.org/10.1145/3219819.3220046.

[45] D.Steinberg, F. Budinsky, and E. Merks.2009.*EMF: Eclipse Modeling Framework*. https://books.google.co.id/books?id=oAYcAAAACAAJ。

[46] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin.2017.FairTest：データ駆動型アプリケーションにおける不当な関連付けの発見。In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*.401-416. https://doi.org/10.1109/ EuroSP.2017.29

[47] M.Völter, T. Stahl, J. Bettin, A. Haase, S. Helsen, K. Czarnecki, and B. von Stockfleth.2013.*Model-Driven Software Development： Technology, Engineering, Management*.Wiley. https://books.google.co.uk/books?id=9ww_D9fAKncC

[48] ジェームズ・ウェクスラー、マヒマ・プシュカルナ、トルガ・ボルクバシ、マーティン・ワッテンバーグ、フェルナンダ・ヴィエガス、ジンボ・ウィルソン 2020.What-Ifツール：機械学習モデルのインタラクティブなプロービング。*IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56-65. https://doi.org/10.1109/TVCG.2019.2934619.

[49] 「Meike Zehlike, Carlos Castillo, Francesco Bonchi, Ricardo Baeza-Yates, Sara Hajian, and Mohamed Megahed".2017.fairness measures： A Platform for Data Collection and Benchmarking in discrimination-aware ML. https:// fairnessmeasures.github.io. https://fairnessmeasures.github.io.

[50] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork.2013.公正な表現の学習.*第30回機械学習国際会議予稿集（Proceedings of the 30th International Conference on Machine Learning）（Proceedings of Machine Learning Research, Vol.28）*, Sanjoy Dasgupta and David McAllester (Eds.).PMLR, Atlanta, Georgia, USA, 325-333. https://proceedings.mlr.press/v28/zemel13.html

[51] ブライアン・フー・チャン、ブレイク・ルモワン、マーガレット・ミッチェル。2018.敵対的学習による望まれていないバイアスの軽減。*Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*.As- sociation for Computing Machinery, New York, NY, USA, 335-340. https： //doi.org/10.1145/3278721.3278779

[52] ジュリアン・ザッカー、ミレーカ・ドゥルーウェン2020.*アービター： A Domain-Specific Language for Ethical Machine Learning*.Association for Computing Machinery, New York, NY, USA, 421-425. https://doi.org/10.1145/3375627.3375858