様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２１ 日現在

機関番号：１２１０２
研究種目：研究活動スタート支援
研究期間：2021～2022
課題番号：２１Ｋ２０２３１

研究課題名（和文）属性間統計差別を抑制したAI学生リスク予測モデルの検証及び開発

研究課題名（英文）Developing and Validating an Equitable AI-Based Model for Predicting Student Risk Across Various Subgroups

研究代表者
柳浦　猛（Yanagiura, Takeshi）

筑波大学・教育推進部・准教授

研究者番号：９０９０２２８９

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：多くの大学が、AIを搭載した早期警告システム（EWS）を使って、学生が最初の学期を始めると同時にサポートを提供している。しかし、EWS技術の基礎となるアルゴリズムが、学生の大学生活の早い段階で公正でな判断を下せるかどうか懸念がある。本研究は日本のある大学のデータを用いて、大学1学期のGPAを予測する機械学習アルゴリズムを開発しその公正さを検証した。本研究の研究成果は大きく二つある。一つは、早期のEWSの導入は属性差別的な判断を行う可能性が高いということである。もう一つは統計的に一般的に用いられている「公正なアルゴリズム」の達成が必ずしも公正な教育結果につながるとは限らないという点である。

研究成果の学術的意義や社会的意義
本研究の学術的意義は以下の3つの点に要約される。一つは、高等教育における公正なアルゴリズムの議論は早期警告システム(EWS)の文脈で語られてこなかったが、本研究では公正なアルゴリズムとEWSの関連性を整理したという点である。もう一つは、公正なアルゴリズムの議論がこれまで統計的な議論に終始していたのに対し、それを公平な教育の関連性との観点から議論した点にある。3点目に、これまで高等教育における公正なアルゴリズムの研究は主に欧米の大学のデータを用いて行う研究が多かったが、そこに日本のデータを用いることによって、日本の観点を提供したということが挙げられる。

研究成果の概要（英文）：Many colleges use AI-powered early warning systems (EWS) to provide support to students as soon as they start their first semester. But there are concerns about whether an algorithm underlying the EWS technology can make fair and unbiased decisions so early in a student's college experience. To examine the algorithm's fairness, we developed a machine learning algorithm that predicts first-term college GPAs by using data from a mid-sized Japanese private university. Our research offers two major findings. Firstly, deploying EWS during the initial phase of the first semester may lead the algorithm to make discriminatory decisions. Secondly, achieving algorithm fairness in a statistical sense does not necessarily lead to fair education outcomes.

研究分野：教育工学

キーワード：ラーニングアナリティクス　Institutional Research　機械学習　学生リスク予測モデル

１．研究開始当初の背景

近年，AI を活用したサービスの社会実装が様々な分野で拡大している．教育分野でも，大規模なデータを 分析・活用して教育の改善を目指す「LA（Learning Analytics）」分野が急速な発展を遂げている（近藤・ 畠中 2016）．LA の目的は，教育機関が収集したデー タを活用することで，教育の質や，学生の定着率，学習意欲を向上させることにある（Sclater et al. 2016）．

LA の一環として，近年，諸外国の高等教育機関で 積極的に実装されているものに「EWS（Early Warning Systems）」がある．EWS は，学生の退学や成績不振 などのリスクを，教育機関が保持する学生に関するデータを用いて AI で予測し，リスクが顕在化する前に支 援を行うという，AI を駆使した教学支援サービスの一 環である（Plak et al. 2022）．日本において実装している大学は限定的であるが（松田・渡辺 2018），欧米の大学では多くの大学で活用されている（Hanover Research 2014）．

このように AI サービスは魅力的である一方，AI が特定のグループに属する人への差別を助長している危険性が，近年様々な分野で指摘されている（Angwin et al. 2016）．人間中心の AI 社会原則において「人々 がその多様なバックグラウンドを理由に不当な差別を されることなく，全ての人々が公平に扱われなければ ならない」（内閣府 p.11）とされているように，倫理的な AI の運用への社会的要請が昨今高まっている．

AI が差別を行うのではないかという懸念は，AI により介入対象者を決定しようとする EWS にとっても無関係ではない．EWS が，学生のリスクを予測する際に差別を行う危険性，そしてその解決方法に関する議論は，「アルゴリズムの公平性」という観点から，欧米を中心にここ数年で急速に進んでいる（Yu et al. 2020, 2021; Kung & Yu 2020; Doroudi & Brunskill 2019; Hutt et al. 2019; Jiang & Pardos 2021; Kleinberg et al. 2018 など）．しかし、差別を抑制した学生リスク予測アルゴリズムの開発研究は未だ進行中である。本研究はこれまでの「アルゴリズムの公平性」先行研究に連なるかたちで、差別を抑制した学生リスク予測アルゴリズムの開発をおこなっていく。

２．研究の目的

本研究は、世界中の高等教育機関で実装中の AI による大学生の学びに関わるリスク予測モデルを、属性間の統計差別という視点から検証し、そしてそこから得られた知見をもとに属性の影響を抑制した予測モデルの開発を目的とする。本研究の核心をなす問いは、「国内外の高等教育機関で実装中の AI 学生リスク予測モデルは特定の学生が不利益を被る仕組みになっているのではないか？」である。モデルが複雑化する中でリスク判定基準の不透明さが強まっているが、AI の負の側面に対しては議論が進んでいない。本研究では、ブラックボックス化しているモデルの構造を属性間統計差別という視点からの分析を通して解明し、新たな学生リスク予測モデルを開発することによって、学術的貢献を目指す。具体的には、次の３つの研究目的をたてた。

研究 1

近年、予測分析に関する文献では、学生の結果予測アルゴリズムの公平性を実証的に調査する研究が徐々に増えてきている（Yu et al.、2020; Yu et al.、2021; Jiang and Pardos、2021; Kung and Yu、2020; Hutt et al.、2019; Kleinberg et al.、2018）。しかし、これらの研究のいずれも EWS の早期導入に関連する公平性の懸念には取り組んできていない。この普及している EWS がバイアス、不平等、または差別的な実態につながる可能性があるかどうかを、文献と実践者の双方が考慮することは重要である。本研究は、EWS の初期学期に使用される予測アルゴリズムの公平性を検証することで、この議論を一歩進めることを目指している。この重要な側面に焦点を当てることで、高等教育におけるより公正な EWS の開発についての示唆を提供していく。

研究 2

学生の不合格リスクを予測する際に人工知能を使用すると、しばしば 0 から 100%の連続変数が生成される。しかし、学生への介入の必要性を判断するためには、この連続変数をリスクの有無を示す 2 値変数に変換する必要がある。しかし、この層別化プロセスでは、データサイエンティストなどの人間が閾値を使用して判断を下さなければならない。しかし、閾値はしばしば任意に選択され（Wynants et al.、2019）、その値はモデルの予測パフォーマンスに大きな影響を与える可能性がある（例：Freeman & Moisen、2008; Hernández-Orallo et al.、2012）。本研究は、最適な閾値を選択する方法が、学習分析の文脈において正確性だけでなく、アルゴリズムの公平性にも影響を与えることを示すことを目的としている。

研究 3

本研究の目的は，高等教育サービスの教学支援として EWS を導入した際，EWS で用いられる予測モデルの公平性が，そのサービスの最終目的である教育の公平性にどのように影響を与えるのか検証することである．また，先行研究で，EWS で用いられる予測モデ ルの公平性が議論の対象となっているが，予測モデルの公平性が教育の公平性を達成するために必要である のか検討する．

３．研究の方法

研究 1

本研究では、学生の一学期の大学 GPA を予測する機械学習アルゴリズムを構築した。これは、多くの大学が重要視している最も早い学業パフォーマンス指標の一つである（Gershenfeld et al.、2016）。データは、日本の中規模の都市型私立大学から取得した。予測アルゴリズムとして、一般的に XGB（Chen and Guestrin、2016）として知られる exTreme Gradient Boosting を使用した。私たちは「早期段階」として、初回入学から 6 週間以内を定義し、Tinto（1993）に従って、早期警告の理想的なタイミングは第一学期の 5 週目から 6 週目の間であるとの主張に基づいている。その時点で利用可能なデータを使用して、一学期の GPA が 2.0 未満の学生の予測モデルを構築し、リスクのあるグループと低リスクのグループの間で予測モデルの公平性を比較した。アルゴリズムの公平性を「分類の均等性」という用語で定義し、予測パフォーマンスのテストサンプル測定結果がサブグループ間で等しいことを意味する（Corbett-Davies and Goel 2018）。文脈におけるアルゴリズムの公平性を理解するために、二学期 GPA を予測する別のモデルを比較対象とした。このモデルでは、二学期開始前に利用可能なデータとして一学期の GPA が予測変数として使用した。また、解釈が困難とされる AI 予測アルゴリズムの内部構造を理解するために、SHapley additive explanations（SHAP）としてもよく知られる SHAP という手法を使用し、特徴変数がモデル内でどのように相互作用するかを調査した（Lundberg and Lee、2017）。

研究 2

日本の私立大学の新入生データを使用して、卒業までにどの学生が卒業するかを予測するモデルを構築し、その予測パフォーマンスが男性と女性の間でどの程度異なるかを比較した。予測モデルとして XGB（勾配ブースティング決定木アルゴリズム）を使用し、そのパフォーマンスを optuna（Akiba et al.、2019）を使用して最適化した。そして 6 つの異なる閾値選択方法を使用した。閾値の選択手法として、50%の固定閾値、Youden 指数、性別に応じた Youden 指数、女性の真陽性率に合わせたカットオフポイントの調整、男女の全体の再現率に合わせたカットオフの調整、およびそれぞれの男性と女性の F1 スコアを最大化するカットオフポイントの使用が含まれた。

モデルの予測の公平性を評価するために、3 つの標準的な検証指標を使用した：適合率（以下 ACC と呼ぶ）、真陽性率（TPR）、真陰性率（TNR）。これらの指標は、学習分析の文献でアルゴリズムのバイアスに対処するために一般的に使用されている（例：Yu et al.、2020 年、2021 年; Jiang & Pardos 2021）。データを 8：2 の比率でトレーニングセットとテストセットに分割し、5 つの交差検証法を使用した。訓練データを使用してモデルを構築し、教師データを使用してモデルの予測パフォーマンスを評価した。このプロセスを 20 回繰り返し、検証指標の信頼性のある標準誤差を得た。100 回のテスト結果（5 つのフォールドを 20 回の反復で実行）の平均パフォーマンスと標準誤差を提示し、モデルの予測の公平性を判断した。

研究 3

EWS で用いられる予測モデルの公平性として，本研究では，Corbett-Davies & Goel（2018）で定義され，Yu et al.（2021）で用いられた，公平なアルゴリズムの定義を用いる。 1 つ目は，機械学習モデルが保護属性を説明変数に含むか否かによって定義される「Anti-classification」、もう 1 つは，保護属性に属する学生と属さない学生のグループ間の予測精度の差によって定義される「Classification Parity」を用いた。

本研究では，まず，学生の 4 年終了時 GPA を予測する機械学習モデルを構築する。そして，モデルを用いて学生の成績不振のリスクを予測し，リスクが高いと判定された学生を介入の対象者とする。その際，「Anti-classification」に従い，保護属性を説明変数に含まないモデルと含むモデルの 2 つのモデルを構築する。また，それらの予測モデルの公平性の評価には，「Classification Parity」の定義を用いる。 次に，2 つのモデルで介入対象者とされた学生に対し効果のある介入を行ったと仮定し，介入前後の歴史的に有利な立場にある学生グループと不利な立場にある学生グループの 4 年終了時 GPA を比較する。そして，介入前のグループ間の 4 年終了時 GPA の格差と介入後のグループ間の 4 年終了時 GPA の格差を比較することで，介入が Kizilcec & Lee（2020）が示したいずれのパターンに該当するかを確認する。それにより，EWS

で用いられる予測モデルの公平性が，教育の公平性とどのような関係にあるのかを検証する。

４．研究成果

研究 1

1 学期の GPA を予測するモデルでは、高リスクグループに属する学生に対して、低リスクグループの学生と比較して有意に高い再現率(Recall)を示した。一方で、低リスクグループの学生に対しては特異度（specificity)が大幅に高くなった。この結果から、このモデルは、リスクグループに属す学生に対してはリスクスコアを一貫して増加させ、低リスクグループに属す学生に対しては減少させる傾向があることを示唆している。この傾向は、個々の学生が属すグループの典型的な特徴をその個人が有しているかどうかに関係なく生じる。一方、2 学期の GPA を予測するモデルは、一学期の GPA を予測変数として取り入れることで、このようなグループ差別的なリスク計算が起こりにくくなることがわかった。一学期の GPA は、二学期の GPA と強い相関関係がある。一学期の GPA が結果変数の予測指標として優れているため、二学期の GPA 予測モデルは、リスクグループに所属するかどうかというメンバーシップ変数に依存する必要が少なくなる。その結果、このモデルは、差別的で不正確な判断をすることが少なくなる。総合的に、本研究は、一学期のデータで構築された学生のリスク予測モデルは、精度の問題だけでなく、公平性の問題も存在することを示した。

研究 2

本研究は、閾値選択方法によってモデルの公平性が大幅に異なることを発見した。私たちの場合、性別に応じた Youden 指数が最も公平な結果を示したが、固定値の 50%では最も公平性に欠ける結果が得られた。先行文献は、閾値の値によって予測の正確性が大きく変動することを示唆している（Freeman & Moisen 2008; Hernández-Orallo et al. 2012)。本研究の分析は、学習分析の文脈においても同様の議論がアルゴリズムの公平性に適用されることを示唆している。本研究は、データサイエンティストが学生のリスク予測モデルを開発する際に、正確性と公平性を考慮し、閾値を任意に選ぶのではなく、データドリブンな形で選択することが大事である点を示唆している。

研究 3

本研究では，高校ランクを説明変数に含む Model1 と含まない Model2 の 2 つのモデルを構築した．そして，「Anti-classification」と「Classification Parity」の 2 つの観点から各モデルの公平性を比較し，EWS で用いられる予測モデルの公平性が，教育の公平性，つまり教育格差にどのような影響を与えるのかについて検証を行った．結果，Model1 と Model2 で予測精度の大きな差はなかった． また，保護属性間での予測精度の差は，いずれのモデルも同程度あり，どちらのモデルでも「Classification Parity」は達成されないことがわかった．そして，Model1・Model2 ともに，教育格差の縮小につながることがわかった．これは，Yu et al.（2021）で考察された内容と同様であり，EWS で用いられる予測モデルが 「Anti-classification」 や「Classification Parity」を満たささない予測モデルであっても，教育の公平につながる可能性があることが示唆された．

参考文献

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July).Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/ machine-bias-risk-assessments-incriminal-sentencing. (参照 2023-02-03)

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Doroudi, S., & Brunskill, E. (2019, March). Fairer but not fair enough on the equitability of knowledge tracing. In Proceedings of the 9th international conference on learning analytics & knowledge (pp. 335-339).

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling, 217 (1-2), 48-58.

Hanover Research. (2014). Early alert systems in higher education. https://www.hanoverresearch.com/wp-content/uploads/2017/08/EarlyAlert-Systems-in-Higher-Education.pd (参照 2023-02-03)

Hernández-Orallo, J., Flach, P., & Ferri Ramírez, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal
of Machine Learning Research, 13, 2813-2869.

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019).Evaluating Fairness and Generalizability in Models Predicting OnTime Graduation from College Applications. International Educational Data Mining Society.

Jiang, W., & Pardos, Z. A. (2021, July). Towards equity and algorithmic fairness in student grade prediction. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 608-617).

Kizilcec, R. F., & Lee, H. (2020). Algorithmic fairness in education. arXiv preprint arXiv:2007.05443 (2020).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018, May). Algorithmic fairness. In AEA papers and proceedings (Vol. 108, pp. 22- 27).

Kung, C., & Yu, R. (2020, August). Interpretable models do not compromise accuracy or fairness in predicting college success. In Proceedings of the seventh ACM conference on learning@ scale (pp. 413-416).

Plak, S., Cornelisz, I., Meeter, M., & van Klaveren, C. (2022). Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment. Higher Education Quarterly, 76(1), 131-152.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education. London: Jisc.

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. International educational data mining society.

Yu, R., Lee, H., & Kizilcec, R. F. (2021, June). Should college dropout prediction models include protected attributes?. In Proceedings of the eighth ACM conference on learning@ Scale (pp. 91-100).

近藤伸彦, & 畠中利治. (2016). 学士課程における大規模データに基づく学修状態のモデル化. 教育システム情報学会誌, 33(2), 94-103.

内閣府. (2019). 人間中心の AI 社会原則. 統合イノベーション戦略推進会議.

松田岳士, & 渡辺雄貴. (2018). 教学 IR，ラーニング・アナリティクス，教育工学. 日本教育工学会論文誌, 41(3), 199-208.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

〔図書〕　計0件

〔産業財産権〕

〔その他〕

現在海外ジャーナルに投稿中

Yanagiura, T., Yano, S., Kihira, M., & Okada, Y. Early Warning Systems and Bias: Examining Algorithm Fairness for First-Term College Grade Prediction Models.

Yanagiura, T., Yano, S., Kihira, M., & Okada, Y. Does the Threshold Value Matter for Algorithm Fairness for Student's Risk Prediction Model?

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関