A Case Study on a Sustainable Framework for Ethically Aware Predictive Modeling

Thomas C.H. Lux[1], Stefan Nagy[1], Mohammed Almanaa[2], Sirui Yao[1], Reid Bixler[1]
Departments of Computer Science[1] and Civil Engineering[2]
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA
tchlux@vt.edu

***Abstract*—Large volumes of data allow for modern application of statistical and mathematical models to practical social issues. Many applications of predictive models like criminal activity heat mapping, recidivism estimation, and child safety scoring rely on data that may be incomplete, incorrect, or biased. Many sensitive social and historical issues can unintentionally be incorporated into predictions causing ethical mistreatment. This work proposes a mechanism for continuously mitigating model bias by using algorithms that produce predictions from reasonably small subsets of data, allowing a human-in-the-loop approach to model application. The benefits offered by this framework are twofold: (1) bias can be identified either statistically or by human users on a per-prediction basis; (2) data can be cleaned for bias on a per-prediction basis. A modeling and data management methodology similar to that presented here could strengthen the ethical application of data science and make the process of cleaning and validating data manageable in the long term.**

Keywords—ethics, predictive models, multi-layer perceptron, decision tree, nearest neighbor, framework, human-in-the-loop

I. Introduction

Data collection efforts are often aimed at providing or improving services for our communities and society as a whole. Large scale predictive models have coevolved with the explosion of available data since the start of the digital age and these models are practically necessary for success in domains as diverse as healthcare, employment, finance, and government. Creators of predictive models were motivated by a critical need to anticipate the future and to take action in advance. When the models succeed they can create life-saving value [1] or change economies [2], but when they fail they can reinforce historical social injustice [3] and leave at-risk children without needed help [4]. It is not surprising then that interpretability is an important research area for commonly used predictive models [5]–[7].

A model is often considered unethical when it violates the expected practices and procedures of a formal agency like a government or business. For example, a model built from anonymous historical data and designed to assist a business in effective employee hiring may recover protected demographic information on the race, sex, or religion of the applicant. However, Title VII of United States federal code prohibits the use of this information in employment decisions. The predictive model would be acting unethically and unlawfully, but the user (the business) may be entirely unaware.

Many researchers have unveiled clear signs of unequal and racial treatment of individuals by predictive models [3], [8], [9]. For example, Angwin et al. analyzed a well-known criminal risk assessment: COMPAS and concluded the outcomes of the model are racially biased [3]. They showed that "Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism, and white recidivists were misclassified as low risk 63.2 percent more often than black defendants." The U.S. government has acknowledged the enormous potential for negative impact in biased models and the White House published several reports highlighting the potential bias that could adversely affect individuals or groups [10]. Research is constantly underway that proposes ethical frameworks and potential solutions to protect principles of ethics and privacy [11]–[13].

In an effort to prevent biased and discriminatory decisions, policy makers and practitioners have classified unwanted discriminatory attributes as protected information that cannot be used as predictors in a model. However, a model made from anonymized unprotected data can become unethical even when it is not provided protected information. The model needs only to infer protected characteristics from the provided data in the prediction process [11], [14]. The possibility of protected inference makes developing unbiased models challenging, particularly when large amounts of data are involved. A burden is placed on users, who must identify ways to ensure predictions remain ethically sound. Generally, there are only two methods to ensure a model is not violating ethical expectations: either carefully monitor and analyze all predictions for ethical mistreatment, or guarantee that the data utilized by the model (and how that data is used to predict) is ethically sound. This work makes an argument for the latter, as a common class of models can make good predictions from human-digestible subsets of large-scale data.

This paper makes an argument for per-prediction ethical validation of data and models via local approximation methods that produce interpretable results. The improvements offered by this methodology are twofold: (1) bias can be identified either statistically or by human users on a per-prediction basis; (2) data can be cleaned for bias on a per-prediction (regular) basis. Modeling techniques similar to those presented in this work could not only strengthen the ethical application of data science, but also make the process of cleaning and validating data manageable in the long term.

II. Related Work

Algorithmic bias has long been a subject of research on machine learning [15]–[17]. Mitchell [18] initially defined machine learning bias as "any basis for choosing one generalization over another, other than strict consistency with the instances". Mooney [19] expanded this definition with the following assertions: (1) that every model bears some inherent bias, and (2) that detecting a model's bias requires comparison against others. In recent years, the broad adoption of machine learning has made algorithmic bias a factor in real-world data discrimination [3], [8], [9].

Several prior works have explored the problem of measuring algorithmic bias [20]–[22]. Calders & Verwer [23] formalized discrimination as, given an input characteristic, the unequal distribution of outputs for different groups. Calders-Verwer (CV) scoring is frequently [24]–[28] used to measure *group discrimination* [29], [30]. For example, if a loan classifier produces dissimilar outcomes for both sexes (e.g. $P(Y = loan \mid S = male) > P(Y = loan \mid S = female)$), then its CV score is measured as the difference in outcomes between those groups (e.g. $P(Y=loan \mid S=male)-P(Y=loan \mid S= female)$). Many works have since expanded on CV scoring for preventing discrimination [26], [31], [32]. Four general approaches exist: (1) *Suppression* – removing attributes most correlating with discrimination-sensitive attributes; (2) *Dataset "massaging"* – altering labels of some objects to mitigate unwanted classifier outcomes; (3) *Reweighting* – assigning data carefully chosen weights to lessen the degree of discrimination; (4) *Sampling* – under- and over-sampling certain groups to compensate.

A recent focus of machine learning and data science research has been on improving model interpretability [33], [34]. More specific goals include transparency for assessing model ethicality [35], [36], augmenting informativeness [37]–[39], and inferring causality [40]. An obstacle to model interpretability is the lack of transparency of black-box classifiers such as deep neural networks [41]–[43]. Post-hoc interpretability [33] (e.g. visualizations or explanations) represents a promising alternative, however, the formalization of "model-agnostic" methodologies remains an ongoing research problem [44].

III. Model Descriptions

In order to construct a predictive model from data, it is usually assumed that the phenomenon being predicted has some underlying function that can be approximated. Many algorithms exist for constructing approximations from data representing unknown functions. The following subsections roughly outline the mathematical formulation of both the (arguably more popular) *global* models and the (arguably less popular) *local* models analyzed in the present case study.

A. Global Predictive Models

Classic machine learning and data science techniques applied today often rely on solving a very specific problem. They create a global predictive model with the aim of capturing trends that exist across (hundreds of) thousands of examples. In general, these global models are constructed given real valued data matrix $X \in \mathbb{R}^{n \times d}$, a truth function $f : \mathbb{R}^d \to \mathbb{R}$, and *labels* $f(x^{(i)})$ for row vectors $x^{(i)} \in X$, $1 \leq i \leq n$. These models find the solution to

$$\min_P \| \hat{f}_P(X) - f(X) \|,$$

where $\hat{f}_P : \mathbb{R}^d \to \mathbb{R}$ is the parametric approximation, $f(X)$ is used to denote the vector with components $f(X)_i = f(x^{(i)})$ and $\| \cdot \|$ is an appropriate measure. The labels may be real numbers, like probability of recidivism estimates, or categories such as "safe" or "not safe" for an at-risk child.

The difficulty with these models is that the minimization search which identifies the parameters for the model is performed over *all* data. Whenever it is time to explain a prediction, the answer is often "all data was used to capture this trend". The models of this form that will be applied are a multilayer perceptron (MLP) and a decision tree (DT).

1) *Multilayer Perceptron:* The neural network is a well-studied and widely used method for both regression and classification tasks [45]. When using the rectified linear unit (ReLU) activation function [46] and training with the BFGS minimization technique [47], the model built by a multilayer perceptron uses layers $l : \mathbb{R}^i \to \mathbb{R}^j$

$$l(u) = \left(u^t W_l\right)_+,$$

where W_l is the i by j weight matrix of layer l. In this form, the multilayer perceptron produces a piecewise linear model of the input data. The computational complexity of training a multilayer perceptron is $O(ndm)$, where m is determined by the sizes of the layers of the network and the stopping criterion of the BFGS minimization used for finding weights. This paper uses the scikit-learn MLP regressor [48], a single hidden layer with 100 nodes, ReLU activation, and BFGS for training.

2) *Decision Tree:* The decision tree is used because of the relatively straightforward interpretation of the prediction process. Model construction is out of the scope of this description, but is a well-studied process [49]. A prediction at a point $z \in \mathbb{R}^d$ for a decision tree constructed over a real vector space is made by traversing nested axis-aligned conditionals of the form

$$\hat{f}(z) = \hat{f}(z | z_{k^{(1)}} \geq v^{(1)}, \ldots).$$

This paper uses the scikit-learn Decision Tree regressor [48], no maximum depth or number of nodes, and the Gini impurity measure of information gain.

B. Local Predictive Models

The construction of approximation functions $\hat{f}_P : \mathbb{R}^d \to \mathbb{R}$ as described for global models can instead be approached on a per-prediction basis. A model is henceforth referred to as *local* when any prediction made at a point $z \in \mathbb{R}^d$ is only a function of a set of points $L \subset X$, where membership in L is determined by a distance metric. The advantage of using a local model is a more compact description of how a prediction is made that is derived from a manageable subset of all known data. *Local* models become particularly useful when predictions regard ethically sensitive issues and need to be rigorously evaluated for bias. The source data for any prediction can be checked on the spot for fairness of representation against any number of protected attributes.

The following sections describe the two local approximation techniques that will be used to predict recidivism likelihood in this work.

1) Nearest Neighbor: A well-studied technique for classification and approximation is the nearest neighbor algorithm [50]. This algorithm (using the 2-norm to measure distance) will represent a baseline for comparison because it is the most mathematically simple local model in this study. A prediction is made for Nearest Neighbor at point $z \in \mathbb{R}^d$ by

$$\hat{f}(z) = f\left(argmin_{x^{(i)} \in X} \|z - x^{(i)}\|_2\right).$$

This approximation technique can be applied in a wide range of applications, however it must be noted that the approximation surface it produces is not C^0 (continuous in value). An extension of this model that is not applied in this work, but could yield useful results is the k nearest neighbor algorithm. This is further mentioned in Section V.

2) Voronoi Mesh: While nearest neighbor inherently utilizes the convex region $v^{x^{(i)}}$ (Voronoi cell [51]) consisting of all points closer to $x^{(i)}$ than any other point $x^{(j)}$. The Voronoi mesh [52] smooths the nearest neighbor approximation by utilizing the Voronoi cells to define support via a generic basis function $V : \mathbb{R}^d \to \mathbb{R}_+$ given by

$$V^{x^{(i)}}(y) = \left(1 - \frac{\left\|y - x^{(i)}\right\|_2}{2\ d(y \mid x^{(i)})}\right)_+ ,$$

where $x^{(i)}$ is the center of the Voronoi cell, $y \in \mathbb{R}^d$ is an interpolation point, and $d(y \mid x^{(i)})$ is the distance between $x^{(i)}$ and the boundary of the Voronoi cell $v^{x^{(i)}}$ in the direction $y - x^{(i)}$. $V^{x^{(i)}}(x^{(j)}) = \delta_{ij}$ and $V^{x^{(i)}}$ has local support. While $V^{x^{(i)}}(x^{(i)}) = 1$, the 2 in the denominator causes all basis functions to go linearly to 0 at the boundary of the twice-expanded Voronoi cell. Note that this basis function is C^0 because the boundaries of the Voronoi cell are C^0. In the case that there is no boundary along the vector w, the basis function value is always 1.

While the cost of computing exact Voronoi cells for any given set of points grows exponentially [53], the calculation of d is linear with respect to the number of control points and dimensions. Given any center $x^{(i)} \in \mathbb{R}^d$, set of control points C ⊆ X, and interpolation point $y \in \mathbb{R}^d$, $d(y \mid x^{(i)})$ is the result of

$$\max_{c \in C \setminus \{x^{(i)}\}} \frac{\left\|y - x^{(i)}\right\|_2}{2} \ \frac{y \cdot \left(c - x^{(i)}\right) - x^{(i)} \cdot \left(c - x^{(i)}\right)}{c \cdot \left(c - x^{(i)}\right) - x^{(i)} \cdot \left(c - x^{(i)}\right)}.$$

The resulting algorithm is capable of producing predictions in $O(n^2 d)$ computation time, which is relatively fast for all but tens of millions of points. Most importantly there is no *training* for this algorithm, so as data is updated the fundamental model itself is changed accordingly.

IV. Data and Results

A case study is presented to demonstrate the comparative performance of *local* models versus *global* models on an ethically sensitive prediction task. The comparison rests on a 3-Year Recidivism for Offenders Released from Prison in Iowa dataset, publicly available on the U.S. Government open data (data.gov) website. This dataset reports whether or not an offender is convicted of another crime within three years of being released from prison. This dataset naively provides a classification task (predict whether or not a released offender would recidivate) or, as will be considered here, it also provides a regression task (predict the probability of recidivism). By collapsing all entries with identical predictor values, this work converts each offender description into a collection of matching individuals with a single real-valued estimate for the probability of recidivism.

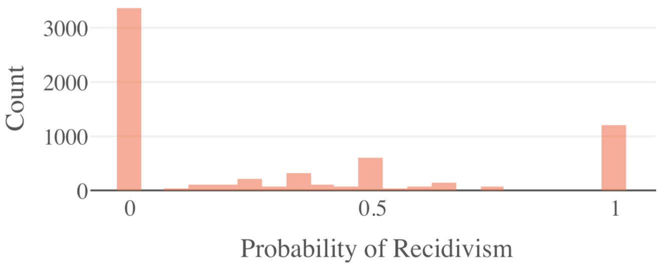


Fig. 1. A 20-bin histogram of the values for recidivism likelihood in the regression task composed of 6,632 samples.

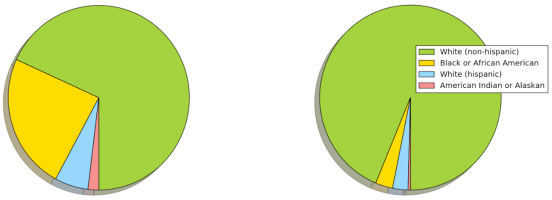


Fig. 2. The distribution across ethnicity groups of the dataset (left) and the actual population in Iowa (right)

This dataset provides 21,646 instances, among which 14,619 instances do not recidivate and 7,027 instances do recidivate. The distribution of ethnicities in this data are distinctly different from the ethnic distribution for the actual population of Iowa in 2016 (See Figure 2). There exists obvious gaps especially for White non-Hispanic (67.4% in data *decreased* from 88.7% of actual population) and Black or African American (23.6% in data *increased* from 2.9% of actual population).

The data provides 17 unique pieces of information related to each individual, including the crime committed in the case of recidivism (blank for non-recidivating individuals). The relevant features to this prediction task can be observed in Table I. The omitted features are either not relevant to the prediction task, or are uniquely determined by one of the chosen features. For this work the "Race – Ethnicity" is reserved as protected attribute (not available to models) and is later used to evaluate bias in predictions. Any categories with fewer than 100 samples are discarded in order to reduce the resulting dimension and prevent predictions from being made with statistically insignificant amounts of supporting data. Categorical features are encoded with c unique categories into

TABLE I. THIS TABLE PROVIDES A SAMPLE OF THE VALUES FOR FEATURES RELEVANT TO THE RECIDIVISM PREDICTION

FEATURE NAME	VALUES
Sex	F, M
Age at Release	Under 25, 25-34, 35-44, 45-54, 55 and Older
Convicting Offense Classification	Aggravated Misdemeanor, Serious Misdemeanor, Sexual Predator Community Supervision, etc.
Convicting Offense Subtype	Murder, Alcohol, Weapons, Drug Possession, Assault, Traffic, Burglary, Forgery/Fraud, Animals, Theft, etc.
Release Type	Special Sentence, Parole Granted, Discharged - Expiration of Sentence, Released to Special Sentence, etc.
Main Supervising District	1JD, 2JD, 3JD, 4JD, 5JD, etc.
Race – Ethnicity	White (non-Hispanic), Black or African American, White (Hispanic), American Indian or Alaskan

$\mathbb{R}^{c-1}$ by mapping each category onto one of the vertices of a regular simplex centered at 0 where all vertices v satisfy $\|v\|_2 = 1$. The feature named "Age At Release" was tested as both the mean age of the range and as a mapped categorical. Results demonstrated insignificant differences in outcome, so the numerical mean age is used for experiments.

In all experiments, predictions are evaluated against the labeled recidivism probability while the race & ethnicity information is used to evaluate model bias. After preprocessing, the original 21, 018 instances with 17 features are reduced to 6, 632 in a 50-dimensional real vector space. The distribution of recidivism probabilities can be seen in Figure 1.

In order to estimate the performance of the different algorithms, *k-fold* cross validation as described in [54] with k = 10 is used. All algorithms are given the same ten folds of randomized training and testing data in order to maintain comparative fairness. Note that in this scheme there will be exactly one prediction made for each data point, meaning all analysis of results is done with the same sized data as described in Section IV.

Figure 3 displays the evaluation of all algorithms. The Voronoi Mesh and Neural Network produce the best results. The prediction outcomes are promising, demonstrating that 50% of recidivism likelihood predictions have less than a 16% absolute error without any attempt at problem-specific tuning.

V. DISCUSSION

As can be observed in Figure 3, the Voronoi Mesh (VM) algorithm can make a prediction based on roughly $2d$ (~100) points from data and compete with the global fitting MLP that uses all data (~7K) points. This demonstrates that local predictive models are capable of producing equally accurate predictions when compared with global predictive models with far more compact support from data.

The benefit of using the VM (or any local model) to make predictions in ethically sensitive applications is that every prediction has a manageable set of source data that describes how a prediction is produced. Addressing the two points in Section I, statistical tests can be run on these source data points to reduce prediction bias in desirable ways (e.g. data could be filtered until the predictive population matches the demographics of the state). Human-readable source for a prediction also opens the door to regular validation of source data, an important aspect of model maintenance.

In order to maintain ethical conscientiousness in predictive modeling, it is vital that predictions can be audited for fairness and adherence to standards. Local predictive models provide a meaningful avenue for pursing the legal right to a representative sample, and the legal right to a fair prediction.

This case study provides only a glimpse at the prospective application of local models. There are far more advanced (and potentially more accurate models) with very similar properties: Delaunay triangulations (simplicial meshes) [55], k nearest neighbor, Linear Shepard [56] (and other Shepard methods), and Box Spline Meshes [52], to name only a few.

VI. CONCLUSION

This paper presents an argument for the application of local models to ethically sensitive prediction tasks. Sample results demonstrate that algorithms which rely only on local support are capable of producing predictions of comparable accuracy to popular global techniques while maintaining an enhanced level of predictive transparency. The potential for operating under more concise legal definitions and meaningful statistical analyses further supports the implementation of local prediction methodologies. A recidivism case study demonstrates that the use of more human-interpretable models for prediction could not only strengthen the ethical application of data science, but also make the process of cleaning and validating data manageable in the long term.

REFERENCES

[1] N. Tomasˇev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," Nature, vol. 572, no. 7767, p. 116, 2019.

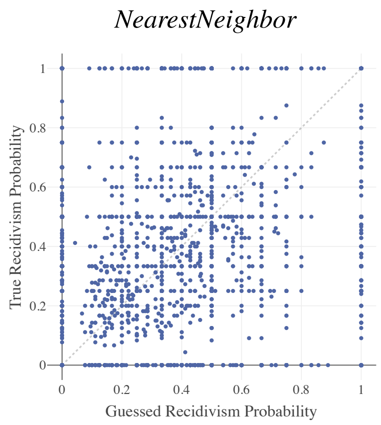
[2] E. Siegel, Predictive analytics: The power to predict who will click, buy, lie, or die. John Wiley & Sons Incorporated, 2016.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016."

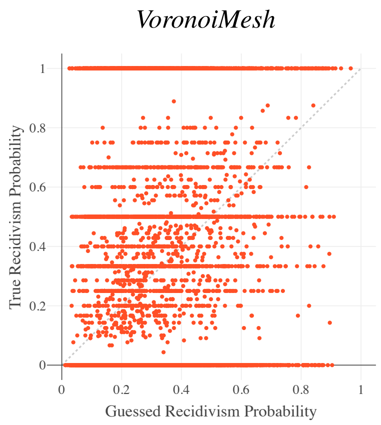
[4] V. Eubanks, "A child abuse prediction model fails poor families," Wired Magazine, 2018.

[5] A. Vellido, J. D. Mart́ın-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in ESANN, vol. 12. Citeseer, 2012, pp. 163–172.

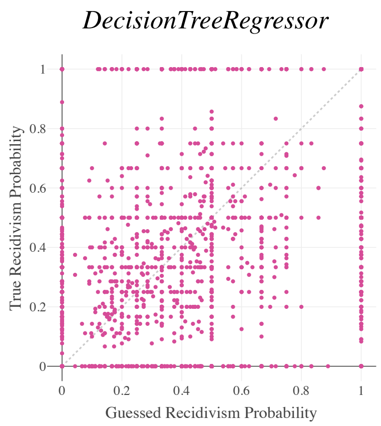
[6] F. Doshi Velez and B. Kim, "Considerations for evaluation and generalization in interpretable machine learning," in Explainable and Interpretable Models in Computer Vision and Machine Learning. Springer, 2018, pp. 3–17.


NearestNeighbor

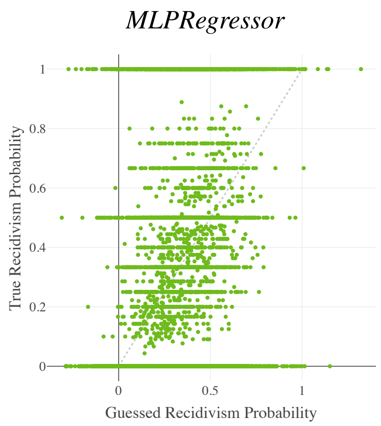
Min	0.000	0.226	White - NH
25th	0.033	0.250	Black - NH
50th	0.167	0.200	White - H
75th	0.400	0.167	AI or NA
Max	1.000	0.306	Asian or PI


VoronoiMesh

Min	0.000	0.148	White - NH
25th	0.072	0.161	Black - NH
50th	0.156	0.125	White - H
75th	0.291	0.129	AI or NA
Max	0.974	0.139	Asian or PI


DecisionTreeRegressor

Min	0.000	0.250	White - NH
25th	0.040	0.286	Black - NH
50th	0.187	0.246	White - H
75th	0.433	0.264	AI or NA
Max	1.000	0.257	Asian or PI


MLPRegressor

Min	0.000	0.147	White - NH
25th	0.068	0.153	Black - NH
50th	0.145	0.134	White - H
75th	0.279	0.113	AI or NA
Max	1.273	0.114	Asian or PI

Fig. 3. These four plots show the true recidivism probability versus guessed recidivism probability for each of the regression techniques. The top two regression algorithms make predictions based only on local data while the bottom two algorithms are global fitting techniques. The left vertical table beneath each figure displays the percentiles of absolute errors when predicting the probability of recidivism with that algorithm. The right table beneath each figure shows the median error in recidivism likelihood for those predictions which were over-estimated (false positives) broken down by race. Notice that without additional constraints, the neural network produces predictions outside of the range [0,1].

[7] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," arXiv preprint arXiv:1808.00033, 2018.

[8] G. D. Squires, "Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas," Journal of Urban Affairs, vol. 25, no. 4, pp. 391–410, 2003.

[9] M. A. Stoll, S. Raphael, and H. J. Holzer, "Black job applicants and the hiring officer's race," ILR Review, vol. 57, no. 2, pp. 267–287, 2004

[10] M. Smith, D. Patil, and C. Mun˜oz, "Big data: A report on algorithmic systems, opportunity, and civil rights," White House Report, Executive Office of the President, 2016.

[11] S. DeDeo, "Wrong side of the tracks: Big data and protected categories," arXiv preprint arXiv:1412.4643, 2014.

[12] E. Vayena, U. Gasser, A. B. Wood, D. O'Brien, and M. Altman, "Elements of a new ethical framework for big data research," 2016.

[13] N. E. Kass, "An ethics framework for public health," American journal of public health, vol. 91, no. 11, pp. 1776–1782, 2001.

[14] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 560–568.

[15] K. Kirkpatrick, “Battling algorithmic bias: how do we ensure algorithms treat us fairly?” Communications of the ACM, vol. 59, no. 10, pp. 16–17, 2016.

[16] S. Hajian, F. Bonchi, and C. Castillo, “Algorithmic bias: From discrimination discovery to fairness-aware data mining,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016, pp. 2125–2126.

[17] R. Baeza-Yates, “Data and algorithmic bias in the web,” in Proceedings of the 8th ACM Conference on Web Science. ACM, 2016, pp. 1–1.

[18] T. M. Mitchell, The need for biases in learning generalizations. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Jersey, 1980.

[19] R. J. Mooney, “Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning,” arXiv preprint cmp-lg/9612001, 1996.

[20] A.RomeiandS.Ruggieri,“Amultidisciplinarysurveyondiscrimination analysis,” The Knowledge Engineering Review, vol. 29, no. 5, pp. 582–638, 2014.

[21] F. Kamiran, I. Zˇliobaite˙, and T. Calders, “Quantifying explainable discrimination and removing illegal discrimination in automated decision making,” Knowledge and information systems, vol. 35, no. 3, pp. 613– 644, 2013.

[22] I. Zliobaite, “A survey on measuring indirect discrimination in machine learning,” arXiv preprint arXiv:1511.00148, 2015.

[23] T. Calders and S. Verwer, “Three naive bayes approaches for discrimination-free classification,” Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277–292, Sep 2010. [Online]. Available: https://doi.org/10.1007/s10618-010-0190-x

[24] T. Kamishima, S. Akaho, and J. Sakuma, “Fairness-aware learning through regularization approach,” in Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, 2011, pp. 643– 650.

[25] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-aware classifier with prejudice remover regularizer,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2012, pp. 35–50.

[26] S. Galhotra, Y. Brun, and A. Meliou, “Fairness testing: testing software for discrimination,” in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. ACM, 2017, pp. 498–510.

[27] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” in Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012, pp. 924–929.

[28] I. Zˇliobaite, F. Kamiran, and T. Calders, “Handling conditional discrimination,” in Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011, pp. 992–1001.

[29] [29] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Learning fair classifiers,” arXiv preprint arXiv:1507.05259, 2015.

[30] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in International Conference on Machine Learning, 2013, pp. 325–333.

[31] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” arXiv preprint arXiv:1507.05259, 2017.

[32] F. Kamiran, T. Calders, and M. Pechenizkiy, “Techniques for discrimination-free predictive models,” in Discrimination and privacy in the information society. Springer, 2013, pp. 223–239.

[33] Z. C. Lipton, “The mythos of model interpretability,” arXiv preprint arXiv:1606.03490, 2016.

[34] C.-F. Juang and C.-Y. Chen, “Data-driven interval type-2 neural fuzzy system with high learning accuracy and improved model interpretability,” IEEE transactions on cybernetics, vol. 43, no. 6, pp. 1781–1795, 2013.

[35] B. Goodman and S. Flaxman, “European union regulations on algorithmic decision-making and a,” 2017.

[36] B. Herman, G. Proksch, R. Berney, H. Dawkins, J. Kovacs, Y. Ma, J. Rich, and A. Tan, “Data science for urban equity: Making gentrification an accessible topic for data scientists, policymakers, and the community,” arXiv preprint arXiv:1710.02447, 2017.

[37] B. Kim, “Interactive and interpretable machine learning models for human machine collaboration,” Ph.D. dissertation, Massachusetts Institute of Technology, 2015.

[38] L. Yang, A. M. MacEachren, P. Mitra, and T. Onorati, “Visually-enabled active deep learning for (geo) text and image classification: A review,” ISPRS International Journal of Geo-Information, vol. 7, no. 2, p. 65, 2018.

[39] T. Wu, X. Li, X. Song, W. Sun, L. Dong, and B. Li, “Interpretable r-cnn,” arXiv preprint arXiv:1711.05226, 2017.

[40] U. Syed and G. Yona, “Enzyme function prediction with interpretable models,” in Computational Systems Biology. Springer, 2009, pp. 373–420.

[41] J. Casillas, O. Cordo´n, F. H. Triguero, and L. Magdalena, Interpretability issues in fuzzy modeling. Springer, 2013, vol. 128.

[42] P. Cortez and M. J. Embrechts, “Using sensitivity analysis and visualization techniques to open black box data mining models,” Information Sciences, vol. 225, pp. 1–17, 2013.

[43] J. Krause, A. Perer, and K. Ng, “Interacting with predictions: Visual inspection of black-box machine learning models,” in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016, pp. 5686–5697.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 1135–1144.

[45] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural networks, vol. 2, no. 5, pp. 359–366, 1989.

[46] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for lvcsr using rectified linear units and dropout,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE, 2013, pp. 8609–8613.

[47] M. F. Møller, “A scaled conjugate gradient algorithm for fast supervised learning,” Neural networks, vol. 6, no. 4, pp. 525–533, 1993.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[49] J. R. Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81–106, 1986.

[50] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[51] G.L.Dirichlet,“U¨berdiereductionderpositivenquadratischenformen mit drei unbestimmten ganzen zahlen.” Journal fu¨r die Reine und Angewandte Mathematik, vol. 40, pp. 209–227, 1850.

[52] T. C. H. Lux, L. T. Watson, T. H. Chang, J. Bernard, B. Li, X. Yu, L. Xu, G. Back, A. R. Butt, K. W. Cameron, D. Yao, and Y. Hong, “Novel meshes for multivariate interpolation and approximation,” in Proceedings of the ACMSE 2018 Conference, ser. ACMSE ’18. New York, NY, USA: ACM, 2018, pp. 13:1–13:7. [Online]. Available: http://doi.acm.org/10.1145/3190645.3190687

[53] M. Dutour Sikiric´, A. Schu¨rmann, and F. Vallentin, “Complexity and algorithms for computing voronoi cells of lattices,” Mathematics of Computation, vol. 78, no. 267, pp. 1713–1731, 2009.

[54] R.Kohavietal.,“Astudyofcross-validationandbootstrapforaccuracy estimation and model selection,” in Ijcai, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[55] T. H. Chang, L. T. Watson, T. C. Lux, B. Li, L. Xu, A. R. Butt, K. W. Cameron, and Y. Hong, “A polynomial time algorithm for multivariate interpolation in arbitrary dimension via the delaunay triangulation,” in Proceedings of the ACMSE 2018 Conference. ACM, 2018, p. 12.

[56] W. I. Thacker, J. Zhang, L. T. Watson, J. B. Birch, M. A. Iyer, and M. W. Berry, “Algorithm 905: Sheppack: Modified shepard algorithm for interpolation of scattered multivariate data,” ACM Transactions on Mathematical Software (TOMS), vol. 37, no. 3, p. 34, 2010.