Exploring Fairness in Machine Learning for International Development

MIT D-Lab
Comprehensive Initiative on Technology Evaluation
Massachusetts Institute of Technology

January 2020

MIT D-Lab
designing for a more equitable world

Comprehensive Initiative on Technology Evaluation

USAID
FROM THE AMERICAN PEOPLE

About USAID

USAID is a leading international development agency and a catalytic actor driving development results. USAID's work advances U.S. national security and economic prosperity, demonstrates American generosity, and supports countries along a path to self-reliance and resilience. USAID believes the purpose of foreign aid should be ending the need for its existence, and provides development assistance to help partner countries on their own development journey to self-reliance – looking at ways to help lift lives, build communities, and establish self-sufficiency. Its efforts are both from and for the American people. USAID demonstrates America's goodwill around the world, increases global stability by addressing the root causes of violence, opens new markets and generates opportunity for trade, creates innovative solutions for once-unsolvable development challenges, saves lives, and advances democracy, governance, and peace.

The USAID Center for Digital Development (CDD) works to address gaps in digital access and affordability and to advance the use of technology and advanced data analysis in development. CDD pursues this mission by: 1) supporting the enabling environment that serves as a foundation for inclusive digital infrastructure and services and 2) building Agency capacity via technical trainings, toolkits, and guidance documents, and by building a network of Mission-based specialists. CDD works to foster market-led innovation and integrate digital technology, advanced data, and geographic analysis, and to align digital development best practices with the strategic planning and design of enterprise-driven programs across the Agency and with partners.

About MIT D-Lab | CITE

MIT D-Lab works with people around the world to develop and advance collaborative approaches and practical solutions to global poverty challenges. The program's mission is pursued through an academics program of more than 20 MIT courses and student research and fieldwork opportunities; research groups spanning a variety of sectors and approaches; and a group of participatory innovation programs called innovation practice.

This document is part of a series of reports produced by MIT CITE. Launched at the Massachusetts Institute of Technology (MIT) in 2012 with a consortium of MIT partners, CITE was the first-ever program dedicated to developing methods for product evaluation in global development. Located at MIT D-Lab since 2017, CITE is led by an interdisciplinary team and has expanded its research focus to include a broad range of global development topics.

Table of Contents

PREFACE 1

CHAPTER 1: INTRODUCTION TO FAIRNESS IN MACHINE LEARNING 2

CHAPTER 2: ML IN INTERNATIONAL DEVELOPMENT 5

CHAPTER 3: FAIRNESS CONSIDERATIONS IN DEVELOPING AN ML PROJECT 11

CHAPTER 4: BUILDING FOR FAIRNESS 18

CHAPTER 5: CONCLUSIONS 46

CASE STUDY 50

APPENDIX 62

On the cover: A clinician in Pune, India, evaluating a patient for pulmonary disease using a mobile stethoscope and mobile app. Photo: Rich Fletcher, MIT D-Lab

Authors

Yazeed Awwad (MIT); Richard Fletcher (MIT D-Lab), Daniel Frey (MIT Department of Mechanical Engineering, MIT D-Lab); Amit Gandhi (MIT Department of Mechanical Engineering, MIT D-Lab); Maryam Najafian (MIT Institute for Data, Systems, and Society); Mike Teodorescu (Boston College Carroll School of Management) (alphabetical).

Acknowledgments

CITE gratefully acknowledges the vast amount of time and energy invested by our partners at USAID including Aubra Anthony, Craig Jolley, Shachee Doshi, Amy Paul, and Maggie Linak. The outcomes of this project were greatly improved by their deep expertise in both the technical substance and the international development context.

This document also benefited from the inputs of participants in the workshop conducted at Ashesi University including Ayorkor Korsah and Boniface Akuku. They helped us to understand how a guide like this can be put to use and also taught us many valuable lessons about ethics, entrepreneurship, and education. The coauthors also thank Varun Aggarwal, Rohit Takhar, Abhishek Unnam of Aspiring Minds for early input on case studies.

The team also acknowledges the contributions of Elisabeth Andrews, who led us through the editing process and made recommendations on how to reorganize the content and make it more accessible. She also generated new content and provided valuable edits on the text. The team also thanks thank Nancy Adams for designing the document and laying out the content, making it more accessible to the reader.

CITE also acknowledges the valuable input provided by external reviewers who shared feedback and raised important questions, which made the document stronger. The team also acknowledges the contributions of Lily Morse of West Virginia University, who provided valuable feedback and contributed ideas.

Finally, the team is grateful to MIT D-Lab Associate Director for Research Kendra Leith for project management, managing relationships with our partners and advisors, organizing a workshop, and providing important advice on the development of this document.

This report was produced by the MIT D-Lab CITE at the Massachusetts Institute of Technology and made possible through the support from the United States Agency for International Development. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the United States Agency for International Development or the US Government.

Suggested Citation

Awwad, Y.; Fletcher, R.; Frey, D.; Gandhi, A.; Najafian, M.; Teodorescu, M. 2020. Exploring Fairness in Machine Learning for International Development. MIT D-Lab | CITE Report. Cambridge: MIT D-Lab.

Preface

This document is intended to serve as a resource for technical professionals who are considering or undertaking the use of machine learning (ML) in an international development context. Its focus is on achieving fairness and avoiding bias when developing ML for use in international development. This document provides guidance on choice of algorithms, uses of data, and management of software development. It also illustrates the application of this guidance through a case study. The focus is on developing ML applications, rather than procuring ready-made solutions, although many of the considerations outlined in this document are also relevant to ML procurement.

This document is meant to be accessible to a wide range of readers, but it does assume some prerequisite knowledge related to machine learning. It is recommended that readers have a basic foundation in computer science.

For a broader introduction to basic concepts of machine learning in the context of international development, readers are referred to USAID's companion document, *Reflecting the Past, Shaping the Future: Making AI Work for International Development* (*Making AI Work*[1]). Developed by the organization's Center for Digital Development, *Making AI Work* identifies issues that may be encountered when implementing ML in international development and provides a summary of findings on the appropriate applications of ML in these settings.

Development practitioners who are addressing fair and responsible use of AI and others concerned about the risks of using AI in development programs may benefit from reading *Making AI Work* before reading this document.

Whereas *Making AI Work* primarily targets development professionals working with technology partners, the present document serves to support technology professionals within the development context. The drafting team built upon *Making AI Work* by describing technical approaches for implementing ML projects in ways consistent with the published USAID guidance. The principles and practices described in this guide, in conjunction with those outlined in *Making AI Work*, aim to support the successful partnerships described by USAID:

> *[D]evelopment practitioners … must collaborate with technology experts to develop these tools for the contexts in which we work. … Many of the projects discussed in this report have involved collaboration between a "technology partner" and a "development partner." In some cases, the development partner may be based in a donor agency or implementing partner (e.g., as an activity or grant manager), while the technology partner is contracted to deliver an ML-dependent tool. Development-technology partnerships can also arise from situations with less formal distinctions. These include academic collaborations, co-creation efforts, or within an in-house interdisciplinary team.*

Drafting of this document was led by MIT D-Lab CITE at the Massachusetts Institute of Technology (MIT). This work was supported initially through USAID's Center for Development Research (CDR) and completed through partnership with USAID's Center for Digital Development.

1. Amy Paul, Craig Jolley, Aubra Anthony. *Reflecting the Past, Shaping the Future: Making AI Work for International Development*. (Washington: USAID, 2018) https://www.usaid.gov/sites/default/files/documents/15396/AI_ExecutiveSummary-Digital.pdf

Chapter 1: Introduction to Fairness in Machine Learning

This chapter introduces the concepts of fairness and bias and how they apply to the use of machine learning. It discusses ethical hazards associated with the use of ML in decision making, including the phenomenon of "ethical fading" in which the use of technology can obscure the ethical implications of these decisions.

The sidebar briefly defines artificial intelligence (AI) and machine learning (ML); for more background information on AI and ML, see *Reflecting the Past, Shaping the Future: Making AI Work for International Development*.

The focus of this guide is on fairness in supervised ML, in which accuracy can be defined and outcomes assessed with respect to a set of labeled training data. Unsupervised ML, by contrast, finds patterns in data without any human-defined examples of "correct" answers. Ethical considerations in unsupervised ML are also challenging but beyond the scope of this report.

Additionally, there are other characteristics of ML deployment, outside of algorithmic considerations, that can impact fairness more broadly but are outside the scope of this document. For instance, fairness considerations may include prioritizing local ML talent over larger technology companies based elsewhere. Teams may wish to ensure that the data being collected and classified are accessible to local populations, particularly if data are being collected with the help of local residents. Involving people impacted by the ML effort in problem definition is another important fairness principle beyond the technical aspects discussed here. The AI Principles[2] provided by the Organization for Economic Cooperation and Development (OECD) are an excellent resource for exploring further aspects of fairness.

A. Defining Fairness in ML

FAIRNESS
Just and equitable treatment across individuals and/or groups.

Whereas bias — the systematic favoring of one group over another — can be measured mathematically, fairness is a flexible and subjective concept that must be evaluated in light of the circumstances and goals of the machine learning project. The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) community offers the following description of the fairness principle:

> *"Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g., race, sex, etc)."*

BIAS
Systematically favoring one group relative to another. Bias is always defined in terms of specific categories or attributes (e.g. gender, race, education level). Many types of bias are socially or ethically undesirable.

The FAT community goes on to explain that it has left this term "purposefully under-specified" to allow it to be broadly applicable because "[a]pplying these principles well should include understanding them within a specific context."

This document takes a similarly flexible approach. Fairness is understood to refer to the pursuit of just and equitable outcomes. In other words, fairness avoids bias that perpetuates or reinforces existing social, economic, political, or cultural advantages or disadvantages. For example, if an algorithm is more likely to disqualify women applicants from receiving loans to start small businesses, regardless of the applicants' traits of creditworthiness, that algorithm could be said to be unfair (or unjust) in its treatment of women (or biased against them). However, it

2. OECD. *OECD Principles on AI.* (Washington: OECD, 2019) https://www.oecd.org/going-digital/ai/principles/

is also possible that, in the pursuit of fairness, an algorithm could deliberately introduce a bias as a means of redressing preexisting inequities. In the small-business loan example, a model could be designed to rate women's creditworthiness higher than that of men with similar credentials in order to help address longstanding barriers for women entrepreneurs. Therefore, while fairness and bias are related, it is not always a one-to-one correspondence.

Fairness may be understood differently in different contexts; it requires the adoption of specific ethical constructs and value system criteria that are used to judge a given set of outcomes and may depend on the rights and responsibilities of individuals determined by local laws or customs. For example, different types of tax systems take different views of fairness - should everyone pay the same amount? Or should they instead pay the same percentage of their wealth or their earnings? Or should those with more wealth pay a greater percentage than those with less? The answers depend on values and beliefs about how people in different circumstances should be treated.

Because different individuals and groups each have unique statuses, abilities, and challenges, an analyst — whether human or computer — must take into account all these factors in order to arrive at an outcome that can be understood as "fair." For example, in a project to develop roads, whose access needs should be considered? Should the roads be equally safe and accessible for people on foot, riding bicycles, or driving cars, or should the needs of people using some modes of transportation be prioritized over others? What about people with physical disabilities that may need more direct access to roadways than people who can walk through undeveloped areas to reach public transportation on a road? Should the plan be designed to provide a basic level of road access to as many people as possible, or to meet the varying needs in the community? The particular interpretation of fairness that is adopted may depend on the local culture, the local government, or the specific group that is in power.

Whether analyses are conducted by humans or computers, there are rarely perfect answers to complex real-world problems. One motivation for pursuing machine learning in international development is the hope that a computer algorithm has the potential to be objective and impartial. However, impartiality can also mean insensitivity to the nuances of the situation, the historical inequities that have resulted in certain groups having disadvantages, and the needs of the different people affected by development efforts.

B. How Fairness Can Fade with the Use of Technology

Research on human behavior suggests that there are important reasons to be concerned about how organizations using ML may lose track of the impact of computer programs' decisions on human beings. The increasing reliance on machine learning rather than human decision making can contribute to a phenomenon known as "ethical fading," wherein individual people or organizations suspend their ethical reasoning and make decisions based on other factors such as financial or practical considerations.[3] Machine Learning may be

ARTIFICIAL INTELLIGENCE (AI)
AI is a field dedicated to creating computers and computer software that are capable of intelligent behavior. In some cases, such goals can be pursued by programming a computer with rules that are understood by humans proficient in those same tasks.

MACHINE LEARNING (ML)
ML, by contrast, is a branch of artificial intelligence in which software learns how to perform a task without being explicitly programmed for the task by humans. In ML, the role of the programmer is to implement algorithms that structure the learning process, rather than encode intelligence through a series of rules. The intelligence of the machine emerges from the combination of learning algorithms and data used to train them.

3. Ann E. Tenbrunsel,, Messick, David, M. "Ethical Fading: The Role of Self-Deception in Unethical Behavior." *Social Justice Research* 17 (2004): 223-236.

used to distance organizational leadership from actions which can cause individual harm, separating the decision maker from the individuals affected, with the result that the ethical implication of the decision "fades" from the mind of the decision maker. With increased automation, impacts on individuals are much less visible, and ethical implications of decisions are more likely to be overlooked.

Another concern is that blame or fault for a problem that arises in an ML application may be assigned to an individual who is not ultimately responsible for the decision making. Elish introduced the term "moral crumple zone" to denote the results of the ambiguity introduced by layers of automation and distributed control.[4] Just as an automobile contains structures within its chassis that can absorb the energy of a collision, people in certain roles may find themselves accountable when something goes wrong in a semi-automated ML system. For example, in a loan program that uses ML to assess creditworthiness, the loan officer may find herself in the moral crumple zone, taking the blame when an application is unfairly rejected.

Ethical fading and moral crumple zones might be observed together, for example, in the use of an ML-backed system for guiding hiring decisions. Suppose a company adopts a new software system that rates applicants. Ethical fading might be a concern if the company's management uncritically accepts the ML system's recommendations to hire disproportionately from a certain group. Further, if the company makes a specific hiring decision that doesn't work out well, the blame is unlikely to be borne by the new software. The employee who ultimately wrote the employment contract may act as a moral crumple zone for the company.

In international development, although the intent is to aid communities, the broader goals of the organization (such as economic or technological development) can nevertheless become distanced from the individual people affected in ways that may cause hardship or distress for both the intended beneficiaries and the people tasked with implementing the program. When ethical fading occurs, individuals simply do not see ethics as their area of responsibility, often assuming that any ethical considerations are adequately addressed by the technology. In moral crumple zones, responsibility for ethical considerations are inappropriately assigned to actors who have limited control over decisions. How can ethical fading be mitigated in the context of a rise in automation? How can moral crumple zones be avoided?

A first step is to understand that automating decisions through the use of ML techniques will not necessarily improve equity in lending, housing, hiring, school admissions, and other decisions with potential life-changing implications. On the contrary, overreliance on ML can obscure biases existing in society and picked up by training data, as well as distance decision makers from those they impact. Rather than allowing accountability to fall to frontline implementers of ML-backed programs, accountability for results of ML implementation must be distributed across software engineers, those who collect and prepare data, those who place the software into use, and others involved in the ML process. Again, the first step is to acknowledge the challenge this situation poses. Additional steps are outlined in the following chapters.

The next chapter will examine some of the ways that ML is used in international development specifically and where bias can emerge in the different steps of an ML project.

---

4. Madeline Claire Elish. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5 (2019): 40-60.

Chapter 2: ML in International Development

This chapter introduces some of the ways in which machine learning is being used in international development and highlights key considerations and guiding principles for its use in these contexts.

A. Examples of ML Applications in International Development

This section highlights some examples of sector-specific applications of ML in international development and illustrates some of the emerging issues related to bias and fairness. While ML has been applied across many sectors in global development, this chapter highlights use cases in a select few sectors that serve as illustrative examples throughout this document. Readers interested in learning more about ML applications are encouraged to read *Artificial Intelligence in International Development*,[5] a recently published discussion paper by the International Development Innovation Alliance (IDIA), which provides a more comprehensive overview of ML applications and was used as a reference for this section.

Agriculture

In the agriculture sector, farmers and agriculture extension agents have used ML approaches such as ICRISAT/Microsoft to monitor soil quality, Plantix to identify plant diseases, and Apollo Agriculture to connect farmers to markets. Additionally, organizations have been using remote monitoring coupled with image processing to provide higher-quality agriculture insurance to farmers.

Education

In the education sector, students are using ML approaches that involve catered content to enable them to learn more effectively. For example, an ML program could use student performance on tests to identify knowledge gaps and provide dedicated content in weak areas. Bolo is using ML to promote literacy by providing local content in local languages, with real-time feedback. Educators are using similar ML technologies to automate the process of assessing student performance.

Governance

In governance, ML is being used to fill gaps in census data and other national surveys, allowing for better identification of community-level problems and allocation of resources. Starling Data is an example of this application could involve using remote sensing for better infrastructure planning. Additionally, private sector and watchdog organizations are using ML to process and interpret natural language to determine toxicity on social media with programs such as Jigsaw, indicative of whether civil or human rights violations are taking place in specific areas.

5. Results for Development. *Artificial Intelligence and International Development*. (The International Development Innovation Alliance, 2019): 1-8. https://observatoire-ia.ulaval.ca/app/uploads/2019/08/artificial-intelligence-development-an-introduction.pdf

Humanitarian

In humanitarian efforts, ML is being used in logistics and planning as well as in identifying and predicting conflict. ML has been used to process and interpret images to examine possible human rights violation sites so that humanitarian agencies can dedicate resources more quickly and effectively; similarly, AIDR has used satellite imagery and social media data to better respond to humanitarian crises. Immigration Policy Lab has applied ML to refugee data to determine where displaced populations could best be relocated.

Healthcare

Several organizations are working in healthcare diagnostics in international development, some of which are highlighted in later chapters. For example, iKure is a social enterprise based in India that provides healthcare solutions for rural communities. Its team has developed a wireless health incident monitoring system that helps connect patients to health centers through rural health workers using a hub-and-spokes model in which workers based at the health centers visit patients in surrounding communities. These rural health workers carry basic bio-medical devices and a mobile app, which allows them to diagnose, monitor, and track patients.

Workforce Development

Addressing barriers to employment is another emerging focus of ML in international development. Aspiring Minds is an Indian company focused on employment skills testing. In the Indian job market, employers tend to evaluate applicants based on their academic credentials rather than their skill sets. Unfortunately, this focus makes it difficult for people who have not had access to higher education to find work in India. Building on the belief that skills are ultimately more important than academic credentials for job success, Aspiring Minds has developed computerized tests and ML-backed assessments to determine applicants' strengths and match applicants with appropriate jobs. Motivated by similar challenges, Harambee, founded in South Africa, seeks to reduce the country's persistently high unemployment rate by using ML to match youth at high risk of long-term unemployment or underemployment with jobs suited to their skills and personality attributes.

Financial Services

Several organizations are using ML to determine the creditworthiness of individuals in areas where formal credit rating mechanisms do not exist. For example, Tala and Branch, used in East Africa and India provide low-interest microloans to individuals in exchange for access to their customer data via mobile payment platforms. These organizations then use ML to determine repayment rates for their users. In the solar sector, several social enterprises are providing solar-powered home systems using a pay-as-you-go model: solar equipment is leased to people with the expectation that they will pay for it over time in monthly installments. Some of these companies, such as Bboxx and Fenix in East Africa, have started using ML to determine creditworthiness based on solar asset ownership and payment history.

Many more sectors are also seeing increases in the use of ML. The following section introduces principles for responsible use of ML that can be applied in any sector.

B. Aiming for Fairness in ML in International Development

To help guide emergent efforts to leverage ML in international development, the following principles, or key considerations, for the development of machine learning can be used to assess and shape ML projects. While these ideals stem from the broader FAT community, they may differ slightly from the ways in which fairness and associated attributes are generally characterized in highly developed contexts. The descriptions provided here are rooted in how these principles present in developing markets specifically.

One important concept for ensuring fair outcomes are achieved is having a "human in the loop." This concept ensures that ML is integrated into human-powered workflows and processes. The technology aids analysis and provides recommendations, but humans ultimately make the decisions.

Many of these principles are interrelated and interdependent, and all contribute to increasing fairness.

KEY CONSIDERATIONS FOR RESPONSIBLE USE OF ML IN INTERNATIONAL DEVELOPMENT

- Equity
- Representativeness
- Explainability
- Auditability
- Accountability

Equity: *Has the ML model been tested to determine whether it disproportionately benefits or harms some individuals or groups more than others?*

As discussed in Chapter 1, it is important to ensure that ML is treating people equitably. Does a specific algorithm fail (misclassify) people belonging to certain groups more often? Does it misclassify different groups in different directions, so that some groups benefit disproportionately while others are disproportionately harmed? Do certain groups have different rates of false positives and false negatives?

Testing the result of algorithms against senstive variables such as gender, race, age, or religious affiliation can prevent the adoption of biased algorithms. These questions and algorithmic approaches to addressing them are discussed further in the coming chapters.

It is also important to understand that accuracy and equity are not necessarily correlated. Algorithms can be technically accurate, yet still inconsistent with the values that organizations want to promote when making decisions such as who should be hired and who should receive medical care. For example, an algorithm may accurately reflect that men in the population have better academic credentials than women, but this trend may be due to women historically being denied educational opportunities. Gaining an understanding of how these outcomes are derived and taking steps to mitigate them — such as measuring skills and abilities rather than academic credentials — is an important element in ensuring that inequitable algorithms are not widely adopted and used.

Representativeness: *Is the data used to train the ML models representative of the people who will be affected by the model's application?*

In order to evaluate representativeness, organizations should consider whether their ML model uses data that are representative of the context in which ML outputs will be deployed. Analysis of this principle should also incorporate representativeness by ensuring that local people and knowledge contribute to consideration of this question. That is, the intended users and beneficiaries of the ML application should be consulted in determination of appropriate training data.

As an example, consider a startup medical diagnostics company that is trying to build a remote diagnostic tool for the rural West African population. High quality, coded datasets from West Africa may not be available, so the startup uses a European dataset to train their models. Some diagnoses may be accurate, but there may be vital differences in the ways that certain diseases present in the West African or European context, which could lead to misdiagnoses that put individuals at risk.

Now consider if the startup instead uses a dataset based on rural East African patient data. While this dataset may more closely match the context for implementation, resulting diagnoses from the model may overlook diseases such as malaria and yellow fever, which tend to be more common in West Africa, again resulting in improper diagnosis.

Finally, consider a startup that uses a dataset from patients from the largest hospitals in West African countries. While this may seem like a good choice because the dataset captures information from a West African population, this dataset would be more representative of urban populations than rural populations, which could also result in improper diagnosis.

It may not be possible to find an exact match for the training data, but it is important to consider all of the ways in which the data may not be representative in order to arrive at the best option. The best option might not be using machine learning until more representative data is available.

Explainability: *Can individual predictions or decisions be explained in human-friendly terms?*

It is important to ensure that the application is explained to end-users in a way that effectively communicates how the outcomes were determined. Communication of how the model works and its limitations is important for whoever is using the model or its outcomes. Individuals and organizations seeking to apply ML outcomes without understanding the nuances of how the models make decisions may use the algorithm outputs inappropriately.

Increasingly, organizations are turning to "black box" machine learning approaches, whose inner workings can range from unintuitive to incomprehensible. In effect, only the computer knows how it arrived at its decisions. In the field of artificial intelligence, "explainability" generally refers to avoiding such "black box" approaches and instead ensuring that the sequence of operations can be understood by developers. For the international development context, the "explainability" term is broadened here to mean that the entire technology-backed prediction or decision process can be explained to the users or ben-

eficiaries of the program. Explanation of why specific decisions were made in individual cases is important when decisions could have significant impacts on people.

Consider the use of ML by a bank to determine if a person should receive a loan. If the bank does not understand how the model is arriving at its decisions, it may inadvertently or even unknowingly serve only the wealthiest customers in the target group, who might receive the highest ratings through the algorithm, despite the organization's intention to assist low-income individuals. For each individual, explainable solutions would provide information on the factors that were considered, why the person was denied a loan, and what that person can do to attain creditworthiness.

Auditability: *Can the model's decision-making processes and recommendations be queried by external actors?*

It is important that ML outputs can be audited externally to show that the model is fair and unbiased and does not introduce new problems for the target population.

In the microloan example, external organizations need to be able to understand ML decisions to ensure fair lending practices. For a medical diagnostic application, the ability to monitor the ML decisions could be a matter of life and death. Consider, for example, if a new epidemic arose that could complicate diagnoses, or if a segment of the population was discovered to present a certain disease differently from previously recorded cases. The ability to query the system and look for missed diagnoses would be vital to protecting patients' health.

Ensuring auditability may require implementing additional organizational infrastructure, such as an institutional framework that requires audits and provides auditors with secure access to data and algorithms.

Accountability: *Are there mechanisms in place to ensure that someone will be responsible for responding to feedback and redressing harms, if necessary?*

An accountable setup ensures there are systems in place for monitoring the use of ML to prevent harmful errors and ensures that specific people are responsible for addressing problems and correcting errors. It is important to make sure that there are human actors engaging with the ML system who are ultimately accountable for its results.

For example, an algorithm might be used to assist in diagnosing medical conditions, but the final diagnosis should still be provided by a trained medical professional. However, there may be more complicated situations that arise; consider if there were a shortage of trained medical professionals during a deadly disease outbreak. Does the risk of misdiagnosis outweigh the risk of not treating people? Even if the decision was made to follow the ML recommendations without case-by-case oversight from a physician, someone must be responsible for that decision and for monitoring its outcomes.

Each of the principles connect to the guiding concept of fairness, whether by ensuring the solutions tackle the right problem, reflect the right people, provide value, or can be understood, monitored, and backed up by responsible human actors. The next chapter examines how concerns around fairness arise at each step in the development of an ML project.

ADDITIONAL PROCUREMENT CONSIDERATIONS

This document focuses on development of machine learning applications. Procurement of existing, pre-packaged machine learning applications may require additional considerations. For example:

RELEVANCE

Is the use of ML in this context solving an appropriate problem?

As the use of ML becomes more widespread, organizations may seek to apply it to their work to distinguish themselves from competitors or increase their appeal to funders. This desire to keep up with technology may influence organizations to implement pre-packaged ML solutions without a clear understanding of whether they are using the right tool to solve the problem.

VALUE

Does ML produce predictions that are more accurate, timely, and actionable than alternative methods? Can the insights be implemented? Is the cost justified?

This principle is focused on the benefit of ML approaches compared to other options. Do the predicted values inform human decisions in a meaningful way? Does the machine learning model produce predictions that are more accurate or efficient than alternative methods? Does it explain variation more completely than alternative models?

For further reading, visit the OECD's guidance on Going Digital.[6]

6. OECD. *Going Digital Toolkit*. (Washington: OECD, 2019) https://goingdigital.oecd.org/en/

Chapter 3: Fairness Considerations in Developing an ML Project

In this chapter, the basic steps of an ML project are briefly explored with respect to their potential for introducing or propagating unfairness. In addition to fairness, other supporting principles discussed in the previous chapter are relevant to certain steps in the process; this chapter aims to highlight where these considerations emerge as a project develops. A hypothetical example of a solar energy company providing microloans and collecting data about its customers is used to illustrate these bias considerations.

Basic Steps of an International Development ML Project

This section introduces the basic steps of an ML project and considers their application in international development and the fairness and bias considerations relevant to each step.

Figure 1 demonstrates how these steps map onto the ML modeling process presented in *Making AI Work*, which highlights three phases of the model development process: Review Data, Build Model, and Integrate Into Practice.

At several points along this process, it will likely be necessary to revisit earlier steps. Insights from model building may lead the team to collect additional data. Model evaluation may lead to the creation of additional models and/or further tuning. Deployment will involve new data collection, which should also lead to further model adjustments, and maintenance requires regular reassessment of the model.

It is also important to involve not only technology developers who understand ML but also international development professionals who understand the implementation context at each stage of ML development.

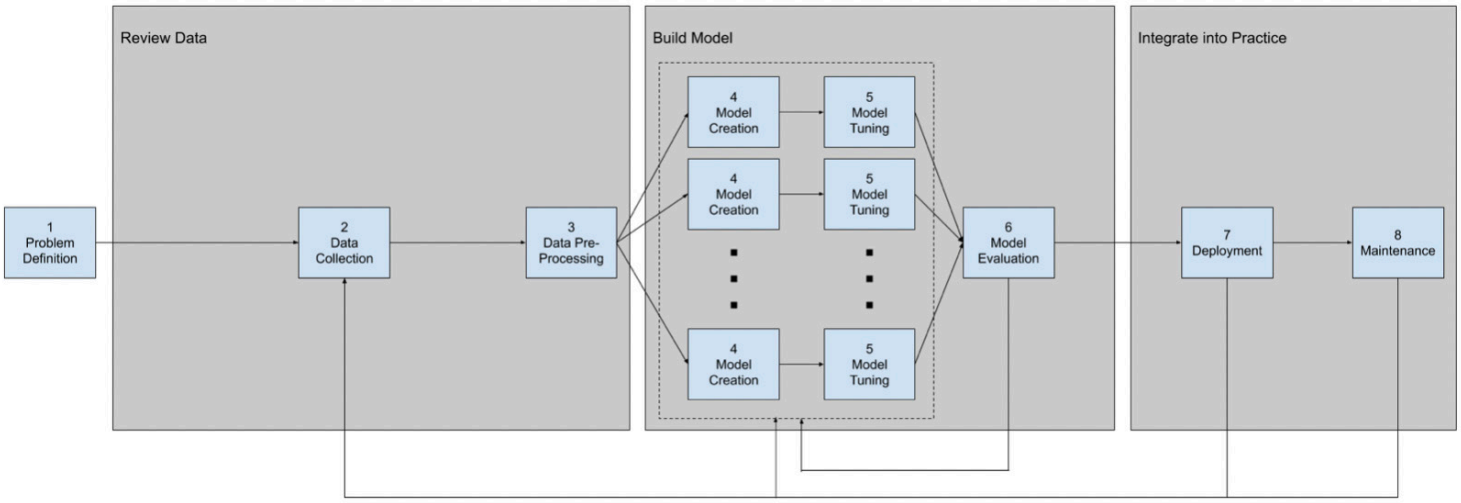


Figure 1 - 8 Basic Steps of an ML Project

STEPS OF AN ML PROJECT

1. Problem Definition
2. Data Collection
3. Data Pre-Processing
4. Model Creation
5. Model Tuning
6. Model Evaluation
7. Deployment
8. Maintenance

SOLAR CREDIT-SCORING EXAMPLE. To illustrate how bias can affect ML projects, this section uses a fictional case study of a company that has been providing household-level solar-powered products to consumers through a pay-as-you-go, rent-to-own model. The company loans the assets (solar panels and solar-powered devices) to the consumer and is paid back over a period of 1 to 3 years. Asset value, product usage, and repayment history are all tracked along with basic demographic information about the consumers and their households. The solar company is looking to expand its business by developing credit-worthiness algorithms that can a) help the company internally decide to whom they should give loans for higher-value solar devices and b) provide this consumer data to external financial institutions so these companies can also consider providing loans to the client base. We consider how each step might evolve for this example.

PROTECTED ATTRIBUTES

Traits that may not be used as a basis for decisions in a machine learning project are features such as gender, age, race, religion, or socio-economic status that the organization does not wish to allow the algorithm to use as the basis for decisions. While in some settings these protected attributes will be prescribed by non-discrimination laws or international treaties, in many international development contexts such regulations are not in place or are not easily implemented. In these situations, it is often up to the organizations making use of machine learning to determine what they will consider to be protected attributes.

1. Problem Definition

Every ML project should begin with a problem-definition phase, wherein the objectives are defined (with the understanding that they may evolve as the project progresses). This step requires gathering and analyzing input from project sponsors and other stakeholders — for example, detailed conversations with in-country staff to understand contextual relevance of the problem and proposed objectives in addition to determining what data are needed to achieve the objectives.

Fairness considerations: Biases on the part of the people defining the problem, project sponsors, and other stakeholders can all be introduced at this stage. For example, how do these various stakeholders address questions regarding representativeness? Who is missing from the data? Are the people defining the problem familiar with the context or are appropriate experts engaged?

SOLAR CREDIT-SCORING EXAMPLE: Assume that the organization defines the problem as determining the creditworthiness of individuals who are using solar lighting systems and appliances. Their objectives are to a) expand the size and scope of solar systems being provided and b) find other lending organizations that want to use their customer data. At this stage, it is important to test their assumptions about who would benefit from additional solar systems — for example, should individuals in off-grid areas be targeted over those who have access to other types of energy and are only supplementing with solar? With respect to other lending organizations, is repayment of loans for solar products an accurate predictor of creditworthiness for other types of loans? For both objectives, has any bias been introduced in the way that the initial round of loans was administered?

2. Data Collection

This stage involves pulling together data that is collected by the organization or acquired from external sources. This may include conducting a study to collect field data, purchasing or obtaining existing data sets, or using custom software to cull data from information

published online, such as social media. It is important for the organizations to collect appropriate data about all the different characteristics that may be associated with the different outcomes of interest, including protected attributes, to ensure that models are not biased.

Fairness considerations: The data collection stage has the potential to introduce a number of systematic biases. First, the person making decisions about which types of data to collect can introduce their own biases, such as beliefs about how the target population uses social media or how they will interpret and respond to survey questions. Second, the data collection design can result in data that are not representative of the target population. Data collection may systematically exclude people who are not well-sampled through the specified data collection format, such as mobile phone records, which exclude those without phones, or surveys conducted only in English. Information about people may also obscured by some data collection strategies. For example, if a mobile phone SIM is shared among several individuals, data drawn from the SIM may be attributed only to the registered user. Third, external datasets may not be appropriate for use in the intended setting. For example, many well-labeled, publicly available datasets used in modeling come from US and Europe image databases, and may not translate well to low-resource settings. Additionally, it may not always be clear how external data were collected and whether data collection biases were introduced by whoever collected those data. Finally, it is important to collect protected attribute data to be able to build machine learning for fairness. If these data are not collected, it is difficult to assess whether outcomes are fair (see the section in Chapter 4 on the pitfalls of "fairness through unawareness").

SOLAR CREDIT-SCORING EXAMPLE: Assume the company decides to collect data on asset usage and repayment history, as well as characteristics about its customers. Depending on how usage is calculated, it may reflect differences in weather and energy needs of the household rather than the customer's actual reliance on the solar-powered system. Usage of the solar device may be shared among multiple people, yet data used in ML models may attribute that usage to a single customer. How repayment history is measured could potentially introduce bias against people with irregular earning cycles, such as agricultural workers whose income can vary seasonally. Data on protected attributes should also be collected. For example, if the organization is dedicated to avoiding gender bias in its determination of credit-worthiness, it should collect data on the gender of its customers so that it can verify that its algorithms are unbiased in later steps.

3. Data Pre-Processing

This step consists of data cleaning and labeling, including extraction and transfer of useful information in a format suitable for ML. Data cleaning refers to the identification and correction (or removal) of corrupt, inaccurate, or irrelevant entries in a database. In supervised ML, labeling refers to the process of assigning tags to data that indicate the quantity the user is trying to predict. For example, when using ML to process and interpret images, labeling could involve tagging every image in a database with a text description.

Fairness considerations: Data pre-processing can propagate biases from the data collection stage or introduce new biases based on the labeling. If data from specific sub-

groups is harder (and more expensive) to label or check for accuracy, organizations may exclude that data, introducing biases. If labeling requires subjective input, individual biases of the data processing team can be incorporated. Furthermore, the labeler's worldview may shape how data is labeled - for example, when looking at labeling satellite maps to label buildings, someone who is only familiar with urban, developed setting may look for angular, glass-and-metal objects and miss the buildings in rural settings that may look different, such as having a round shape or grass roof.

SOLAR CREDIT-SCORING EXAMPLE: Assume that data for loan repayments and client information is collected differently based on the tools used by organizational staff: data can be collected electronically or in handwritten records in either English or local languages. Electronic data is easy to process and build a model around, whereas transcription of handwritten records presents a challenge and has the potential to introduce errors. Furthermore, translation of information from a local language to English may present a further barrier in terms of cost and accuracy; labeling decisions could reflect inconsistencies in language or difficulties in translation. The company may at this point choose to use only the electronic data. If the collection methods are uniformly distributed across the population, this may be a reasonable solution. However, in the more likely case that the data collection distribution is non-uniform — perhaps electronic data are only gathered from wealthier areas with computer access — using only electronic data will result in a non-representative sample.

OVERFITTING

A modeling error in which a model is too closely fitted to training data. As a result, the model "learns" idiosyncratic features of the training data and cannot generalize well to new data. Overfitting is more likely to be a problem when data are limited in either volume or diversity, a common problem in international development contexts.

4. Model Creation

This step is the core of the technical process and involves selecting and developing potential models using data from Step 3. This step begins with expression of the problem in a form amenable to ML techniques. Informed by the problem definition (Step 1), the analyst can choose a suitable ML algorithm. This step involves the separation of the dataset into a training set for training the candidate ML models and validation sets and test sets for comparing model performance. At this stage, avoiding overfitting should be a consideration, especially with small data sets. Usually, more than one algorithm is examined to see which approach best suits the problem. Chapter 4 will explore this step in greater detail and present a Fairness Methodology to guide criteria for algorithmic fairness.

Fairness considerations: Analysts should be familiar with how different algorithms work to ensure that the team does not implement an algorithm that can propagate known biases in the data. It is, therefore, important to consider the problem definition and the types of data collected before determining the type of algorithm to use. As every algorithm has limitations, it is possible to compound bias in the data by using an algorithm not well suited for the data. For example, using a Naive-Bayes approach, which assumes that predictors are independent of each other, might be incorrect if the predictors are not actually independent, (for instance, would attending college be independent from future earnings potential?).[7]

7. The Appendix discusses limitations for different types of algorithms in greater detail. For the interested reader, the following textbooks are recommended: Stephen Marsland. *Machine Learning: An Algorithmic Perspective* (New York: Chapman and Hall/CRC, 2014) and Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning*. (New York: Springer, 2013).

SOLAR CREDIT-SCORING EXAMPLE: Consider one potential approach to determining whether an individual should receive a loan for additional solar equipment: a k-nearest neighbors (k-NN) classifier. A k-nearest neighbor approach would look at a given number of people (k) in the dataset who are most similar to the individual in question and see whether those people were given a loan or not. If the majority were given a loan, then the algorithm would classify the individual as someone who should get a loan. Even if gender is not explicitly included, this type of algorithm could consider a woman applicant to be nearest to other women applicants, based on features that correlate with gender. If historical data are gender-biased such that the organization provided loaned assets at a higher rate to men than to women, female applicants may be placed closer to unsuccessful applicants from the past, leading to a gender-biased determination that the applicant is not creditworthy. In contrast, a logistic regression model would generate coefficients corresponding to features such as gender or income. The higher the coefficient, the stronger the effect of that feature on the outcome. While both models can be biased, it may be easier to identify bias in a logistic regression model because the features used in determining outputs are more explicitly presented. Chapter 4 discusses some of the ways to achieve this fairness.

NAVIGATING ALGORITHM CHOICE

Descriptions of specific algorithms and the bias considerations associated with each are presented in the Appendix.

5. Model tuning

Often, ML models will contain different elements such as threshold values and hyperparameters that control the learning process itself. For example, in the k-nearest neighbor approach described in Step 4, changing the value of k (a hyperparameter indicating the number of "nearest neighbors" to group together) will change how the machine learns from the data. Successful ML requires selecting appropriate threshold and hyperparameter values. Therefore, the previous steps of training and evaluation may need to be repeated many times to identify values that improve performance of the chosen algorithms on the chosen data set.

HYPER-PARAMETERS

Values that control how an algorithm learns. While an algorithm's parameters select which features the algorithm should learn about, hyperparameters provide input about how that information is incorporated in the learning process.

Fairness considerations: Tuning the hyperparameters in the model changes the underlying model. Hyperparameter tuning is done to improve performance, and specific performance metrics (accuracy, statistical parity between groups, etc.) need to be defined by the ML implementer. Additionally, the selection of thresholds for decision-making based on model results are also determined at this step. Many problems will also have tradeoffs between fairness and performance or between how fairness is implemented for different groups. The selection of performance metrics and thresholds, and decisions about how an analyst deals with tradeoffs, can reflect individual or organizational biases.

SOLAR CREDIT-SCORING EXAMPLE: Assume that the company decides to implement a logistic regression model. When used in classification, the analyst must determine a threshold value at which to approve or reject an individual for a loan. The company's data indicates that men and women have different loan default rates, which the analyst interprets as different likelihoods of repaying a new loan. The goal is to correct for this bias to make the loan decision fairer. The analyst's choice of threshold values can either minimize errors (misclassifications of creditworthiness) in women, minimize errors in men, or equalize errors in both groups (as shown in Figure 2). Equalizing errors for both groups in this case will increase false positive errors for one group while increasing false negative errors for the other.

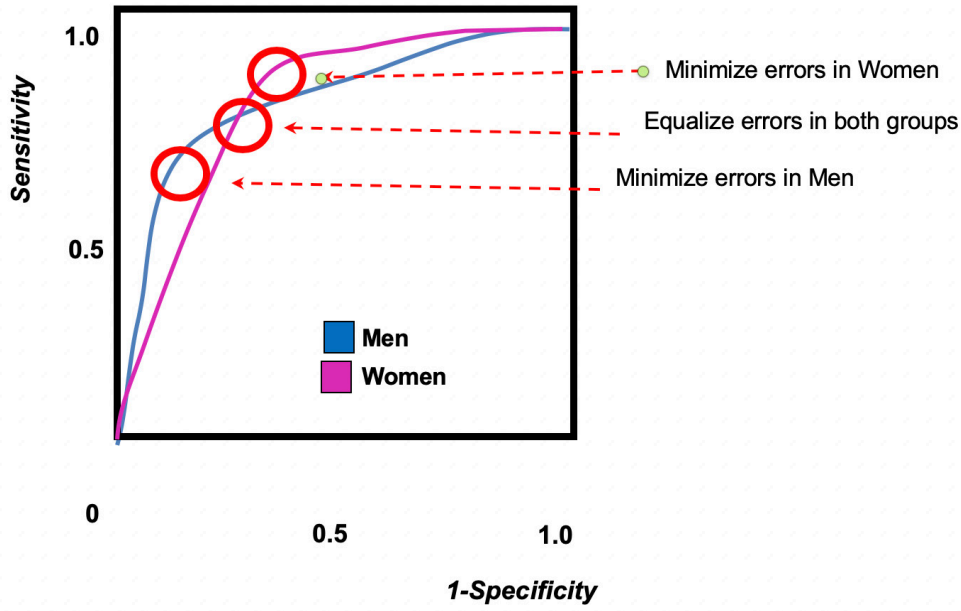


Figure 2 - Sample ROC curve for multiple groups, showing different possible operating points for the algorithm

6. Model Evaluation

A key step in every ML project is model evaluation. Models are evaluated against predetermined metrics — such as accuracy, performance, and speed — to decide which approach is most suitable. At this point, if models do not meet criteria for deployment, the implementer may revisit the model creation step or even an earlier step.

Fairness considerations: The choice of criteria to compare between models should reflect the values of the organization and should be well documented to promote external auditing or explainability. For example, different models may emerge as preferable depending on whether a greater priority is preventing false negatives or false positives. The choice of training data might also need to be revisited to avoid replicating biases reflected in the data.

SOLAR CREDIT-SCORING EXAMPLE: Suppose the data shows a link between default rates and gender. A strict evaluation criterion for the highest accuracy approach would introduce bias into the system because of this gender difference (discussed in more detail in Step 5). In order to implement fairness methods (discussed further in Chapter 4), the organization may choose to sacrifice accuracy in favor of creating fairer rules for evaluating male and female candidates. Rather than using a strictly representative sample, the company could consider creating a training data set with equal numbers of default rates among men and women.

7. Deployment

During this step, the ML algorithms are deployed in the field — often first in beta (small-scale, controlled, and manually audited), then increasingly scaled up after the model is verified to be working correctly It is important to note that deploying the model to a new region/ context should also be beta-tested and scaled slowly. In the context of international development, this step often involves integrating the ML system into a decision-making process.

Fairness considerations: The integration of a model into the real world is a complex process that presents many ethical challenges. The appropriate use of the model, including its limitations and processes for accountability, must be clearly communicated to users. This communication includes technical information about the model (the model approach, accuracy, errors, and tradeoffs made in the design) as well as where, how, and for whom the model should be used based on the representativeness of the training data.

SOLAR CREDIT-SCORING EXAMPLE: Assume that the organization was using a credit-worthiness model to determine which of its clients should be offered loans for purchasing a solar-powered television. The model made use of information about who was most likely to purchase and use a solar-powered television, which determined that specific individuals living in off-grid communities were good candidates. For logistic simplicity, the organization may choose to distribute solar-powered televisions to regions that have more of these individuals, which unfairly impacts qualified individuals living in other regions.

8. Maintenance

As machine learning models are deployed and additional data is collected, the team will continue to learn more about the problem, including potential shortcomings. The maintenance team is responsible for keeping the model updated with new data and retraining the model as needed. Newer and better algorithms may also become available and can be incorporated as the project evolves.

Fairness considerations: Even if the training dataset is highly representative of the target population when it was collected, as time goes by, the models can become less accurate given changing circumstances for the population. As a result, it is important to regularly audit the models to check for and rectify any biases and unintended negative consequences introduced by changes on the ground. Lack of bias in the model evaluation and interpretation stage does not guarantee lack of bias at scale.

SOLAR CREDIT-SCORING EXAMPLE: Imagine that a country in which the company is operating imposes a regressive kerosene tax that adversely affects the rural population of solar asset users. While the population adjusts to the tax, the model may reject loans for those affected because their available disposable income has gone down. This outcome would present at just the wrong time, given that access to solar energy could reduce dependence on kerosene and improve applicants' financial outlook. Adjusting the model to account for this change would create fairer outcomes.

Chapter 4: Building for Fairness

This chapter explores the application of the principles for responsible ML development discussed in Chapter 2 to the steps of ML project development outlined in Chapter 3. It considers practical approaches to problem definition (Step 1), mitigating bias in data sets (Steps 2 and 3), model creation, tuning, and evaluation (Steps 4, 5, and 6), and ML deployment and maintenance (Steps 7 and 8).

A. Asking the Right Question for ML

PROXY

Value that is measured as a substitute for the real quantity of interest. Proxies may be used to make predictions, or as a direct stand-in for things that are hard to quantify (e.g., potential or risk).

This section focuses on key aspects of Step 1, Problem Definition. When any ML model is used in decision making, it is important to consider first whether the right question is being asked and secondly whether ultimately a machine should be making this determination.

Too often, errors occur because the framing of the problem is incorrect. Programmers must consider what it is exactly that they wish to know and which problem they want to solve. Although ML may not be capable of answering a question directly, some proxies are better than others for delivering fair and useful outcomes. For example, if the goal is to help unemployed people find jobs, the ideal but impractical question may be, "Who is best suited for the open positions?" It is difficult to measure suitability directly for candidates who don't yet hold those jobs as their performance in them can't be known. To approximate suitability in a way that prioritizes fairness, it may be important to ask, "Who is likely to succeed in the open positions?" rather than "Who has held positions like these before?" The latter question may reproduce existing inequities, whereas the former question, depending on how "likelihood" is defined, could identify a broader range of capable individuals.

Before proceeding, it's also important to ascertain whether it's appropriate for a computer to answer the question. Although neither computers nor humans are perfect (and, after all, humans write the computer programs), certain moral and ethical questions in development projects may have consequential impacts on people's lives, leading programmers to consider at the outset if they really want a computer conducting the analysis. In an emerging refugee crisis, for instance, determining who can access different types of resources and services may demand the compassionate wisdom of experienced leaders rather than computer algorithms.

B. Mitigating Bias in Data Sets

This section explores Steps 2 and 3: Data Collection and Data Pre-Processing, to introduce methods of addressing bias in data sets. Examples of bias in data sets include under-sampling for racial, cultural, and gender diversity in image recognition, such as categorizing wedding photos only when the bride is wearing clothes of a specific color in accordance with cultural norms.[8] The issue of image datasets underrepresenting certain ethnicities is also known in facial recognition, where classification accuracy suffers when images of underrepresented minority individuals are analyzed.[9] In a third example, voice recognition

8. Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, JimboWilson, D. Sculley. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World." In *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017.

9. Larry Hardesty, "Study finds gender and skin-type bias in commercial artificial-intelligence," *MIT News Office*, February 11, 2018, systemshttp://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

systems are well known to perform more poorly for non-native English speakers than native speakers,[10] which results in incorrect answers to questions posed to popular voice-based assistant systems.

When bias arises in a data set, methods for addressing this include addressing the sampling of the data, cleaning the data and labels, or adding, removing, diversifying, or redistributing features.

Table 1 provides further resources for the more technical approaches discussed in this section.

Table 1 - Further Reading for Addressing Bias in Datasets

APPROACH	RECOMMENDED RESOURCES
Data Augmentation	Fedor Kitashov, Elizaveta Svitanko, Debojyoti Dutta. "Foreign English Accent Adjustment by Learning Phonetic Patterns." ArXiv (2018): 1-5. https://arxiv.org/abs/1807.03625
Feature-level Reweighting	Stefano M. Iacus, Gary King, and Giuseppe Porro. "Causal Inference Without Balance Checking: Coarsened Exact Matching". Political Analysis 20, no 1 (2012): 1-24, http://j.mp/2nRpUHQ
Resampling through Randomization of the Minority Class	B. Efron. "Bootstrap Methods: Another Look at the Jackknife." The Annals of Statistics 7, no. 1 (1979): 1-26. www.jstor.org/stable/2958830. Gerdie Everaert and Lorenzo Pozzi. "Bootstrap-based Bias Correction for Dynamic Panels." Journal of Economic Dynamics & Control 31, no 4 (2007): 1160–1184.
SMOTE: Synthetic Minority Over-sampling	Nitesh V. Chawla, Kelvin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-Sampling Technique". Journal of Artificial Intelligence Research no 16 (2002): 321-357. https://doi.org/:10.1613/jair.953
Adversarial Learning Approaches	Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning".In proceedings of AAAI/ACM Conference on AI, Ethics, and Society (New York: Association for Computing Machinery, 2018): 335-340. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In proceedings of 27th International Conference on Neural Information Processing Systems 27, Montreal, December 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture For Generative Adversarial Networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR, 2019): 4401-4410

10. The Economist, "In the world of voice-recognition, not all accents are equal," *The Economist*, February 15, 2018, https://www.economist.com/books-and-arts/2018/02/15/in-the-world-of-voice-recognition-not-all-accents-are-equal

Gathering more diverse data is the preferred option when cost and time permit. Having a more diverse data set almost always helps ML algorithms make decisions more accurately and more fairly. The outcomes observed in the examples above are essentially a side effect of what statisticians would call "sampling bias" – when a project attempts to model a phenomenon for an entire population but accidentally selects a sample not representative of that population. For example, trying to create a model that would respond to all English speakers but obtaining training samples only from British and American speakers will result in poor performance for speakers from other places. However, it is possible that even with a representative sample an algorithm may under perform on minority groups. The solution would be to oversample minority groups, which would then increase accuracy of the algorithm for those minority groups.

Data augmentation[11] refers to a family of techniques that increase the size and diversity of the training data without actually collecting more data. The effectiveness of data augmentation has been demonstrated in image recognition tasks through techniques like cropping, rotating, flipping, and darkening input images. For example, one technique for data augmentation in natural language processing is to generate training sets with accented versions of the words available in the base dialect of the training set.

Feature-level reweighting[12] describes a family of approaches in which features are assigned weights (multiplied by scalar values) to make the data more representative. The weights are usually adjusted so that the classifier algorithm meets some criteria of fairness. A common approach is to repair data by more heavily weighting selected features of the sensitive group to equalize misclassification error rates. This is a standard technique in statistics and is often useful for other purposes, including matching for causality. In the natural language processing example, one could reweight features correlated with protected attributes in order to deemphasize them.

Resampling through randomization of the minority class[13, 14] is a variant of the "bootstrap" technique from statistics, which is used for bias correction in estimators and can be used to boost the number of elements of the minority class by sampling more of that minority class through random sampling with replacement. This approach effectively increases the number of elements of the minority class as seen by the classifier. Of course, this solution is not as good as increasing the sample size by collecting more data for the minority, as suggested in the first solution in this list (gathering more diverse data). In the natural language processing example, assuming that fewer samples were collected from non-native English speakers than from native English speakers, this approach would involve randomly resampling from the non-native English speaker samples to increase their representation in the data set.

SMOTE (Synthetic Minority Over-sampling Technique)[15] is a generalization of the resampling approach described above that further includes a k-nearest-neighbor approach in order to generate "synthetic" minority class members. This approach goes beyond a simple

11. Fedor Kitashov, Elizaveta Svitanko, Debojyoti Dutta. "Foreign English Accent Adjustment by Learning Phonetic Patterns." *ArXiv* (2018): 1-5. https://arxiv.org/abs/1807.03625

12. Stefano M. Iacus, Gary King, and Giuseppe Porro. "Causal Inference Without Balance Checking: Coarsened Exact Matching". *Political Analysis* 20, no 1 (2012): 1-24, http://j.mp/2nRpUHQ

13. B. Efron. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7, no. 1 (1979): 1-26. www.jstor.org/stable/2958830

14. Gerdie Everaert and Lorenzo Pozzi. "Bootstrap-based Bias Correction for Dynamic Panels." J*ournal of Economic Dynamics & Control* 31, no 4 (2007): 1160–1184.

15. Nitesh V. Chawla, Kelvin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-Sampling Technique". *Journal of Artificial Intelligence Research* no 16 (2002): 321-357. https://doi.org/:10.1613/jair.953

bootstrap: instead, the nearest neighbor of the resampled element is computed and this information is used to generate a new element using a random value within the range given by the difference between the resampled element and its nearest neighbor. Generating this new element effectively creates a new "synthetic" member of the minority class that provides more variation for the classifier by supplying a new data point.[16, 17, 18]

Adversarial learning[19] approaches address biases in prediction due to protected attributes by applying adversarial learning. In this approach, there are two machine learners – one predicting the output, and the other predicting the protected attribute — in order to converge on a model that predicts the correct outcome independent of the protected attribute. Adversarial models have been popular in image classification. These emerging techniques are not entirely generalizable just yet, but provide a good opportunity for future research. If this approach were to be applied to the natural language processing example, it would involve one ML program predicting the words spoken, and another ML program predicting the status of the speaker (native or non-native English) based on the output of the language processing program. If the status of the speaker could be predicted based on the output of the language processor, it would indicate bias in prediction.

ALGORITHMIC FAIRNESS
Design of algorithms to achieve fair outcomes.

The next section delves into different criteria for evaluating fairness in an algorithmic model and how these criteria can be selected and applied in machine learning.

PROBABILITY
Likelihood of an event occurring.

C. Model Creation, Tuning, and Evaluation for Algorithmic Fairness

This section focuses on fairness in Steps 4, 5, and 6: Model Creation, Model Tuning, and Model Evaluation. This statistical perspective on fairness is the predominant focus of this section. Therefore, some background on statistical reasoning is explained in the first section to provide conceptual context for the reader.

Background on Probability and Conditioning

This section introduces the basic concepts of probability and conditioning on which the discussion of algorithmic fairness relies. Readers familiar with these concepts may be comfortable skipping to the next section; however, this section illustrates the application of these concepts to decision-making in a way that may be useful in the selection of algorithms discussed in the appendix.

CONDITIONING
Specifying certain conditions, often as a given set of data, under which probability will be evaluated.

Definitions of algorithmic fairness and criteria for evaluation of algorithmic fairness are often stated in terms of probabilities in general and conditional probabilities in particular.

A probability can be viewed as a measure of the likelihood of an event. The probability of a six on a roll of a die is 1/6 because there are six sides and they are all equally likely to appear. A probability can also be conceived as a frequency of an event within a sufficiently large

16. Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." In International Conference on Intelligent Computing, August 2005, edited by D.S. Huang, X.-P. Zhang, G.-B. Huang, 3315-3323, New York: Springer. https://sci2s.ugr.es/keel/keel-dataset/pdfs/2005-Han-LNCS.pdf

17. Tomasz Maciejewski and Jerzy Stefanowski. "Local neighbourhood extension of SMOTE for mining imbalanced data." In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (New York: IEEE, 2011): 104-111. https://doi.org/10.1109/CIDM.2011.5949434

18. Chumphol Bunkhumpornpat, Krung Sinapiromsaran, Chidchanok Lursinsap. "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." In Pacific-Asia conference on knowledge discovery and data mining, 2009, edited by T. Theeramunkong, B. Kijsirikul, N. Cercone, T. Ho, 475-482. New York: Springer. https://doi.org/10.1007/978-3-642-01307-2_43

19. Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning".In proceedings of *AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery, 2018): 335-340. https://doi.org/10.1145/3278721.3278779

population or sequence of similar opportunities. In this frequency conception, the probability of a six on a roll of a die is 1/6 because, in the past, very nearly one out of six rolls were observed to be six whenever a large enough sample was evaluated. In this section, the frequency conception will be emphasized for its relevance to the example scenario.

The concept of conditioning relates to the connection between probabilities and data. A conditional probability is the probability of an event occurring under a specified condition, i.e. given some relevant data.

For example, assume a 40-year-old woman, Beth, is considering whether to get a mammogram. She learns that 1% of women aged 40 to 50 have breast cancer. The mammography procedure she is considering has a 9.6% false positive rate and an 80% true positive rate. Beth decides to get a mammogram screening and receives a positive mammogram result.

Statistical procedures tell us that, prior to the test, Beth had a 1% probability of having breast cancer. The conditioned probability that she has breast cancer, given the data of her positive test result, reveal that Beth has a 7.8% probability of having breast cancer. Figures 3 and 4 and the step-by-step process below explain why.

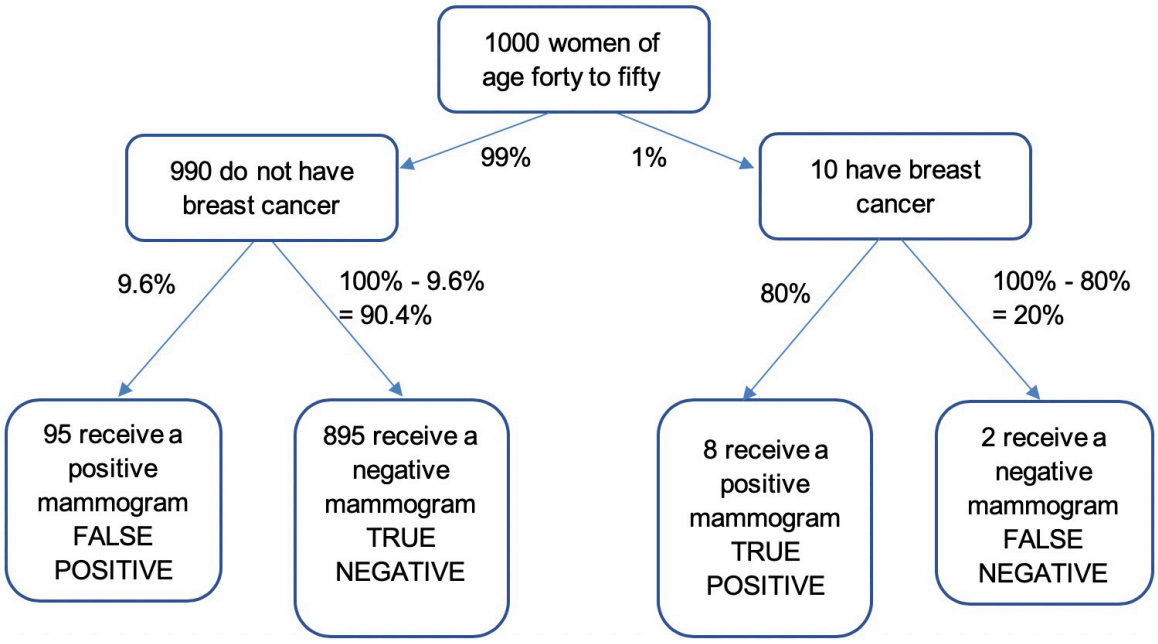


Figure 3 - Out of a population of 1000 women, 990 women can be expected to be cancer-free and yet 95 of them will screen positive (false positive). Among the 10 women who do have cancer 8 will screen positive (true positive).

Assume a population of 1000 women aged 40 to 50. On average, 1% of women in that age range will have breast cancer. This means that the 1000 women can be partitioned into two sets: the 10 who have the disease and the vast majority (990 women) who do not. If there are only two groups, and everyone must fall into one group or the other, this process can be described as a collectively exhaustive partitioning of the set. Collectively exhaustive partitioning can be applied in this case because the example makes a binary distinction between people who have cancer and those who do not.

This example requires one more layer of partitioning. Because the mammography test result will be relevant to the case, the next layer of events is based on the two possible test results: positive and negative.

For the women who do not have breast cancer, there are two possibilities: a false positive or a true negative result. Because the population of women who do not have breast cancer is large and the false positive rate is substantial (9.6%), there is a large number of false positives in our population (95 women). For the women who do have breast cancer, there are two possibilities: a true positive or a false negative result. Because the test has a good power (ability) to detect the disease (80%), there is also a considerable number of true positives in our population (8 out of the 10 women with breast cancer).

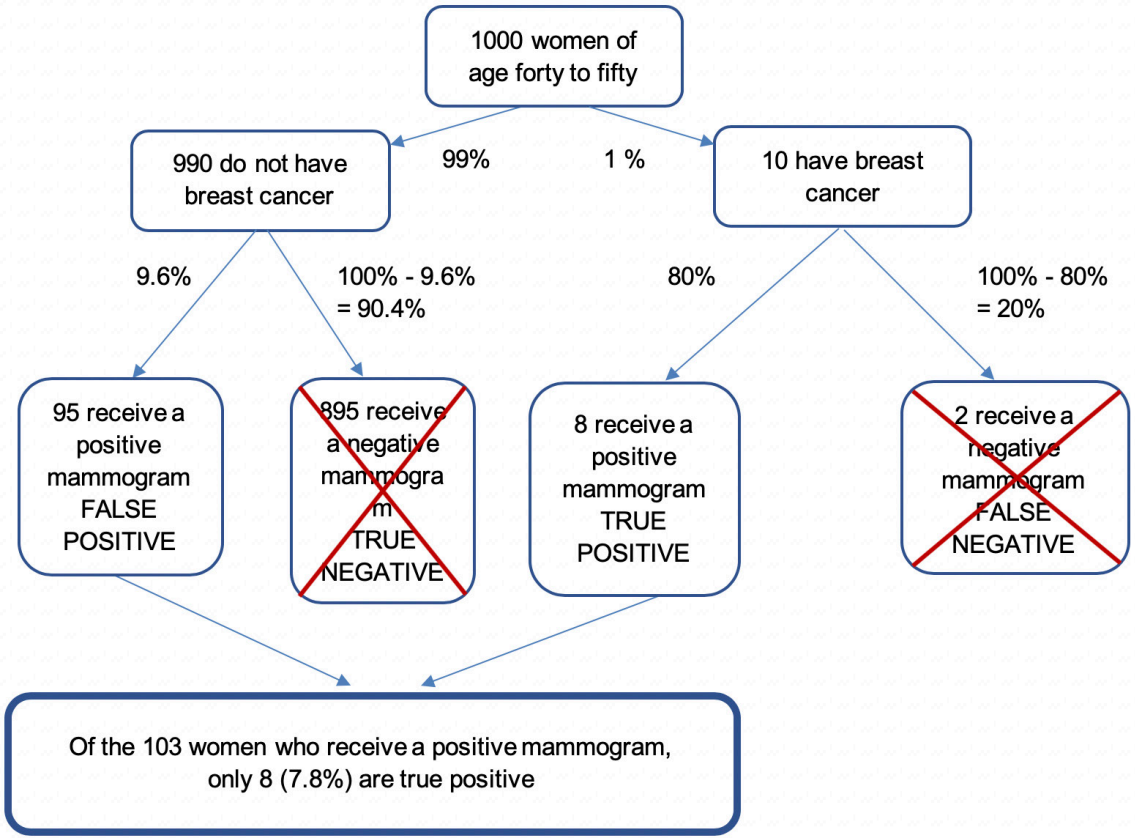


Figure 4 - A total of 103 women received a positive result from their cancer screenings. Only 8 of these women actually had cancer. The probability of having breast cancer given a positive result is therefore 8/103 = 7.8%.

Beth's mammography test result enables us to condition the events in our example. We don't know for sure whether Beth has cancer or not, but we do know that she received a positive test result, so we can rule out the false negative and the true negative categories. The only remaining events after conditioning are the 95 women in the population with a false positive and the 8 women with the true positive results. Given those proportions, we can infer that the probability that Beth has breast cancer is 8 out of (95+8), which is 7.8%.

As this example illustrates, conditioning a probability based on data is a process of elimination followed by assessment of proportions.

To compute a conditional probability accurately and reliably, the following three steps are recommended:

1. Enumerate all the events that were possible prior to having the new data. List every event in such a way that they are finest grained (all the layers of known possibilities are outlined), mutually exclusive (an event can have only one of the possible finest-grained outcomes), and collectively exhaustive (each event must achieve one of the possible outcomes). Place number labels on each event that accurately reflect the proportions as you understand them to occur in the general populations (aka the "base rates").
2. Take account of the new data that came to light. Eliminate all the events that are ruled out given the new data.
3. Recalculate the relevant frequency by inspecting the proportions among the remaining possibilities.

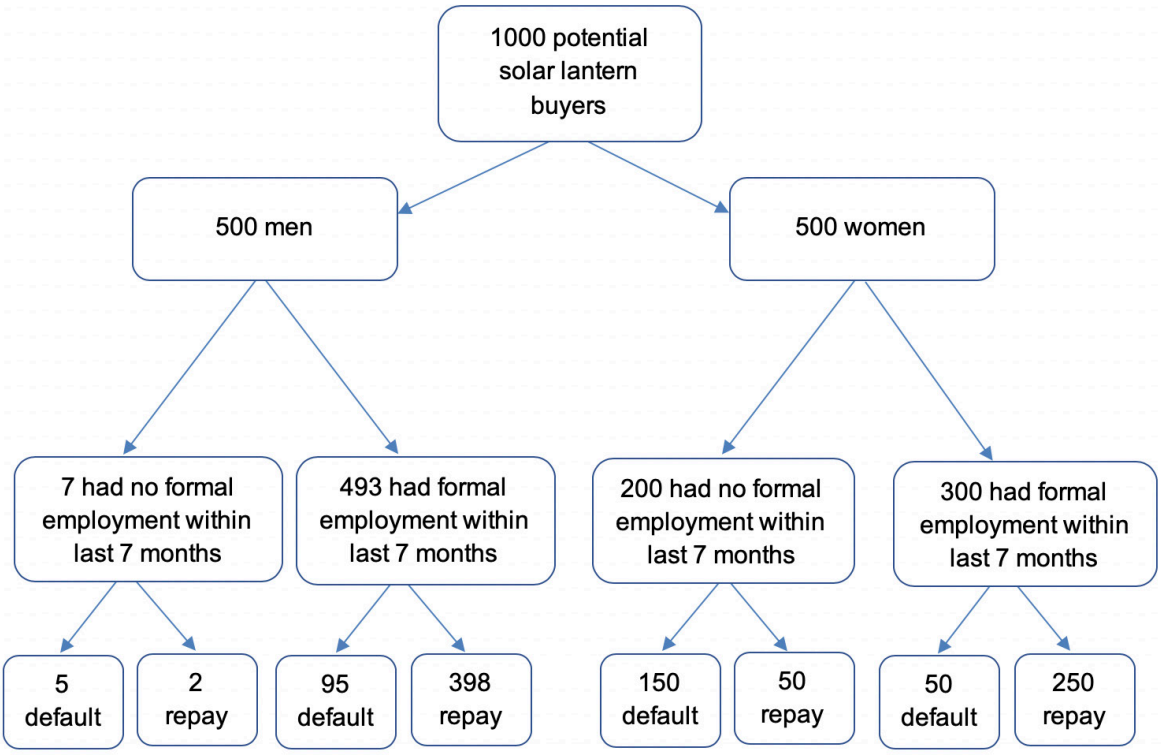


Figure 5 - In this notional example, 100 men defaulted and 200 women defaulted which could lead to a biased lending decision. However, women were far less likely to have held formal employment within the last 7 months. Women and men experiencing similar employment conditions defaulted at similar rates.

A Statistical View of Algorithmic Fairness

The previous section explains the thinking behind predictions based on conditional probability. Machine learning holds out the promise that these types of patterns in data can be discovered by computers and used for making predictions. However, without human oversight, machine learning may introduce or replicate biases that result in unfair outcomes.

Consider, for example, the challenge of making loans to people who want to buy solar lanterns. Assume that a company applied ML to a data set and trained an algorithm to predict who is going to default on their loan payments. Subsequently, the company finds that the algorithm appears to be treating women unfairly. Out of 500 men to whom the algorithm is applied, 100 were predicted to default. Out of 500 women to whom the algorithm is applied, 200 were predicted to default. Given the disproportionate outcome and potential unfair treatment between two groups, the company investigates the model further.

The company discovers that the machine learning algorithm was trained on data that had proportions as indicated in Figure 5. In this hypothetical case, the machine learning system produced predictions that exactly matched the proportions in the training set. In that sense, the ML system was accurate. If the patterns in future loans are similar to those in past loans, the ML algorithm might have a good record of predicting future defaults. Nevertheless, the company might still be deeply concerned about how unfairly women are being treated because they are rejected at twice the rate of men, potentially based on embedded cultural and systemic biases in historic data.

Upon further analysis of this data, the company recognizes that default rates were affected by the history of formal employment. Women were far less likely to have held formal employment within the past 7 months. Moreover, women's conditional default rates, given their employment status, were comparable to men's: Among applicants who lacked formal employment in the past 7 months, 71.4% of men defaulted compared to 75% of women; among applicants who had no gaps in employment, 19.3% of men defaulted compared to 16.7% of women.

Uncovering discrepancies in employment and making this link to conditional probabilities is a key take-away of this example. If the solar lantern company consistently uses the ML algorithm for lending decisions, there will be an unfair distribution of benefits. The conditional probability of rejecting a qualified loan application given the borrower is male is 20%. The conditional probability of rejecting a qualified loan application given the borrower is female is 40%.

However, if the company considers the unequal employment conditions that the applicants in the training data experienced, it finds that women were no more likely to default than men in equivalent circumstances. The next section explores ways to mitigate such problems.

Algorithmic Criteria for Fairness

This section reviews some of the most widely used, historically important, or key emerging algorithmic criteria for fairness. The discussion focuses on four core approaches known widely in the computer science literature: demographic parity, equalized opportunity, equalized odds, and counterfactual fairness. All of these approaches seek to improve the fairness of outcomes by establishing a criterion for fairness and modifying an algorithm to meet goals with respect to that criterion. However, enforcing a fairness criterion does come at the

cost of accuracy, which is to be expected as a fairness criterion is an additional constraint (see Zafar et al 2018 in the Further Reading list at the end of this chapter for examples of this trade-off).

It is important to acknowledge that fairness may not be 100% achievable in every ML application. While the goal may be to design a completely fair ML application, failure to do so should not be a reason to abandon ML entirely but rather a motivation to continue improving over time. Those employing ML in international development contexts should ensure frank and transparent discussions of what is possible within the existing constrains and what will be considered acceptable in terms of achieving fairness for a specific project at its various stages.

The advantages and disadvantages of each approach are summarized in Table 2 and discussed within each subsection. Although "fairness through unawareness" is presented first, this criterion is not recommended for ML applications. The discussion begins by explaining why this common approach introduces unacceptable risk in terms of bias.

Table 2 - Advantages and Disadvantages of Criteria for Algorithmic Fairness

CRITERION	ADVANTAGES	DISADVANTAGES
Fairness through unawareness	• Simple to implement	• Not effective unless some unusual criteria are satisfied (no correlated attributes)
Demographic parity	• Conceptually simple • Can have legal standing (disparate treatment)	• Does not address individual-level fairness • May unacceptably compromise prediction accuracy
Equalized opportunity	• Appeals to a reasonable interpretation of fairness • Can be a good option if the true positive rate is most consequential factor	• Disparate false negative rates may remain between two populations • Requires lots of labeled historical data
Equalized odds	• Appeals to a reasonable interpretation of fairness	• May not address group disparities sufficiently • Can be inconsistent with high levels of accuracy

ADDITIONAL FAIRNESS CRITERIA

This section does not seek to be comprehensive because there are so many options for fairness approaches available in the literature. Some excellent criteria have emerged in recent years, such as use of equality indices from economics.[20] Some of the best recent work is beyond the scope of this chapter; describing its advantages would require prerequisite explanations not covered here. For example, Spelcher et al. (2018) proposed a means of formulating objective functions for automated decision making based on formulations of cardinal social welfare in economics. All criteria have limitations; for example, the work of Heidari et al. does not address individual fairness. In this section we focus on the four core algorithmic fairness criteria commonly cited in the computer science community (demographic parity, equalized opportunity, equalized odds, counterfactual fairness) and rebut fairness through unawareness (which is an ineffective approach to fairness). Readers are encouraged to explore the "Further Reading" list on page 31 to learn more other criteria (for example, Verma and Rubin 2018).

Not Recommended: Fairness Through Unawareness

One approach for addressing fairness in ML is to remove protected attributes from the data set, commonly known as "fairness through unawareness." While this approach may seem intuitive, due to correlations in the data, it is possible — and in practice frequent — that removing the protected attributes actually increases unfairness. The primary underlying flaw of the fairness-through-unawareness approach in ML is that the protected attributes whose overt labels are removed from the data sets are often correlated with other features that remain in the training data.

Generally, the design of the machine learning solution should include clear steps to avoid discriminating based on protected attributes and algorithm implementation should include checks after the training of the algorithm is complete to ensure that biases from the training data were not picked up inadvertently. For example, researchers have shown bias in word embeddings, a natural language processing technique of growing use and popularity, as programs tended to pick up gender stereotypes embedded in words that did not explicitly specify gender, such as associating "computer programmer" with "man" and "homemaker" with "woman."[21, 22]

This result could be due to historic patterns in the data that correlate with past discrimination trends – for example, districts or neighborhoods correlating with ethnicity or socio-economic status – or due to "redundant encodings." In redundant encodings, other variables essentially encode the same information as the removed protected attribute. An example of such a redundant encoding could be belonging to groups or organizations that are gender specific.

---

20. Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. *A Unified Approach to Quantifying Algorithmic Fairness: Measuring Individual and Group Fairness via Inequality Indices*. (London: KDD, 2018)

21. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In proceedings of *27th International Conference on Neural Information Processing Systems* 27, Montreal, December 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets

22. Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture For Generative Adversarial Networks." In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR, 2019): 4401-4410

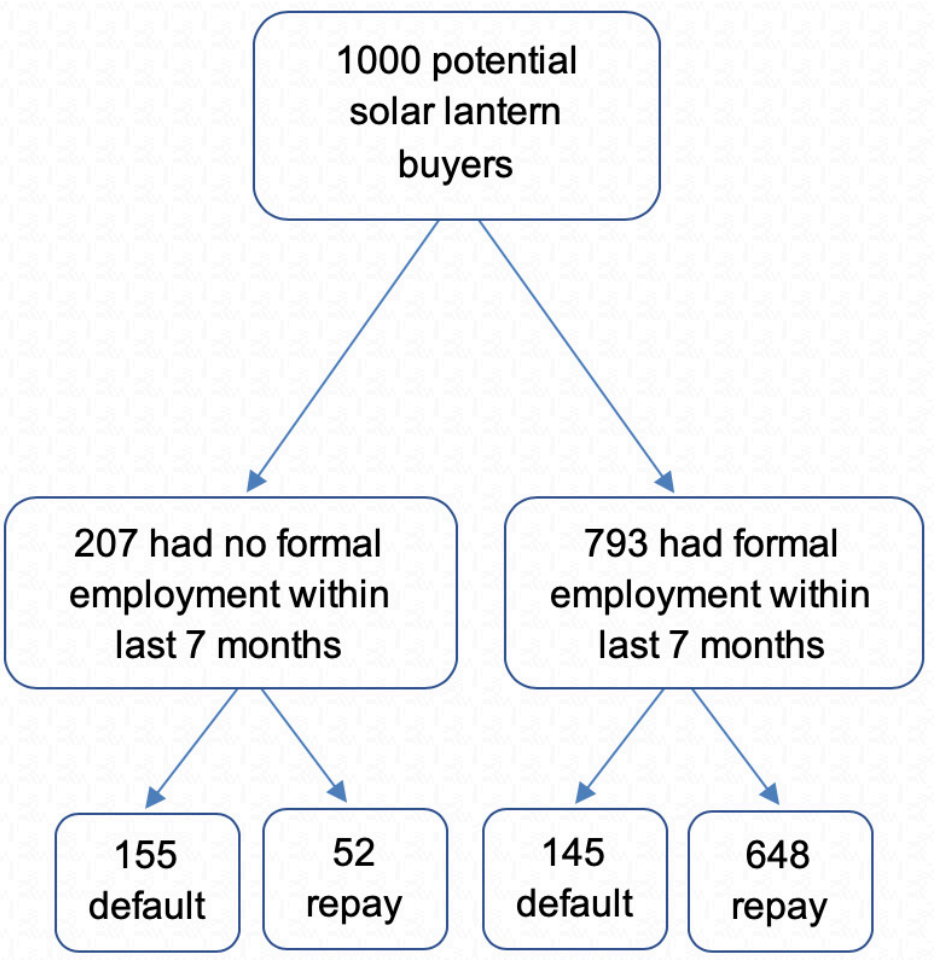


Figure 6 - Fairness through unawareness of gender applied to solar lantern training data. Here the label of "man" and "woman" has been removed. The default rates in the training data can be seen to be higher among applicants who lacked formal employment in the past 7 months.

To illustrate for the international development context, Figure 6 returns to the solar lantern case to describe what happens if "fairness through unawareness" of gender is pursued. The figure depicts the data on which the ML algorithm is trained if the protected label of gender is removed. It reflects the very same data set as in the previous section, but aggregates the numbers across the genders. From this data set, the machine learning algorithm should learn that conditional probability of defaulting given the borrower lacked formal employment is 155/207 or nearly 75%. The conditional probability of defaulting given the borrower had formal employment within the past 7 months is 145/793 or about 18%.

The problem, in this case, is that women tended to lack a record of formal employment in the recent past. Of the people who had no formal employment, 200/207 or about 97% were women. Conversely, of the people who had no gap in formal employment, 300/973 or about 30% were women.

If the solar lantern company consistently uses the "fairness through unawareness" ML algorithm for lending decisions, there will still be an unfair distribution of benefits. Figure 7 applies the known facts about gaps in employment to disaggregate rejected applicants by gender. The conditional probability of rejecting a loan application given the borrower is

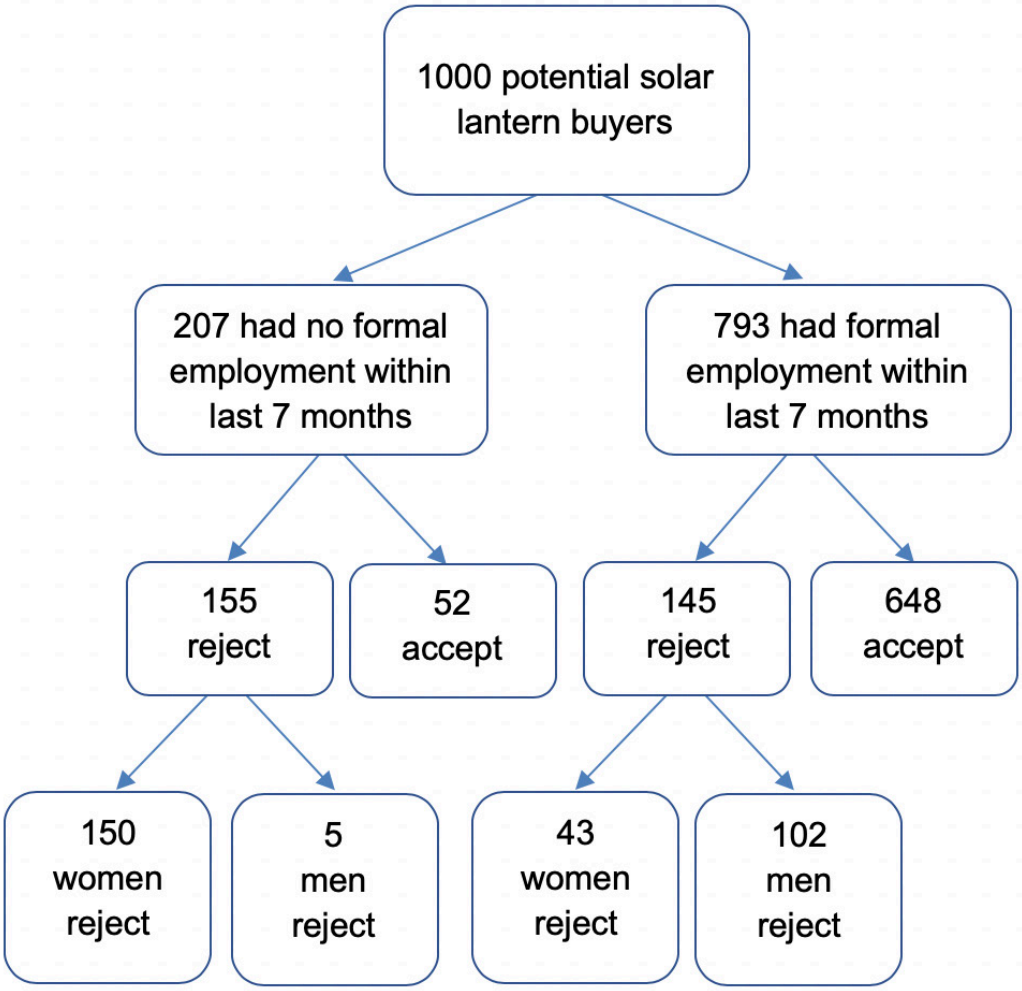


Figure 7 - Fairness through unawareness results for gender. Removing the gender label does not resolve the bias against women applicants because a key attribute in the data set — formal employment — is associated with gender.

male will be 107/500 or about 21%. The conditional probability of rejecting a loan application given the borrower is female is 193/500 or about 38%. The fairness through unawareness approach left the 2:1 disparity in rejection rates nearly unchanged.

Figure 7 illustrates how the fairness-through-unawareness approach reproduces the bias against women applicants, because a lack of formal employment in the last 7 months is closely associated with gender. Using this approach and rejecting applicants who are predicted to default, 193 women would be rejected for loans in contrast to 107 men -- nearly the same proportions as seen in the gender-labeled data (see Figure 5).

The hypothetical example using solar lantern sales was developed to make a point, but it is not an unrealistic scenario. Researchers at Carnegie Mellon University revealed that Google ad listings targeted to those seeking high-income jobs were presented to men at nearly six times the rate they were presented to women,[23] despite the fact that gender had been treated as a protected attribute. In the Amazon resume tool example, bias emerged from the data used to train the ML algorithms, which consisted of actual resumes submitted to Amazon over a 10-year period. Because historically men were more likely to have been hired/successfully evaluated, they were understood (by the model) to be better applicants.

As a result, the computer system learned to penalize unprotected features associated with women, such as downgrading candidates who graduated from all-women's colleges. In this example, the bias against female job applicants was a function of the training set. Amazon edited its algorithm to explicitly instruct it to be neutral toward such factors as all-women's colleges. This is, in effect, an extension of the fairness-through-unawareness approach to eliminate not only the explicit label for the protected class but also data elements that are clear correlates with the label. The company acknowledged that this is no guarantee for avoiding discrimination, and it is indeed a very limited remedy.

The Google ad listing tool and Amazon resume tool examples demonstrate that the fairness-through-unawareness-approach may support inequality if the underlying model relies on historical datasets that contain hidden prejudices against underrepresented groups. For example, prior political histories of countries can still result in pervasive inequality even after anti-discrimination laws are implemented. Entrenched disparities in wealth and opportunities can persist, resulting from systemic, institutionalized processes such as racial segregation, forced migrations, gender-restricted education,[24] and discrimination based on caste.

In contrast with fairness-through-unawareness, other, more interventionist remedies to ML fairness take a more proactive approach that explicitly addresses data imbalances. Again, fairness through unawareness is not recommended and the examples given above serve as a warning that it is an ineffective approach to fairness. The four criteria examined in the next subsection require a deeper level of intervention than fairness through unawareness and present a more promising approach to building for fairness.

Fairness Through Awareness

The remainder of this section addresses "fairness through awareness," in which protected attributes are explicitly employed in the ML models. Each subsection covers a different criterion for algorithmic fairness: demographic parity, equalized opportunity, equalized odds, and counterfactual fairness.

RECOMMENDED RESOURCES: MATHEMATICAL FORMULAE & FURTHER READING

The mathematical formulae associated with these four fairness criteria in the computer science literature are provided in Table 3. Readers are referred to the publications in the right-hand column for discussions of these formulae. Full references are provided in the Further Reading list.

23. Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." In proceedings on *Privacy Enhancing Technologies*, 1 (2015): 92–112. https://doi.org/10.1515/popets-2015-0007

24. OECD. *Social Institutions and Gender Index*. (Washington: OECD, 2020) https://www.genderindex.org/

25. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. "Fairness through Awareness". In proceedings of the *3rd Innovations in Theoretical Computer Science Conference*. (New York: Association for Computing Machinery, 2012): 214-226. https://doi.org/10.1145/2090236.2090255

Table 3 — Mathematical Formulae for Algorithmic Fairness Approaches

CRITERION	FORMULA	REFERENCES
Demographic Parity	$P(A = 0) = P(A = 1)$	Hardt 2016; Verma 2018
Equalized Opportunity	$Pr\{\hat{Y} = 1 \mid A + 0, Y = 1\} = Pr\{\hat{Y} = 1 \mid A = 1, Y = 1\}\}$	
Equalized Odds	$Pp(A = 0, Y = y) = p(A = 1, Y = y, y \exists \{0,1\}$	
Counterfactual Fairness	$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a$	Kusner et al 2017

Further reading on algorithmic criteria for fairness

The following publications are recommended for further reading on this topic:

- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In *Advances in Neural Information Processing Systems 29*, Barcelona, December 2016, 4349-4357.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. "Fairness through Awareness". In proceedings of the *3rd Innovations in Theoretical Computer Science Conference*. (New York: Association for Computing Machinery, 2012): 214-226.
- Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." In Advances in Neural Information Processing Systems, edited by D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett (New York: Curran Associates Publishers, 2016): 3315-3323.
- Matt Kusner, Joshua Loftus, Chris Russell and Ricardo Silva. "Counterfactual fairness." In *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017: 4066-4076.
- Sahil Verma and Julia Rubin. "Fairness definitions explained." In *IEEE/ACM International Workshop on Software Fairness* (New York: ACM, May 2018): 1-7.
- Muhammad Bilal Zafar, Isabel Valera, Gomez M Rodriguez, and Krishna P. Gummadi. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." In proceedings of the *26th International Conference on World Wide Web*, Perth, Australia, (Geneva: International World Wide Web Conference Committee, April 2017): 1171- 1180.

Demographic Parity

Demographic parity is the simplest of the widely known mitigation strategies for bias in ML. The approach is to establish a small collection of pre-defined groups and then require parity of some statistic of the outcome across these groups.[26, 27]

If demographic parity is applied to the solar lantern case, the process begins with an algorithm that learns the relationship between loan defaults and employment gaps as before. However, with demographic parity, the ML algorithm is given access to the protected class label — in this case, gender. That label is used to modify the rejection rates based on gender until they are equal for men and women. The approach of demographic parity, as applied in this case, rejects some men who would otherwise have been accepted and accepts some women who otherwise would have been rejected One solution is presented in Figure 8. In this case, 150 women are rejected and 150 men are also rejected.

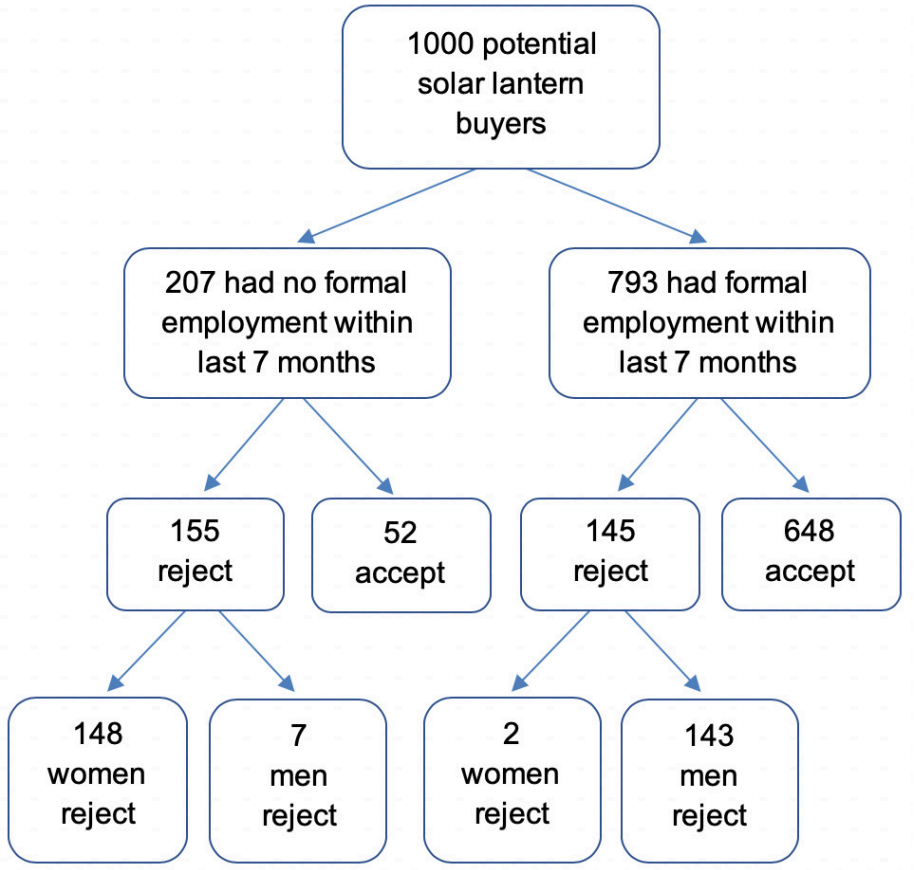


Figure 8 - Demographic parity for gender applied to solar lantern case The algorithm was altered to ensure that resulting rejection rates (here 150 out of 500) were equal for men and women.

26. Faisal Kamiran and Toon Calders. "Classifying without discriminating." In proceedings of *2009 2nd International Conference on Computer, Control and Communication*, Karachi, Pakistan, (New York: IEEE, 2009): 1-6.

27. Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. "Fairness-aware Learning through Regularization Approach." In *2011 11th IEEE International Conference on Data Mining Workshops* (New York: IEEE, 2009): 643-650. http://dx.doi.org/10.1109/ICDMW.2011.83

Among the problems that arise from implementing demographic parity is that it forces equality even when the data set is fundamentally unequal. Although demographic parity establishes a form of group fairness, it fails in some reasonable tests of individual fairness.

For example, imagine that the people who had applied for access to solar lanterns became aware that employment history was a key variable being used in credit decisions. Across those with formal employment history, almost all of those who were rejected were men. So, an individual man whose loan application has been rejected and had formal employment in the past 7 months could argue that he was not treated fairly.

Demographic parity carries some long-term risk to those with the protected class labels if the approach is implemented poorly. Consider, for example, if demographic parity were applied to equalize credit access for urban and rural loan applicants. If the long-term effect is that loan default rates among rural borrowers were elevated as compared to loan defaults among urban borrowers, then bankers might come to distrust rural loan applicants.

Equalized Opportunity

Equalized opportunity is an approach to fairness in ML wherein the true positive rates are forced to be the same between the protected group and everyone else. This has the effect

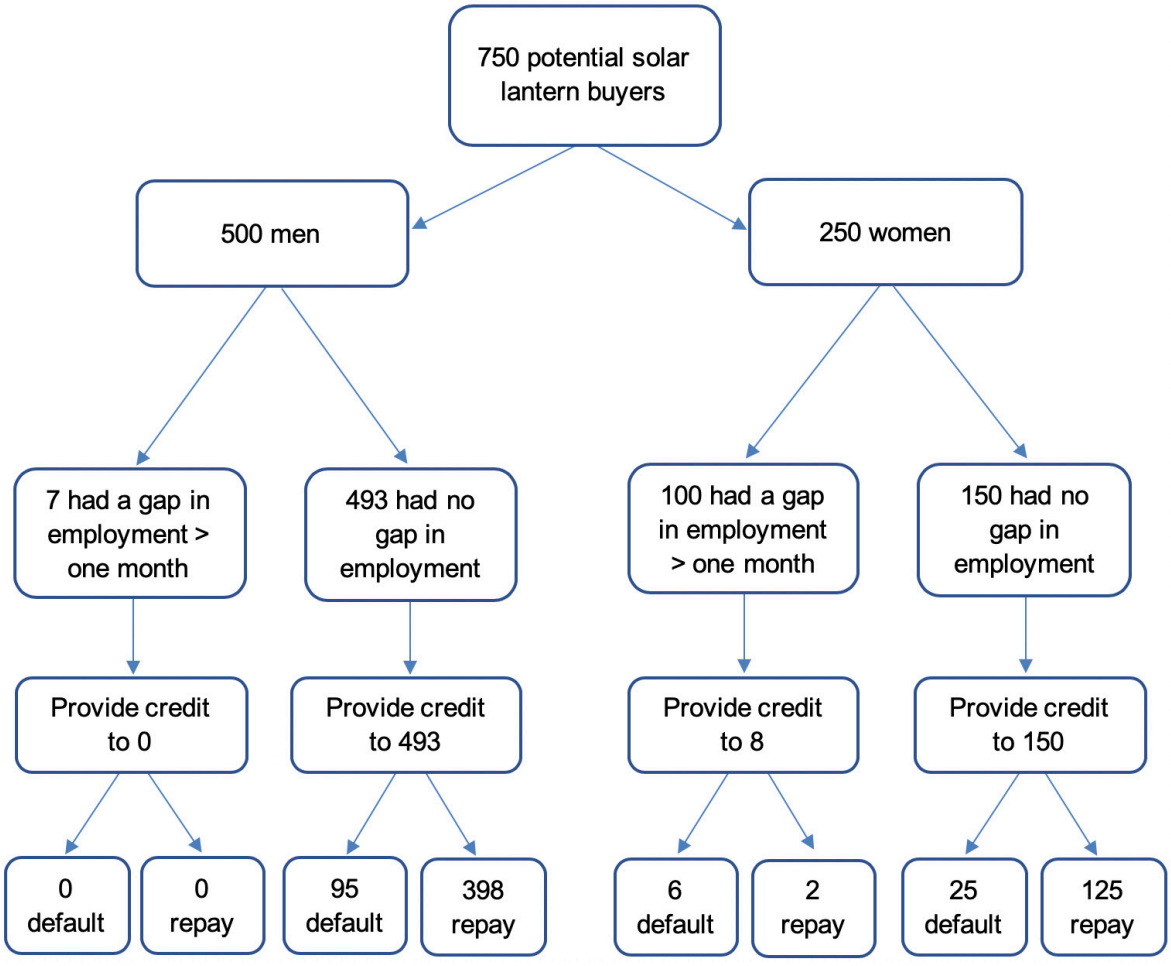


Figure 9 - Equalized opportunity applied to solar lantern case. The algorithm produces a similar, approximately 80% predicted repayment rate on loans granted to men (398 of 493) and women (127 of 158).

of providing social benefits to individuals of both groups at the same relative frequency and it also implies that risks and losses related to providing those benefits are more fairly distributed. To implement equalized opportunity, it is essential to begin with a substantial quantity of labeled historical data. That is, there must be an adequate data set enabling the true positive rates to be estimated and subsequently equalized.

Whereas in demographic parity the predicted outcome is equalized across protected attributes for the entire data set, in equalized opportunity the equality constraint is enforced only on subsets of the population with the positive value of the outcome.

To illustrate equalized opportunity, consider a proposed use of ML to support lending decisions that enable solar lantern sales. Imagine a training data set that established which people actually repay their loans. In the chart below, the proportions of women who lacked formal employment history and the proportions paying back the loans are the same as in the earlier example, however for the purpose of illustration the proportion of women applying for credit is different from that in the previous example.

The "equalized opportunity" framework was used to establish the numbers of loans made to each of the four subgroups: men who had formal employment, men who lacked formal employment, women who had formal employment, and women who lacked formal employment.

MATHEMATICAL PRESENTATION OF EQUALIZED ODDS

Table 4 - Terms for Describing Accuracy

NOTATION	TERM NAME	TERM MEANING
Y	Outcome Variable	The target, i.e. value we are trying to predict (actual value)
$\hat{Y}$	Predicted Outcome	The predicted target value using our model
A	Protected Attribute	A variable in the dataset which encodes a protected attribute
X	Predictor Variable	Any variable used in the model to predict the outcome, i.e. for n predictors our model is $Y=f(X_1,X_2,...,X_n)$
TP	True Positive	$Y= \hat{Y} =1$ correctly classified as positive
FP	False Positive	$Y= 0; \hat{Y} =1$ correctly classified as negative
TN	True Negative	$Y= \hat{Y} =0$ correctly classified as negative
FN	False Negative	$Y = 1; \hat{Y} = 0$ incorrectly classified as negative
TPR	True Positive Rate	TPR = TP/(TP+TN)
FPR	False Positive Rate	FPR = FP/(FP+TN)
ACC	Accuracy	ACC=(TP+TN)/(TP+TN+FP+FN)

The approach of equalized odds relies on having a training data set for which the outcome *Y* is known with certainty. In Table 4 all values of the outcome variable Y and the associated values of its predictors $X_1,X_2,\ldots X_n$ are known, as are the values of the protected attribute *A*, i.e. in the training set $(Y,X_1,X_2,\ldots X_n,A)$ are known.

The *predicted outcome* $\hat{Y}$ is compared with the actual value of the outcome *Y* for the same values of $(X_1,X_2,\ldots X_n,A)$ to determine if there was a misclassification error. For a graphical representation, see the confusion matrix below. The *accuracy* of the algorithm is defined in Table 4 and can be computed from the four cells of the confusion matrix.

Enforcing a fairness criterion does come at the cost of accuracy, which is to be expected as a fairness criterion is an additional constraint (see Zafar et al 2018 in the Further Reading list on page 31 for examples of this trade-off).

Rather than equalize the predicted outcome across protected attributes for the entire data set, equalized odds essentially forces the same equality constraint as demographic parity, but only for subsets with the same value of the outcome Y. Equalizing the odds means to equate the True Positive Rate and the False Positive Rate for different values of the protected attribute, such that the algorithm performs equally well across categories of the protected attribute (see Hardt et al. 2016 in the Further Reading list on page 31.

Equalized odds expand the approach in equalized opportunity to being fair to both those with a positive outcome (in the loan repayment example, those who repay the loan) and to those with negative outcome (those who default). This expansion of fairness criteria is justified if bias leads to unfairness through a false negative outcome (e.g. denying loans to people who would repay) in addition to unfairness through a false positive outcome (e.g. granting loans to people who will not repay).

		PREDICTED	
		Negative	Positive
ACTUAL	Negative	**True Negative (TN)**	**False Positive (FP**
	Positive	**False Negative (FN)**	**True Positive (TP**

The Confusion Matrix

Both criteria of equalized odds and equalized opportunity allow for a perfect predictor (where $\hat{Y} = Y$), whereas demographic parity does not (see Hardt et al 2016). Thus, if the goal is to achieve higher levels of fairness and optimize for accuracy, equalized odds and equalized opportunity are better approaches than demographic parity. Relaxing the equalized odds constraint to just the "advantaged group" – i.e. non-defaulters (Y = 1), yields the equalized opportunity constraint, which is easier to satisfy in practice than equalized odds.

As stated previously, "equalized opportunity" requires that the true positive rate is the same between the men and the women. "Equalized opportunity" was accomplished in this case because 398 out of 493 men repaid, meaning that the probability of repayment given that the loan was made to a man was slightly over 80%. Similarly, because 127 out of 158 women repaid, the probability of repayment given that the loan was made to a woman was slightly over 80%. Under this "equalized opportunity" scheme, the company chooses to accept similar risks in both applicant pools. From that perspective, the approach appears to be fair.

Equalized odds

Equalized odds (aka predictive value parity) is an approach to fairness in ML that is similar to equalized opportunity but places an additional constraint on the algorithm. In equalized odds, both the true positive rates and the false negative rates are equalized between the protected groups. Equalized odds most often drives the ML algorithm to sacrifice accuracy in order to satisfy additional criteria of fairness.

To understand the difference between equalized odds and equalized opportunity, consider the solar lanterns case. In the illustration of equalized opportunity, 7 men were not provided credit. If any of these 7 men would have repaid, these instances would be considered false negatives. The false negative rate can be estimated because all of the men who were denied a loan lacked formal employment. In the training population of men with no formal employment, 2 of 7 paid the loan, meaning that the probability of payment given that the loan was denied to a man was about 28%.

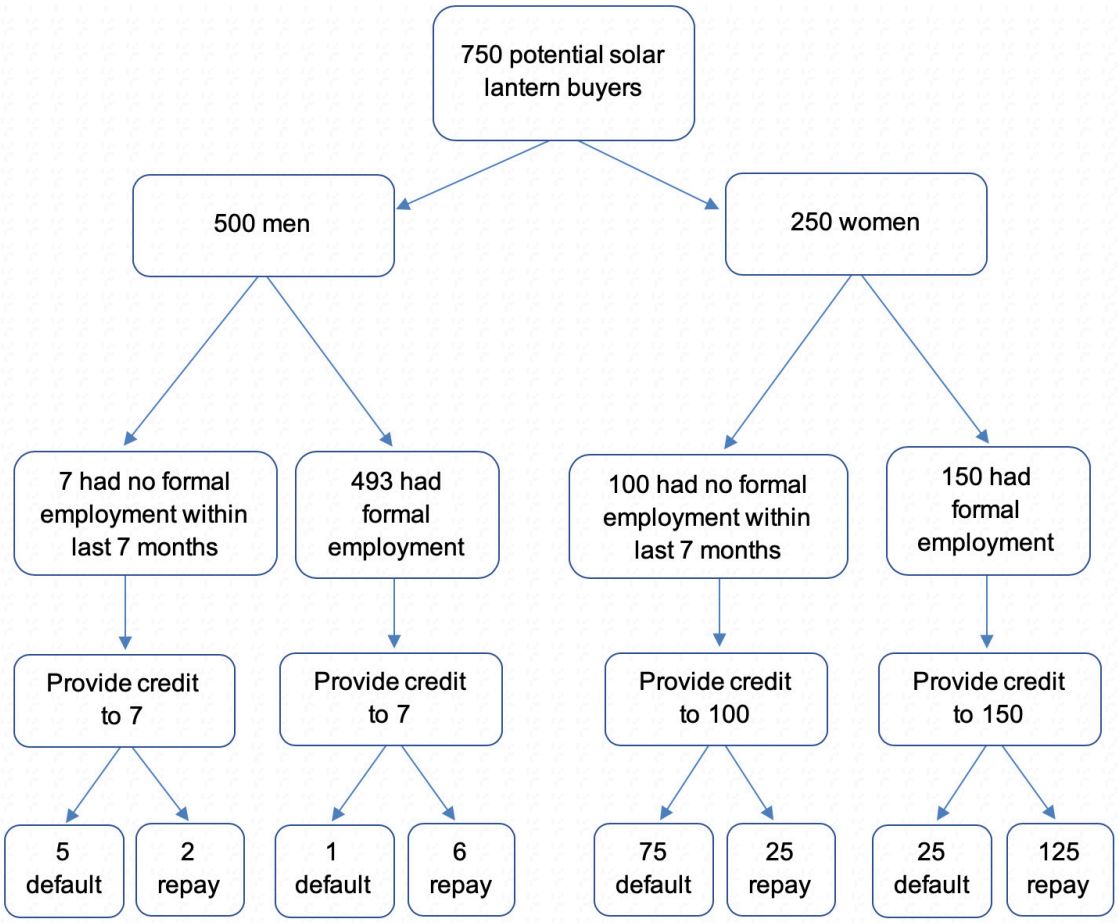


Figure 10 - Equalized odds applied to solar lantern example. The false positive and false negative rates are both equal, but the default rates are high.

In the illustration of equalized opportunity, 94 women were not provided credit. For any woman among them who would have repaid, this denial of credit is a false negative. That false negative rate can be estimated because 64 of the women who were denied a loan lacked formal employment. In the whole population of women lacking formal employment, 50 of 200 paid the loan, meaning that 16 of the women with no formal employment would have repaid. In addition, 78 of the women who were denied a loan had a history of formal employment in the past 7 months. In the whole population of women with formal employment, 250 of 300 paid the loan, meaning that 65 of the women with formal employment would have repaid. Therefore (16+65) or 81 women who were denied loans would have repaid. The probability of payment given that the loan was denied to a woman was 81 out of 94 or about 86%.

It can be argued that the equalized opportunity framework was not fair because it accepted an 86% chance of denying loans to women who would have repaid, whereas it accepted only a 28% chance of denying loans to men who would have repaid.

How could the algorithm for determining provision of credit accomplish equalized odds in the case of the solar lanterns? Changes would have to be made in the numbers of people provided credit across each of the four subgroups: men who had a gap in employment, men who had no gap in employment, women who had a gap in employment, and women who had no gap in employment.

The "equalized odds" framework was used to establish the numbers of loans made to each of the four subgroups: men who had a gap in employment, men who had no formal employment, women who had a gap in employment, and women who had no gap in employment. As stated previously, "equalized odds" requires that the true positive rate is the same between the men and the women. This goal was accomplished in this case because 8 out of 14 men repaid, meaning that the probability of repayment given that the loan was made to a man was about 60%. Similarly, because 150 out of 250 women repaid, the probability of repayment given that the loan was made to a woman was also about 60%. The true positive rates have parity, so that seems fair.

Next, the false negative rates are reviewed. In this illustration, 486 men were not provided credit. Of these men denied a loan, none had a gap in employment. In the training population of men with no gap in employment, 398 of 493 paid the loan, meaning that the probability of payment given that the loan was not made to a man was about 80%. In the illustration of equalized odds, 1 woman was not provided credit and she had no gap in employment. In the training population, 250 out of 300 women with no gap in employment would have repaid the loan. The probability of payment given that the loan was denied to a woman was about 80%. The equalized odds framework was fair in the sense that it accepted an 80% chance of denying loans to women who would have repaid and also accepted an 80% chance of denying loans to men who would have repaid.

However, there was a downside to the equalized odds approach as compared to the equalized opportunity approach applied to this solar lantern case. Under equalized odds, there were 106 defaults out of 264 loans for a default rate of about 40%. Under equalized opportunity, there were 126 defaults made out of 651 loans made for a default rate of about 20%. The equalized odds framework was fairer, but it was also far less accurate, and the cost was a doubling of the default rates on the loans.

Counterfactual Fairness

Among the newer and more complex methods in ML is counterfactual fairness. In this approach two groups are guaranteed the same predicted outcome if the protected class status were different, all other things being equal — for example, if all the genders were switched. In this approach, data sets are altered to represent the counterfactual circumstance that an individual belongs to one group when, in fact, they actually belong to the other (but only in cases where the explanatory variables are equivalent).

This algorithmic construct is the same as that of demographic parity except only under the quite challenging demands of a realistically constructed counterfactual scenario. Consequently, this approach requires a mapping of the causal relationships among variables, selection of exogenous variables, and regression analysis of those relationships. When a protected attribute is flipped to the counterfactual value, that change must be propagated to the other variables that are deemed causally dependent on the protected attribute that was changed, which is why the counterfactual analysis must precede implementation.

It is challenging to work out, in realistic terms, what is meant by the counterfactual "had they been male instead of female." For example, opportunities for formal employment are causally connected to gender in many societies. So, to be fair, in the counterfactual fairness approach, an effort is made to sort out any plausible connections relevant to the decision.

Consider using the counterfactual framework for the solar lantern case study. For half of the loan applications made by men, the gender label is switched from "man" to "woman." For any individual application in this group, because this application now comes from a woman, there is some probability that this newly created "counterfactual person" would have been denied an opportunity for formal employment. As a result, some fraction of the data elements for those men with the counterfactual gender label would have to be altered so that they lack formal employment history. In the end, the decision algorithm has to treat men no differently whether they retain the original gender label or whether the label and associated data are altered.

The causal relationships analyzed within counterfactual fairness are precisely what is neglected in fairness-through-unawareness. The two approaches would give similar results if the causal structure were the same in both groups, but they rarely are.

An important advantage of the approach of counterfactual fairness is that it appeals to a deep notion of what fairness requires. By directly addressing the causes of bias and inequality, counterfactual fairness does a more thorough job of accounting for the systemic nature of biases that surround commonly protected attributes of gender, employment, race, and other factors.

However, counterfactual fairness is an approach rather than a single, well-defined technique. The results depend critically on many implementation details. In practice, the counterfactual fairness approach provides diagnostic tools rather than prescriptive solution methods.

Fairness Methodology

This section proposes a methodology for choosing among the fairness criteria covered in the previous section. The methodology presented in Figure 12 is largely data-driven with a central role played by the protected attributes. To streamline the methodology, the questions asked of the implementation team are designed to be relatively few. A general discussion of how to apply this model is followed by an example of its application to the solar asset loan case.

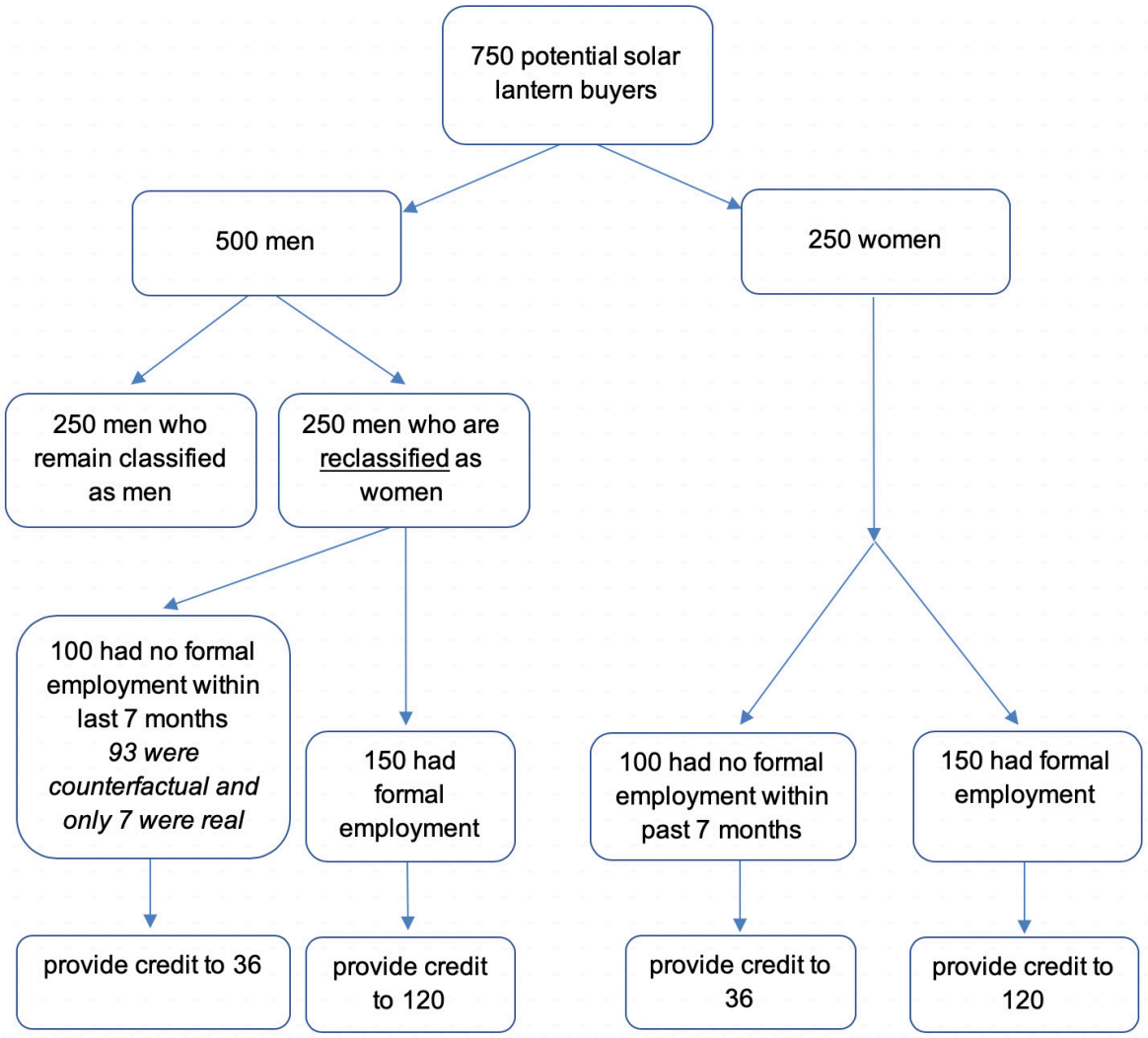


Figure 11 - Counterfactual fairness applied to solar lantern example. Half of the men were reclassified as women in the data set, with conditions that applied to women (likelihood of formal employment) applied to those reclassified men. Doing so produces the same credit outcomes for the reclassified men as for the women.

As shown in Figure 12, if one is compelled to demonstrate statistical parity by law or by policy (e.g. hiring equal numbers of men and women), then demographic parity should be pursued, but with a careful consideration of other fairness criteria as far as the law allows. It is likely that the demographic parity goal does not fully determine the machine learning implementation. There will still be many choices to make and those can be informed by some of the more nuanced fairness frameworks and tools.

If demographic parity is not required, a question for the implementation team to address early on is whether there is a reasonable prospect of creating a causal model for the relevant fair prediction scenarios. Fundamentally, any effort to design a system consistent with both individual and group fairness will have to address the reasons that bring about unfair outcomes. If that higher standard can be met within reasonable time and budget constraints, then the implementation team should attempt to pursue that avenue through the framework of counterfactual fairness.

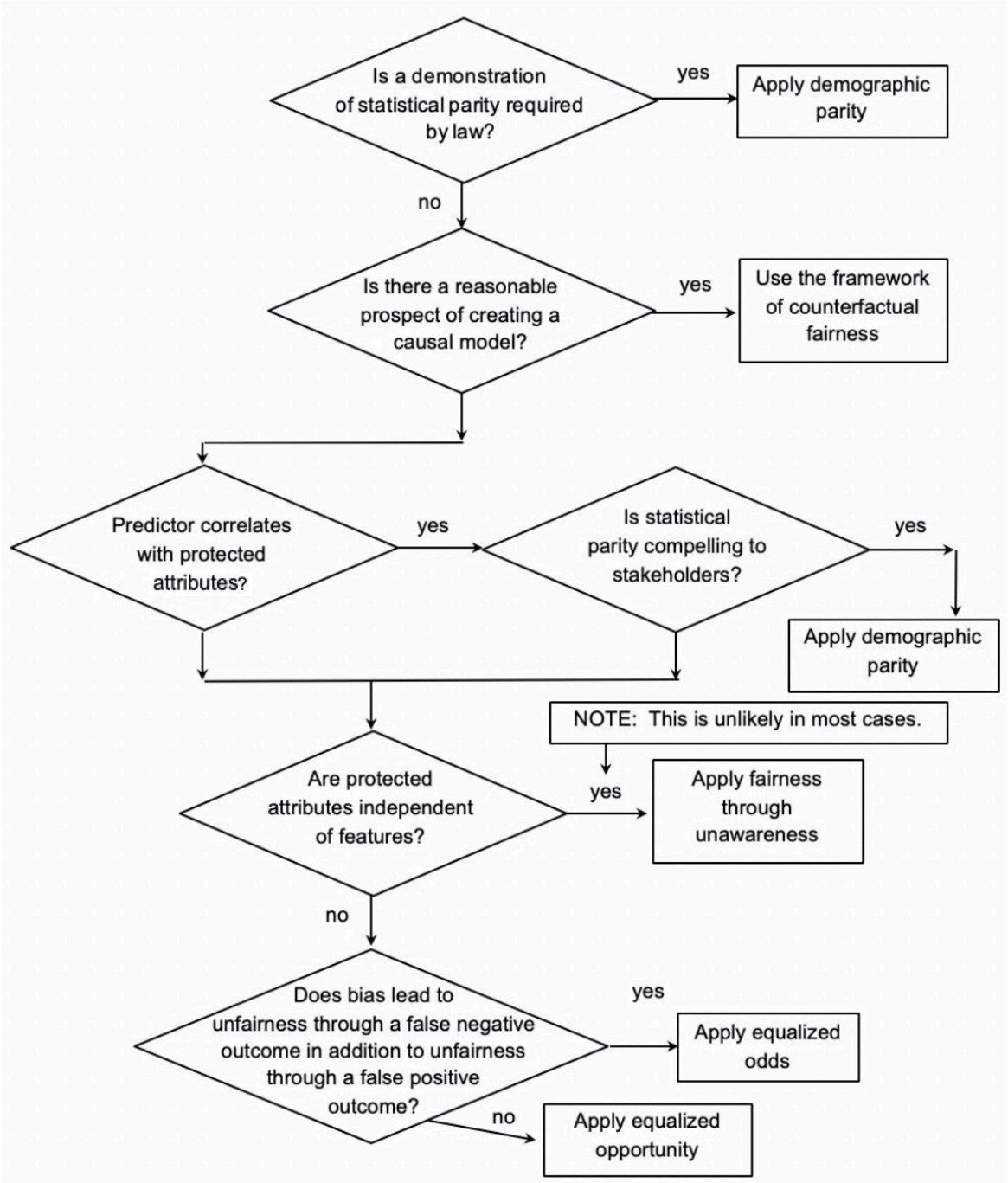


Figure 12 - Decision tree for selecting an appropriate fairness criterion

If a causal model is viewed as too complex or costly to build and validate, then some more frugal alternative can be pursued. For example, the team could adopt demographic parity criteria. However, that step can lead to objections regarding individual fairness when pair-wise comparisons are made between similar decision cases that differ only in the group membership and the outcome. The team therefore has to decide if the demographic parity approach is compelling for the stakeholders.

If there is not an adequately compelling case for demographic parity, then further alternatives should be considered. A principal determination that should be made is whether protected attributes are independent of other features or predictors. However, most research and most experience from practice has shown such independence to be very unlikely: protected attributes are almost always correlated with other features in the data, for example gender with employment or socioeconomic status with residential district. In the rare case that this criterion is met, fairness through unawareness can be applied by simply removing the labels that indicate membership in the protected class. Vigilance should be exercised to ensure that outcomes are fair for the protected class.

At this point in the flow chart in Figure 12, the team should determine if bias leads to unfairness through a false negative outcome (e.g. denying loans to people who would repay) in addition to unfairness through a false positive outcome (e.g. granting loans to people who will not repay). If so, the team should apply the algorithmic criteria of fairness for equalized odds. However, this approach can significantly degrade the accuracy of the model; it should be established with a high degree of confidence that the negative outcomes have a large influence on fairness. If negative outcomes are not a significant concern — for example, if the priority is to grant loans to men and women with similar prospects of repayment but it is acceptable to deny loans to men and women with differing prospects of repayment — then the team should apply the algorithmic criteria for equalized opportunity.

Applying the Fairness Methodology - Solar Lantern Example

Consider the solar lantern example. To simplify, imagine gender is the only protected attribute of interest (in practice, additional attributes would most likely be considered, such as age).

Is a demonstration of statistical parity required by law?

The process begins with checking to see if there is any legal framework that requires statistical parity across protected variables. Assume that, in the country in which the intervention is being implemented, legal frameworks exist to prevent discrimination on the basis of gender, but there are no specifications for if and how they apply to loans and microfinancing. Because the organization is not legally required to apply demographic parity, the answer is no and the organization can proceed with implementing another fairness method that may yield better results.

Is there a reasonable prospect of creating a causal model?

Theoretically it may be possible to build a causal model to account for certain differences affecting women's employment histories, such as having gaps in employment due to childbirth. However, there are several other considerations for gender inequality that may affect the model. For example, there may be an education gap between men and women in the region of implementation, which may result in women having lower paying jobs. Additionally, in the region, women may also work informal jobs at higher rates than men, which may result in employment data variations. Building a causal model to account for these considerations may not be possible or may be prohibitively expensive. Therefore, the organization establishes that it is not a reasonable prospect to build such a model and proceeds to exploring other options.

Is statistical parity compelling to stakeholders?

For the funders and key stakeholders, is there a clear benefit derived from attaining statistical parity? The decision to achieve statistical parity is often easier to communicate externally to funders, beneficiaries, governments, or other stakeholders because the motivation and process for employing a more complex fairness criteria may be challenging to understand. However, if statistical parity has detrimental impacts on populations that are traditionally disadvantaged, like women in financial inclusion, it may make sense for the organization to implement alternative fairness methods that are more favorable to the traditionally disadvantaged.

Are protected attributes independent of other features?

The data from which the machine learning model is being built includes a variety of factors to determine credit-worthiness, but let us assume that they are all uncorrelated with gender. While this scenario may be unlikely, fairness through unawareness would be an admissible fairness strategy because these correlations could not cause the protected attribute to be inferred after it is removed from the dataset.

Does bias lead to unfairness through a negative outcome in addition to unfairness through a positive outcome?

When answering this question, a key value judgment will be implicitly made about whose interests will take priority. For this example, this question is asking whether it is sufficient to give loans to men expected to repay at the same rate as women expected to repay or whether the model also needs to ensure that men expected to default are denied loans at the same rate as women expected to default. If the solar lantern company answers the question with "No," this implies that the positive outcome (qualified applicants accessing loans) is more important to the stakeholders than the negative outcome (unqualified applicants being denied loans). Before answering "no," the company should be reasonably confident that bias does not lead to unfairness through a negative outcome and therefore the organization is justified in applying **equalized opportunity**. One additional layer of complexity is that stakeholders may not necessarily have the same incentives when it comes to implementing fairness criteria or see the same cost for false positives and false negatives. For example, the lender might be more interested in reducing the number of loans improperly approved (which would be actual defaults) whereas the applicants would be more interested in reducing the number of loans unfairly denied.

The case study section provides another example of application of this Fairness Methodology and describes a real-world approach to mitigating bias in a health care intervention.

D. Deployment + Maintenance

This section covers important considerations in Steps 7 and 8: Deployment and Maintenance.

Documentation

Writing high-quality ML programs is good, but not sufficient. Organizations also need thorough and well- written documentation of ML codes and data sets.

This documentation can serve several main purposes: guiding the programming process, conveying information about training, aiding resolution of problems during use, and facilitating knowledge transfer to other developers.

ML documentation should cover the following topics, at a minimum:

- When and by whom the ML program was developed
- The purpose and intended users of the ML program, including specification of circumstances under which use is recommended and any cases for which use is discouraged
- Requirements for hardware and environments on which the application will run
- Essential facts about the data such as the origins of the training data, labeling procedures, and provisions for data quality
- Assumptions made by the development team
- Algorithms employed and their limitations
- Data collection methods including sampling strategies employed and assurances that sample sizes are adequate to the task yet efficient
- Steps that have been taken to mitigate bias

Additionally, updates and revisions of the ML program should be documented in real time, if possible, or recorded immediately after they are made.

Testing

Before an ML system is put into use, there must be clear evidence that the system meets the requirements that the organization and the review committee have specified.

While user testing is ideal and should be employed for the system's most critical requirements, it can be costly. Simulation, analysis, and peer review can also be used to address less critical requirements. To illustrate the distinction, consider an ML system that is designed to support the mental health of refugees. The ways that the system responds to users that are having a disagreement with a friend might be checked by peer review. The ways that the system responds to users that are experiencing suicidal ideation probably need to be validated through a realistic set of tests with actual users that are overseen by qualified psychiatric staff.

Prior to launching an ML system, the following steps must be completed:

1. Validate that the requirements of the ML system match the stakeholders' needs. The requirements must include considerations of fairness as discussed in this document.
2. Establish a testing plan to ensure that requirements are being met. The testing plan must be sufficiently detailed to ensure that inputs to the system have been described and the corresponding outputs are adequately characterized.

   The testing plan must answer the questions:

   - What kinds of input /output relationships would be evidence of bias?
   - How will such bias be measured?
   - How will bias be addressed?
3. Complete and document the testing procedure. Record and resolve any anomalies observed during the ML systems tests

METHODS TO EVALUATE FAIRNESS

CITE emphasizes that even if the recommendations in this chapter are followed carefully, final checks on the outcomes are an essential part of an end-to-end development process.

A system should not be assumed to be fair unless it has been demonstrated to be fair. The following methods can be used to evaluate fairness:

» Targeted adversarial testing is known to be effective at identifying bad outcomes even for relatively infrequent outputs. For this approach, an independent group challenges an organization to improve its effectiveness by assuming an adversarial role or point of view. In software development, it is beneficial to organize a pool of trusted, diverse testers who can test the system in this way and incorporate a variety of adversarial inputs into unit tests. For example, in a credit lending application, the loan program should be subjected to adversarial testing by people who understand how the credit system works, perhaps including people who were unfairly rejected for loans.

» Design testing metrics that are likely to reveal differences in outcomes across subgroups of users. False positive and false negative rates can be particularly helpful when applied across different user classifications.

» Stress-test the system on difficult cases. Stress-testing refers to tests that run the model through extreme situations. The effectiveness and limitations of such methods have been demonstrated through their widespread use by financial regulators. A stress test, in financial terminology, is an analysis or simulation designed to determine the ability of a given financial instrument or financial institution to deal with an economic crisis. As an alternative to financial projection using averages or optimistic scenarios, a company or its regulators may do stress testing to determine how robust a financial instrument is during bleak scenarios like crashes. Similarly, ML developers should seek to identify particularly harmful or problematic circumstances and expose the system to these circumstances, perhaps via simulations. Such stress tests should be updated frequently to reflect the latest emerging challenges faced within a field.

» Consider the possible effects of feedback loops within the ML system and within interconnected systems that combine with the ML system. In some cases, biases and unfair outcomes can be amplified over time due to such interconnections. For example, an employment matching ML system may send new prospects to certain companies that preferentially selected from one ethnic group because the company leadership was more comfortable interacting with people of that ethnicity. That preference may be reinforced and legitimized when the ML system responds to and follows that established pattern. This example highlights a negative consequence of feedback loops. When such loops are implemented thoughtfully, they are more often beneficial. Guidance on effective construction and use of digital feedback loops is available in USAID's Guide to Digital Feedback Loops.

Accountability

As discussed in the ML implementation principles in Chapter 2, accountability is essential for responsible use of ML. Central to the concept of accountability is the requirement that people are being held accountable. This observation motivates the concept of a "sign-off" on all major decisions related to ML development programs in international development. A well-articulated formal process for sign-off approved by the review committee and involving documentation should be developed prior to any ML implementation. A sign-off need not be by a single person only, but in many cases one person would suffice. Related to the sign-off is an important question of availability. The responsible persons must be available to devote time and attention to the sign-off process. In a USAID context, it's common to have a project supervised by a COR/AOR (Contracting or Agreement Officer's Representative) on the USAID side and a Chief of Party on the implementer side. In many interventions involving ML, CITE recommends that both of these individuals be involved in a sign off process and that these individuals recognize that it is a major commitment to develop sufficient familiarity with the implementation details.

In an international development context, strong institutional mechanisms for accountability are especially important. Because international development programs are often designed to assist the poorest and most vulnerable people, it is important to recognize that those targeted by ML interventions in international development are also those least likely to have the resources and organizational power to seek redress when they are harmed. Poverty, inequality, and underrepresentation in governance can limit the ability of individuals and communities to organize and advocate for countermeasures when interventions have negative consequences. Accountability mechanisms must be designed in such a way that the organization engages in self-monitoring and takes action to protect those affected by its programs when problems occur.

Chapter 5: Conclusions

This chapter describes a set of recommendations that go beyond the development of ML models to include governance structures and broader practices that drive toward fair use of ML in development.

RECOMMENDED RESOURCES

Interested readers can find additional details and perspectives within Responsible AI Practices by Google AI and the website Principles for Digital Development. Further, the United States led the development of the OECD Recommendations on Artificial Intelligence, which includes principles for the responsible stewardship of trustworthy AI. These principles consider issues like fairness, transparency, and accountability. The OECD is now developing implementation guidance to help move from these high-level principles to practice. The guidance, and other reference materials, will be posted on the OECD AI Policy Observatory. Some of the same recommendations made by these organizations are also promoted in this chapter.

Human-Centered Design

CITE strongly advises that developers of machine learning systems employ a "human-centered" design approach. The International Organization for Standardization (ISO) defines human-centered design as:

> *an approach to interactive systems development that aims to make systems usable and useful by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, usability knowledge, and techniques.*[28]

Specifically, CITE recommends that system designers:

- **Engage a diverse population of potential users** of the system to develop the systems' specifications. By ensuring that a wide variety of people have expressed their needs, organizations can avoid unintentionally designing the system for only a narrow sub-population. For example, an ML system for credit scoring would benefit from inputs from both urban and rural users who may have very different needs for loan products.
- **Employ a variety of different use-case scenarios.** A use-case is a description of a set of interactions between a user (usually a human) and a software or ML system. By diversifying the use-cases employed in system design, organizations can make the system fairer by respecting the ways that different people are likely to use the system. For example, use cases for an employment-matching ML system could range from a large company launching a new factory to a service that supports many small companies.
- **Disclose data collection.** Users should be clearly informed about their choices and options with respect to how their data are used. When users are interacting with an ML-enabled application, ensure that disclosures are made at relevant times alerting the user that the user's sensitive data is being collected and allowing the user to opt out when possible. This disclosure-and-decision principle helps enhance the transpar-

28. ISO. *ISO 9241-210:2010 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems*. Geneva: ISO, 2010.

ency of the software's process, and enables greater control by the user. For example, many web sites inform users when cookies will be used and some request consent for that type of data collection. The distinction between disclosure and consent should be emphasized and genuinely informed consent (thoroughly explained, understandable, and assessed) and is strongly preferred to a pro forma collection of a signature.

» **Favor user control over automation.** Whenever possible, design the ML system to provide augmentation of a user's capabilities and assistance of users in a task as opposed to automation of tasks. For example, rather than providing a single answer or suggested solution step to the user, provide a list of options instead. This advice has an important technical basis as studies have shown that the accuracy of a system's predictions is often enhanced when that precision is defined over a set of answers rather than a single-point solution. For example, an ML system for supporting decisions on which crops to plant should provide a list, such as five crops that are likely to thrive rather than just a single most highly-rated choice. Subsequent evaluation of the ML tool will be more robust because a greater variety of crops would be in consideration on the basis of the ML system's recommendations.

» **Prepare for potentially adverse (problematic) feedback** early in the design of your software system. By explicitly recognizing and anticipating unintended outcomes and planning to mitigate their effects, a system becomes more fault tolerant. Live testing and feedback with even a small group of potential users can help to identify these adverse feedback scenarios and validate the countermeasures against them.

Implementing Fairness Strategies

Chapter 3 introduced an 8-step overview of the process implementing machine learning solutions in international development and highlighted the potential for fairness considerations to arise at each step. Figure 13 builds on the figure introduced in Chapter 3 to demonstrate how and where different strategies for supporting fairness can be implemented during the ML project lifecycle.

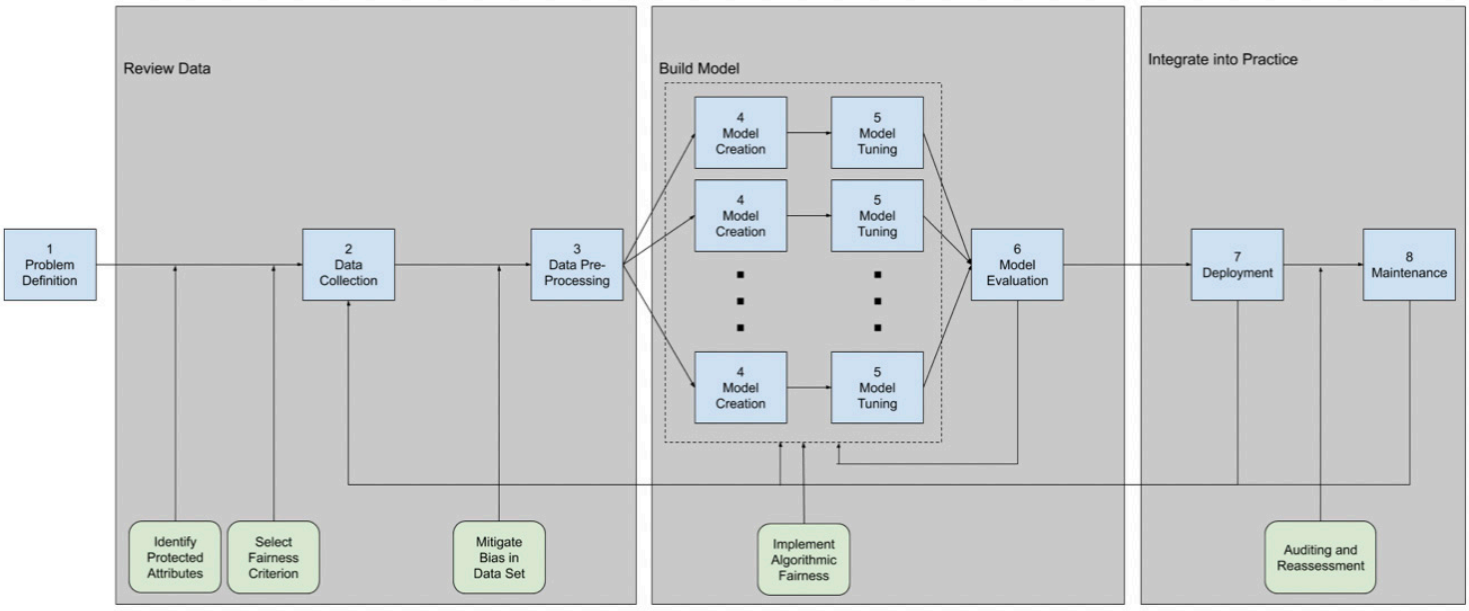


Figure 13 - Addressing fairness and bias throughout the ML process

Initial considerations: Problem definition

Teams should review the guiding principles discussed in Chapter 2 prior to implementing ML solutions in practice. When a team employs ML in an international development context, it is essential to ensure relevance of the ML effort. Relevant applications of ML have clear value to stakeholders and address the priorities of the impacted communities. If a simpler approach can solve the problem adequately and cost effectively, then ML should not be used.

Prior to data collection, it is important to identify which variables in the data are protected attributes. As mentioned in Chapter 3, certain variables may be designated as protected attributes by legal restrictions. However, in most cases in the international development context, legal protections are not robust enough to provide protection for all groups that may be marginalized, and organizations will need to identify the appropriate variables to ensure fair outcomes. Chapter 1 delves into definitions of fairness in greater detail.

When ML is the chosen approach, organizations should form or draw on existing ethical review committees. These committees should comprise multiple individuals who work together to evaluate the ethical standing of all aspects of the ML use process. Ethical review committees can serve as a natural check-and-balance for ML, allowing humans and machines to work in concert.

One imperative from the outset of a project is to ensure that the design and impacts of the ML system are communicated clearly to all stakeholders, particularly to those who will be affected by the ML-augmented intervention. Auditability and accountability are also especially important when ML systems affect people living in poverty. The technical professionals on the team should frequently ask how the model's decision-making processes can be queried or monitored by external actors. The team's leaders must ensure that someone will be responsible for responding to feedback and redressing harms, if necessary. These issues are discussed further in Chapter 2.

Phase 1: Review data

After choosing protected attributes, it is important to observe if any features are correlated with protected attributes before determining which fairness criterion is most appropriate to implement. Features that are highly correlated with protected attributes should be treated as protected attributes, and correlations in general may need special attention. Based on these correlations and other information about the problem, an appropriate fairness criterion can be selected to implement fairness. Choosing among the various approaches available is a central task for the design team. Chapter 4 details the selection of these fairness criteria.

Curation, cleaning, and labeling of data is also central to the success of ML in general and is essential in an international development context. The technical professionals involved must evaluate representativeness of the data for its intended use. For example, using data from one region to train an ML system that is then deployed in another region can result in inaccurate and unreliable models. Identifying issues of representativeness and balance in datasets will often require domain expertise from international development professionals, as the ML implementing team may need additional, contextually relevant information.

In most real-world cases, data will have gaps and errors and may not be ideally representative or balanced. Teams will need to determine if and how to overcome these challenges and making choices about balancing different tradeoffs. The preferred approach is always to gather a more diverse data set, However, this step may not be feasible due to high costs and long timelines. Other approaches to overcome limitations including data augmentation, resampling, and generation of synthetic data are discussed in Chapter 3. For example, in the data preprocessing stage, individuals should make adjustments that normalize prejudiced historical data before the algorithm is formulated.

Phase 2: Build model

The model building phase includes the creation, tuning, and evaluation of one or more ML models. ML implementers should pay special attention to the types of algorithms that are chosen in the model creation, particularly their strengths and weaknesses. Appropriate algorithm choice can reduce the risk of bias and optimize for fairness considerations. Further discussion of bias and fairness considerations for specific algorithms can be found in the Appendix.

Phase 3: Integrate into practice

In the final phase, the results of the ML models are validated. Model outputs should routinely be checked for errors and biases. The results from these findings should be used to calibrate the classifier such that desired true positives, false positives, true negatives, and false negative rates are achieved and are not reflective of data biases.

As time passes after the ML model is initially created, the likelihood increases that the model will become inaccurate. In international development, such changes affecting the accuracy of models commonly include population demographics; new taxes, subsidies, or government policies; and introduction of technology. If underlying assumptions in building the model are no longer accurate, the model needs to be corrected. Reassessing the fairness of the ML solution should be part of a regular schedule, similar to maintaining the codebase of the ML implementation.

In addition to these technical considerations, organizations should implement initiatives that emphasize ethics and encourage human involvement throughout each stage of the use of ML outputs. Organizations can also conduct training seminars that educate employees on the strengths and weaknesses of different fairness algorithms or offer skill-building activities, such as role-playing exercises that teach individuals to evaluate the extent to which ML models are fair or discriminatory.

Ethics in machine learning is a growing area of research, and fairness and bias are two important aspects of this larger field of discussion. ML implementers are encouraged not only to use this framework and guiding principles, but to also to keep up to date with newer techniques as they continue to emerge. The resources at fatconference.org, ainowinstitute.org, and ai.google/responsibilities are great starting points for further reading and reflection.

Case Study: Machine Learning and Bias in Medical Diagnosis

This chapter offers a case study exploring a real-world approach to bias in machine learning. It focuses on a case involving diagnosis of pulmonary diseases in Pune, India that required exploration of bias with respect to gender and socio-economic status. The chapter first offers some context for the case study by discussing some of the challenges with using machine learning in the medical field. Two brief examples are explored before presenting the detailed case study.

A. Background on ML within Global Health Efforts

In developing countries, a lack of electronic medical records has delayed the implementation of data science. In addition to a variety of domestic public and private health organizations, the health ministries in many developing countries also collaborate with a variety of international organizations, such as the World Health Organization (WHO), which often create guidelines that are used by health personnel to diagnose and treat disease. Furthermore, the emerging and widespread use of mobile phones and mobile apps has begun to create new opportunities to apply data science and machine learning to the delivery of health care services in low-resource areas.

Due to high density of population and lack of infrastructure, a common application of machine learning in global health screening for diseases, which is often performed by community health workers or by organizations that conduct specific health camps. In this application, the goal is to identify individuals who are at high risk of having a specific disease and then refer them to clinics in the health system to seek a diagnostic test and follow-up treatment.

In health facilities, machine learning is also being considered as a decision support tool for general-practitioner doctors or nurses. Future and emerging uses of machine learning include: automatic analysis of radiology images (e.g., X-Rays), automatic interpretation of genetic testing, epidemiology, logistics and operation of medical supplies or personnel.

Challenges of Health and Biomedical Data

The use of health and biomedical data involves a number of challenges and considerations, discussed below.

Human impact of decisions

Decisions regarding health can have a significant impact on individuals and their loved ones. Many aspects of health affect well-being, mobility, and the ability to work and care for oneself and others — and certain medical decisions can have drastic consequences. Errors that lead to misdiagnosis or incorrect treatment can be catastrophic for patients, and can also impact the reputation and future viability of outreach efforts and clinics.

Privacy and legal concerns

Individuals' health and biomedical data are particularly sensitive due to their intimate and personal nature. This type of data is often tightly protected by local government regulations. In the United States, for example, patient data systems are regulated by specific regulations and standards, such as the Health Insurance Portability and Accountability Act (HIPAA). Such regulations also exist in many developing countries and this landscape is changing rapidly with emerging regulatory and data privacy frameworks such as India's Digital Information Security in Healthcare Act (DISHA). Such regulations often have mandates that concern the anonymity of the data as well as the storage, use, and handling of medical data. These constraints add additional complexity to any machine learning project, and can also affect how the data can be used.

Complexity of disease and diagnostic models

The cause and etiology of disease is often complex. Risk factors of diseases may be numerous and are not always known. Genetics, behavior, and environment can conspire with other factors to determine an individual's disease risk. As a result, the data used for building machine learning models for disease and diagnostics models are often incomplete or non-exhaustive, which can lead to surprising or erroneous results.

While it is often necessary to simplify a machine learning model to create binary features, (e.g. smoking = yes/no, breathlessness = yes/no, stress = yes/no), it is important to keep in mind that the field of medicine is very much an analog science with continuous variables. Many diseases, both infectious and non-communicable, (e.g. malaria, pneumonia, asthma) have varying levels of severity, which introduces a degree of variability into the model (if all severities are treated equally).

Difficulty of labelling data accurately

In developing a machine learning model, the process of supervised learning requires training data. This training data requires additional effort on the part of doctors and clinical staff to manually label the data. While some labelling can be accomplished by a simple diagnostic lab test (to confirm YES/NO if the patient has tuberculosis for example), other types of labelling are subjective and rely on the doctors interpretation and experience. For example, some doctors listening to a lunch sound might interpret the sound as a wheeze and some may not. The subjective nature of certain data labelling tasks is important to keep in mind when deciding if a specific problem is appropriate for machine learning analysis.

Genetic predispositions

While it is hoped that diagnostic algorithms and machine learning could be applied fairly and equally to people of all racial and ethnic groups, this ideal is complicated by the fact that there exist significant disparities in the prevalence of certain diseases among different racial and ethnic groups (e.g. cardiovascular disease in the South Asian population or type 2 diabetes in the African-American population). There are also significant physiological variations across racial and ethnic groups. For example, the South Asian population has smaller lung capacity than Caucasian European population. Knowing this, it is possible to see why a given algorithm may produce different results for different groups, with different rates of false positives and false negatives in each group. As medical applications such as pharmacological therapies and cancer treatments are becoming more personalized, it is likely that machine learning algorithms will eventually need to be tailored and personalized as well.

Interpretability

Given that health outcomes and disease can be critically dependent on factors of race, ethnicity, and demographics, in these cases, it is generally important to employ machine learning models that can be interpreted, so doctors can make a better connection between specific risk factors and disease. For this reason, opaque algorithms, such as neural networks, are particularly problematic when applied to disease prediction. However, at the same time, opaque algorithms such as convolutional neural networks (CNNs) can provide very good performance for specific biomedical tasks, such as automatic image segmentation and cancer tumor detection in X-ray images.

Because of these challenges, health and biomedical data is a risky domain in which to apply ML. However, growing populations in developing areas and increasing health costs are driving the exploration and adoption of machine learning to all levels of health care systems in international development.

Brief Examples of Bias Considerations in ML Health Applications

The ethical questions around ML can be subtle and depend on individual perspectives and beliefs. However, examples can also illustrate common themes such as using appropriate proxies and ensuring that training data is relevant to the patient population.

Example #1: Hospital Admission for pneumonia patient

One famous example of machine learning applied to hospital patient admission was published by Caruna [2015].[29] This example involved a study in the 1990s to use a com-

29. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospi tal 30-day readmission." Proceedings of the *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York: Association for Computing Machinery, 2015): 1721-1730. https://doi.org/10.1145/2783258.2788613

puter algorithm to decide which pneumonia patients should be hospitalized and which should be sent home for outpatient care. The original algorithm predicted the 30-day probability of death for each patient; those with a higher probability of death would be admitted to the hospital.

Unfortunately, the results of the algorithm were problematic. The computer algorithm determined that patients with other respiratory ailments and co-morbidities — such as asthma, chronic obstructive pulmonary disease, or chest pain — had a lower probability of dying and that, therefore, these patients should not be admitted to the hospital. Reanalysis of the data revealed that arriving pneumonia patients with respiratory ailments do indeed have a lower probability of death, and the explanation was that these groups of patients sought medical care sooner and thus had a lower severity of pneumonia infection and thus a lower probability of dying. But this reasoning ignored the fact that such patients are also more vulnerable in terms of potential complications. This risk was not considered.

From this example, we can see that the problem was not the algorithm per se, but rather the *design* of the algorithm and the *specific question* that the algorithm was asking. While the probability of death would be a very reasonable question to consider for a health insurance company, it was perhaps the wrong question to ask in this context. What the hospital really wanted to know was the probability that the patient would develop complications, and "death within 30 days" was used as a proxy for that measure. However, that proxy was affected by things other than the variable of interest (i.e. developing complications), such as the patient's awareness of having co-morbidities. A better question to ask might be about the severity of the infection and, based on the severity and risk factors such as age and co-morbidities, the algorithm could then recommend which patient should be admitted to the hospital. The available patient data, such as the level of fever (temperature) and comorbid respiratory ailments, could have been used to predict the level of infection severity and risk of complications, but this was not done.

This example reveals not only the need for proper algorithm design, but the importance of including people with domain knowledge in the algorithm design process. Consultation with a range of people with experience in the context of interest — such as pulmonologists, emer-

RACIAL BIAS IN FACIAL ANALYSIS

Another well-known machine learning bias example, publicized by Joy Buolamwini [2017],[30] concerns the performance of facial analysis algorithms when applied to people of different skin colors. Facial analysis models for automatic gender classification created by IBM and Microsoft were shown to perform surprisingly poorly (accuracy < 40%) when tested on dark-skinned women. Because the machine vision features that such computer algorithms use to analyze a human face are dependent on the levels of pixel contrast and average brightness, such algorithms are critically dependent on skin color and lighting. If the facial analysis algorithm is not trained using dark-skinned faces, it is perhaps not surprising that the algorithm will perform very poorly when presented with a dark-skinned face. Fortunately, there have been improvements in some of the commercially available face recognition algorithms subsequent to the public release of the research results exposing bias.[31]

Example #2: Predicting Wound Infection from Photographs

Fletcher et al. (2019)[32] published a study showing how a computer algorithm could predict the infection of a

30. Joy Buolamwini. "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers" PhD diss., (MIT, 2017). http://hdl.handle.net/1721.1/114068

31. Inioluwa Deborah Raji and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." In proceedings of the *2019 AAAI/ACM Conference on AI, Ethics, and Society* (Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 2019): 429-435. https://doi.org/10.1145/3306618.3314244.

32. Richard Ribon Fletcher, Olasubomi Olubeko, Harsh Sonthalia, Fredrick Kateera, Theoneste Nkurunziza, Joanna L Ashby, Robert Riviello, and Bethany Hedt-Gauthier. "Application of Machine Learning to Prediction of Surgical Site Infection." In proceedings of *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC) (New York: IEEE, 2019): 2234-2237. https://doi.org/10.1109/EMBC.2019.8857942.

surgical wound very accurately using only a color photograph of the wound (Figure 15). This study involved approximately 500 patients from rural Rwanda who had given birth through Cesarean section. The photographs were taken by the community health workers 10 days post-surgery. Although the infection prediction algorithm performed extremely well when tested on Rwandan women, it would be unreasonable to expect the algorithm to perform equally well if tested with lighter-skinned women — from Ethiopia or Europe, for example. The use of this algorithm is acceptable if it is restricted to the domain on which it was trained, which was rural Rwanda; but it would be technically and ethically wrong to present this algorithm as a general solution for infection prediction in all racial groups.

This example not only reveals the specificity of machine learning algorithms, which are trained to reflect particular data sets, but also underscores the need for domain knowledge, transparency, and oversight to ensure these algorithms are applied appropriately. Just like the facial analysis algorithms studied by Buolamwini (see sidebar), color-based algorithms can be expected to perform poorly when applied to patients with a different skin color.

It is important to understand that tailoring an algorithm to a particular patient group is not in itself a problem. Problems occur when algorithms are oversold in terms of their capabilities, applied beyond the bounds of appropriate use, or employed without an understanding of how the algorithm arrives at its decisions.

Figure 14 - (left) Community health worker capturing an image of a surgical wound. Figure 15 - (right) The top row contains sample images from infected wounds, and bottom row are sample images.

With this background in mind, the chapter now turns to the detailed case study.

B. Case Study: Exploration of Bias in Health Diagnostic Data

Building on the previous discussion of applying machine learning in a global health context, this section illustrates some ways that a health diagnostic model can be examined for bias, and also demonstrates some of the inherent difficulties in working with health data.

Clinical Study Description

The data used for this example is in the domain of pulmonary disease and was collected as part of a diagnostic prediction study conducted by MIT and the Chest Research Foundation in Pune, India.

The purpose of the study was to create a simple diagnostic algorithm that could accurately detect the presence of three different pulmonary diseases — asthma, chronic obstructive pulmonary disease (COPD), and allergic rhinitis (AR) — as well as combinations of these diseases.

320 subjects, including healthy controls, were tested using a mobile phone diagnostic kit that included a simple questionnaire and a peak flow meter.

The total numbers of patients presenting with the pulmonary diseases and combinations were as follows:

Efforts were made to recruit equal numbers of women and men, resulting in 171 male and 132 female patients.

In order to provide labels and training data for the machine learning algorithm, every subject in the study was also administered a complete battery of pulmonary function tests, which included spirometry, body plethysmography, impulse oscillometry, and lung gas diffusion testing. Based upon these tests, an informed diagnosis was given by an experienced chest physician.

Applying the Fairness Methodology

To determine an algorithmic approach to fairness, the Fairness Methodology from Chapter 4 was applied (see page 25).

Is a demonstration of statistical parity required by law?

Beginning with the first question in the framework, demonstration of statistical parity was not required by law — in this case the goal was not to diagnose the same rates of illness across groups but rather to diagnose across groups with the same accuracy.

Is there a reasonable prospect of creating a causal model?

As the discussion will illustrate, smoking emerged as a likely causal factor, but because there were no women smokers, it was clear that smoking could not be the only cause. Because there remain myriad unknown influences on pulmonary health, the answer to this question was also no.

Is statistical parity compelling to stakeholders?

Again, it is not in this case, because the goal is not to diagnose the same number of cases across the different groups but rather to ensure that all groups are receiving accurate diagnoses.

Are protected attributes independent of other features?

In this case the question is whether gender and socio-economic status are functionally tied to other features in the data set. As the discussion below illustrates, these protected attributes are in fact tied to behavioral differences (smoking).

Does bias leads to unfairness through a negative outcome in addition to unfairness through a positive outcome?

In this case, the concern is both with false negatives (missed diagnoses) and false positive diagnoses, which would misallocate scarce medical resources. The methodology therefore calls for equalized odds.

In practice, this was achieved by separating the data sets according to the different gender and socio-economic status (SES) classes and creating independent models for men and women and for low and high SES groups.

Algorithm Development

A machine learning algorithm was created using logistic regression, which is highly interpretable. This type of machine learning model was chosen because the parameters generated by logistic regression enables the analyst to conduct coefficient analysis and subsequent bias analysis. (See the Appendix for a discussion of different algorithms and their benefits and drawbacks with respect to bias.)

A separate model was created for each disease (asthma, COPD, and AR). Because some patients had comorbid conditions (e.g. asthma+AR or COPD+AR), the training data for each model included only patients that had a single disease rather than a combination (e.g. Asthma only, COPD only, and AR only). This training methodology produced the highest accuracy.

For model development, approximately 75% of the data was used for training and 25% of the data was reserved for testing. The median area-under-the-curve (AUC) accuracy of the three models (COPD, asthma, AR) was

85%, 75%, and 95%, respectively. This AUC measure provides a summary of the discrimination ability of the model across the entire range of inputs.

Bias Analysis – General Considerations

For the purpose of bias analysis, the goal was to examine if a given algorithm would favor or penalize members with a certain protected attribute, such as race or gender. The question posed was:

Is there a significant difference in prediction accuracy between specific subgroups?

In order to explore this question, it is important to note some general observations about machine learning:

» The accuracy of an algorithm will generally depend on the size of the test set as well as the homogeneity and variance of the data in the test set.

» In general, the accuracy of a machine learning model will improve with an increasing amount of training data. However, there is a point of diminishing returns.

» The amount of training data needed to achieve the optimum level of accuracy also depends on the quality of the training data. If the quality of training data is poor, and the data contains a great deal of noise or random variability, the inferences that can be drawn from the data will also be weak, so a greater amount of data may be required. In general, quality is more important than quantity.

In terms of the mechanics of machine learning, it is also important to note that because there is some variability in health data, it is common practice to run many iterations of the same model, each time changing the patients that belong to the test set and the training set. After many iterations, the median values are incorporated into the model. With logistic regression, the median value of the coefficients may be used. The number of iterations required depends on the amount of variability in the model.

In this case of pulmonary diseases in India, running 1000 iterations of the model produced a sufficiently low

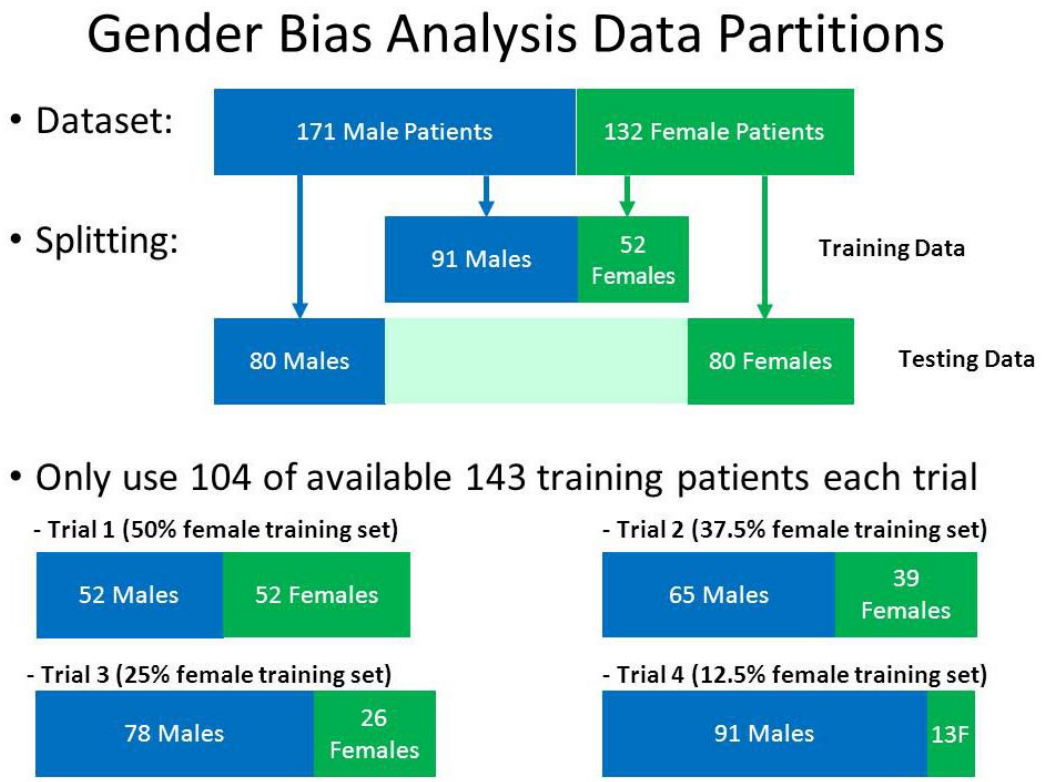


Figure 16 - Data partitions used for gender bias analysis. The size of the test set and the size of the training set were kept constant, but the proportion of males and females was varied in the training set.

variance that enabled comparison across models. This process illustrates that even though algorithms such as logistic regression are deterministic (determined by the parameter values and initial conditions), there is a stochastic (random) component to model development, because the members of the training set and test set

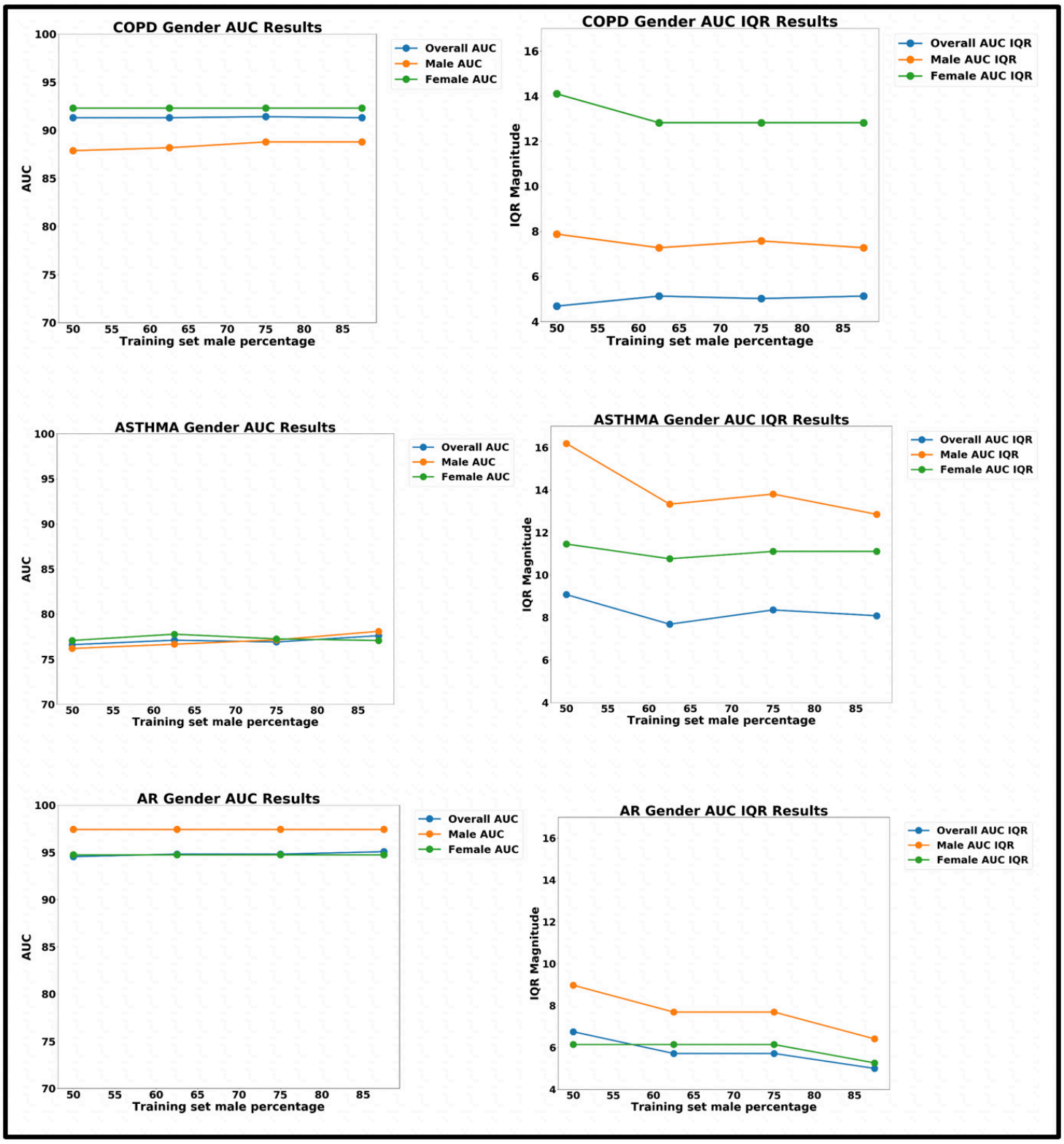


Figure 17 - Plots of AUC Accuracy (left) and InterQuartile range (right) results of gender bias analysis for three different disease diagnostic models: (top) COPD; (middle) Asthma; and (bottom) Allergic Rhinitis (AR). Results are shown for different proportions of males vs females in the training set. The horizontal axis represents the proportion of males in the training set.

are selected at random. Thus, it is important to measure variations in the model accuracy across different iterations of the model.

For smaller data sets (N< 250), the method of leave-one-out cross-validation is sometimes used, which leaves out only one data point from the training set and uses it for prediction. However, for larger data sets, it is preferable to separate the data into training and test data and use a held-out test set for analysis that is never mixed with the training set.

Gender Bias Analysis – Methodology

In order to explore possible gender bias in the algorithm, the algorithm accuracy was separately tested on male and female patients and the results compared. In order to examine how sensitive the algorithm was to differences in gender, different models were created that were trained on different proportions of male and female patients.

In order to conduct this analysis, a pool of 160 patients were defined to be used as the test set. The size of this test set was kept fixed and was equally divided between male (N=80) and female (N=80) patients. The remaining 143 patients (91 males, 52 females) were used as a general pool from which to select training data. The data partitioning used for this bias analysis is illustrated graphically in Figure 16.

In order to see how the algorithm depends on gender, four different data sets were defined, each with a size of N=104, but with each having a different proportion of males and females: 50% female, 37.5% female, 25% female, and 12.5% female. For each iteration of the model, a different set of patients would be randomly selected to be part of the N=104 test set. One thousand iterations of the model were calculated.

To maintain consistent results, a held-out test set was used: the patients in the test set were isolated from the rest of the data throughout all our analysis and were never mixed with members of the training set.

Gender Bias Analysis – Results

The results of the gender bias analysis are shown in Figure 17 for each of the three pulmonary diseases studied (asthma, COPD, and AR). The accuracy as defined by the ROC AUC is given as well as the variation expressed as Interquartile Range (IQR), a pair of values for the parameter covering from the 25th percentile to the 75th percentile of the observed distribution.

From the plots, it is clear that the accuracy of all three models does not change significantly as a function of the proportion of women in the training set. This indicates that sampling bias is not a significant concern for these diseases.

However, the results do show that there is a systematic diagnostic gender bias between males and females for COPD, and a small systematic bias for AR. For asthma the model performs equally well for male and female patients.

The plots of IQR reveal that there is significant variability in the COPD and Asthma patients, with the female patients having the highest variability for COPD and the male patients having the highest variability for asthma. For AR, there was low variability for both males and females.

Gender Bias Analysis – Discussion

Although the machine learning model exhibited minimal gender bias for asthma and AR, a significant diagnostic bias is noted for COPD. Because this bias persists even when the training data is equally divided among male and female patients, it is clear that this disparity is not due to sampling bias in the training data.

What, then, is the cause of this bias?

In order to examine this question further, various risk factors for COPD were explored to examine which factors may be dependent on gender. It is well-known that one of the greatest risk factors for COPD is smoking cigarettes (a cause of emphysema that contributes to COPD). This observation was also confirmed by performing a coefficient analysis on the logistic regression model and seeing the coefficient value of the smoking variable.

By examining the proportion of male and female patients in the study that smoke cigarettes, a large difference was observed between genders. As shown in Figure 18, all of the smokers are male and none of the female patients are smokers. From this observation, it can be hypothesized that the gender bias in the algorithm is due to the large disparity in the smoking status between men and women. Other features did not show any significant gender disparity.

The smoking data also help explain why the model accuracy is higher for women compared to men. Because none of the female patients smoke cigarettes, there is less variance in the female patient population, and thus the model is better able to predict COPD and achieve a higher accuracy. However, among the male patients, approximately 45% of the male patients smoke and 55% do not, which creates significant variability in the results. Thus, the accuracy for male patients is lower.

Based on this analysis, one can also speculate that it may be possible to achieve higher accuracies in the COPD model if the data were stratified by smoking status. For example, it would be possible to create a separate model for smokers and for non-smokers. The resulting models should not only have higher accuracies but also exhibit less bias across genders.

Socioeconomic Status Bias Analysis – Methodology

In order to explore potential socio-economic bias, a similar methodology was used to partition the data and isolate socioeconomic status (SES) as a variable.

Similar to the gender bias example, a pool of 58 patients was defined to be used as the test set. This test set was equally divided between male (N=29) and female (N=29) patients. The remaining 245 patients – 175 high income (high SES), 70 low income (low SES) – were used as a general pool from which to select training data.

In order to see how the algorithm depends on SES, four different data sets were defined, each with a size of N=140, but with each having a different proportion of low SES and high SES: 50% low SES, 37.5% low SES, 25% low SES, and 12.5% low- SES. For each iteration of the model, a different set of patients would be randomly selected to be part of the N=140 test set. One thousand iterations of models were computed and the median was taken.

In order to maintain consistent results, a held-out test set was used, keeping patients in the test set the same throughout all our analysis.

The data partitioning used for the bias analysis is illustrated graphically in Figure 19.

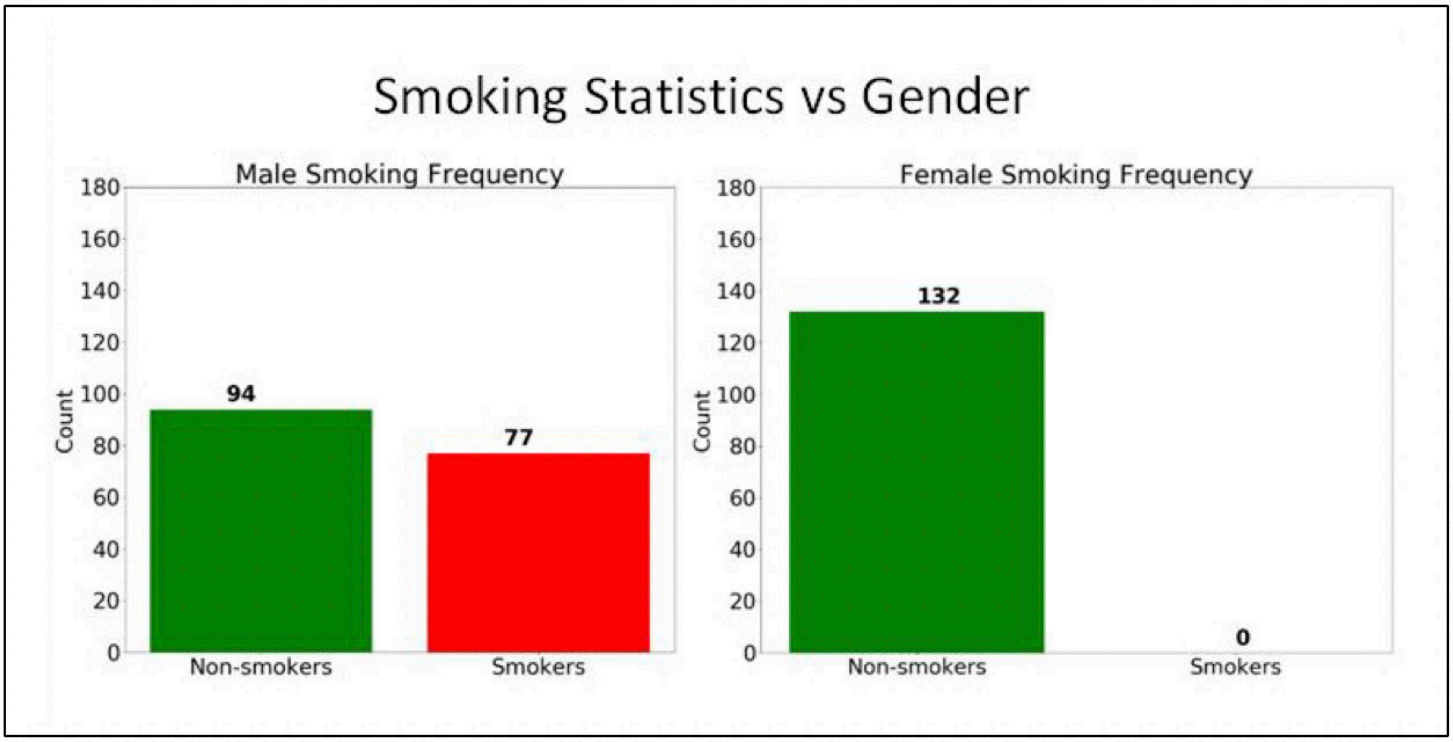


Figure 18 - The number of smokers and non-smokers in each gender group.

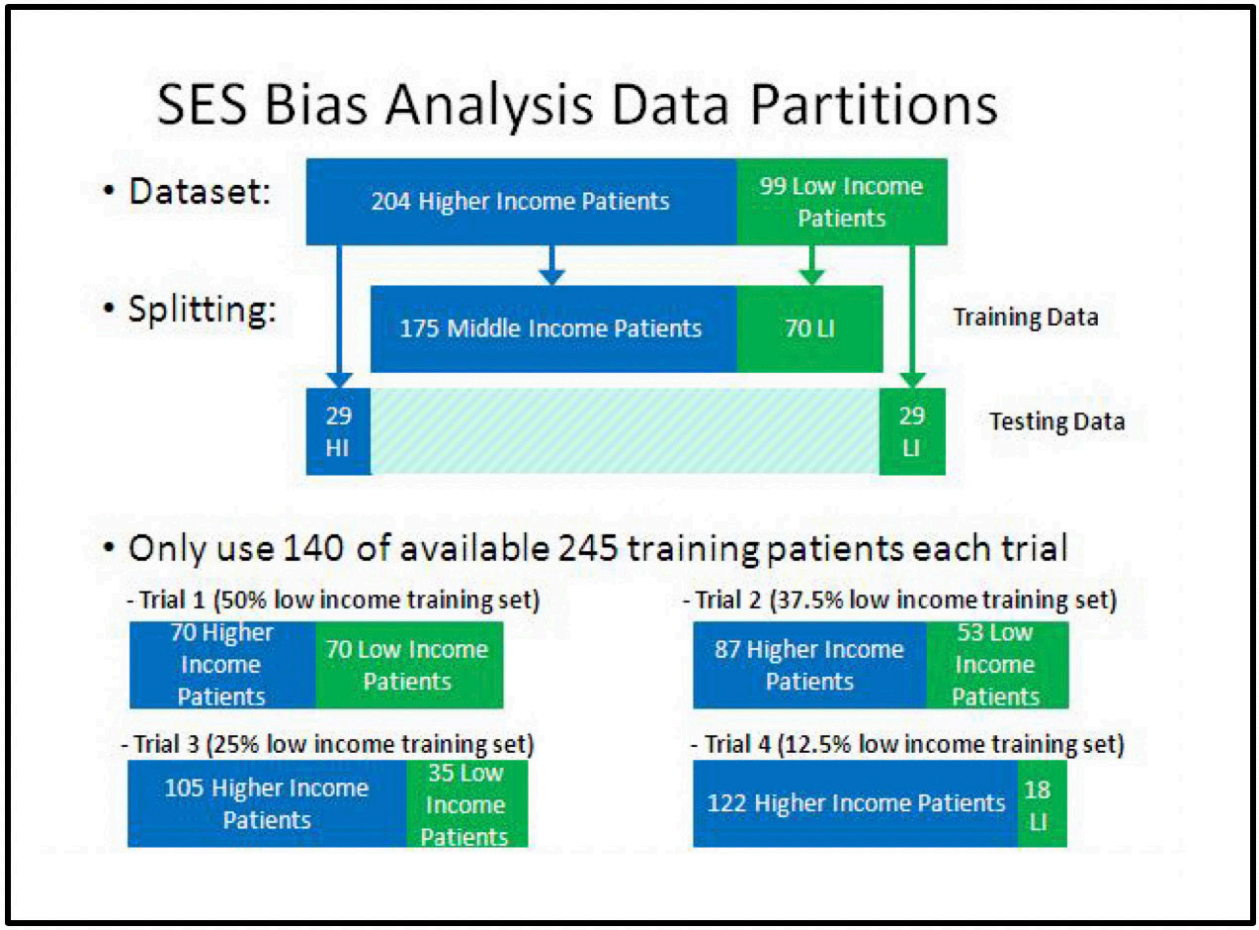


Figure 19 - Data partitions used for gender bias analysis. The size of the test set and the size of the training set were kept constant, but the proportion of males and females was varied in the training set.

Socio-Economic Status (SES) Bias Analysis – Results

The results of the SES bias analysis are shown in Figure 20 for each of the three pulmonary diseases studied (asthma, COPD, and AR). The figure shows the accuracy, in terms of ROC AUC, as well as the variation in the accuracy expressed as the Interquartile Range (IQR).[33]

For all three disease models, the accuracy remains fairly consistent as the proportion of low-SES patients is varied. In the variability analysis, however, the IQR value for allergic rhinitis (AR) increases significantly as the proportion of low-SES patients is reduced.

Socio-Economic Status (SES) Bias Analysis – Discussion

Based on these results, it is clear that for COPD and Asthma, there is little sampling bias present in these models, in terms of SES. In other words, the proportion of low-SES patients in the training data has little effect on the accuracy and variability of the model.

In the model for AR, however, the accuracy of the model degrades significantly as the proportion of low-SES patients is reduced, which indicates that there are some SES-dependent features that contribute to the model. In this case, the model for AR does appear to be very sensitive to sampling bias. Unless low-SES patients are included in the training data, the variability of the AR model will be unacceptably high for low-SES patients. There is little systematic diagnostic bias between low-SES and high-SES patients, so the same model can be used for both groups; however, in order to maintain low variability in the performance, it is important to include equal proportions of low-SES and high-SES patients in the training data.

33. Variance is not generally used because this error does not have a normal (Gaussian) distribution.

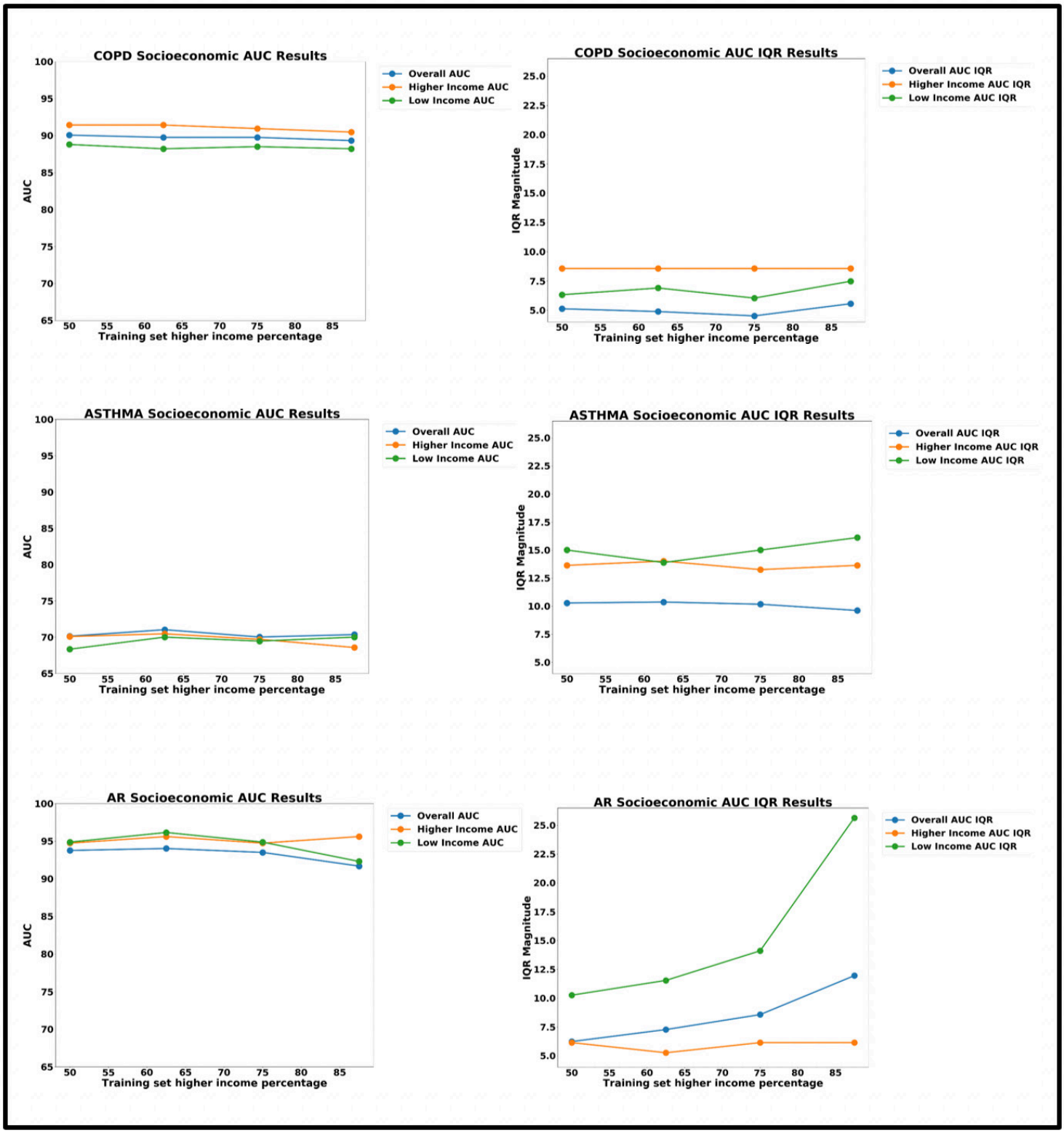


Figure 20 - Plots of AUC Accuracy (left) and InterQuartile Range (right) results of Socio-economic (SES) bias analysis for three different disease diagnostic models: (top) COPD; (middle) Asthma; and (bottom) Allergic Rhinitis (AR). Results are shown for different proportions of high vs low SES in the training set. The horizontal axis represents the proportion of high SES patients in the training set.

In terms of systematic diagnostic bias, there is little disparity in diagnostic accuracy between the high-SES and low-SES groups. However, for COPD, some small amount of disparity can be observed. In order to explore this further, we can once again examine which features in the model may have a disparity across SES groups.

As with the gender bias analysis, it is clear from the logistic regression coefficient analysis that cigarette smoking is a feature with high predictive value. Figure 21 shows the number of high-SES and low-SES patients that smoke cigarettes. From these data, it is clear that the high-SES group comprises predominantly non-smokers, whereas the low-SES group is roughly evenly divided between smokers and non-smokers. Based on this observation, it can be hypothesized that the small disparity in the accuracy between high-SES and low-SES patients is primarily due to the disparity in the smoking prevalence among these patients' groups.

As with the case of the female patients in the gender bias analysis, the COPD model produces a higher accuracy among high-SES patients, mostly likely because these patients are more homogeneous.

Summary

In addition to demonstrating how a bias analysis may be conducted in a real-world machine learning application, this case study presents and highlights some important considerations that should be examined whenever machine learning is being applied to health care. Key considerations include:

- Health and disease are complex, analog processes with many risk factors, often including hidden variables. Machine learning analysis generally requires creating crude approximations to these processes, which must be done carefully and with the proper domain expertise, in order to avoid introducing errors and false conclusions.
- The domain of health can reflect genetic, environmental, and behavioral differences across different genders, racial or ethnic groups, and socio-economic classes that affect disease prevalence and present additional challenges for ensuring fairness. If it is revealed that a particular algorithm consistently produces very different results for one patient group vs another, it is generally best to design a separate algorithm for each group rather than try to create a universal algorithm that will very likely perform poorly on both groups. Alternatively, a more complex and flexible model (such as random forests) might have more ability to "act like" different models in different situations, at the cost of requiring more training data.

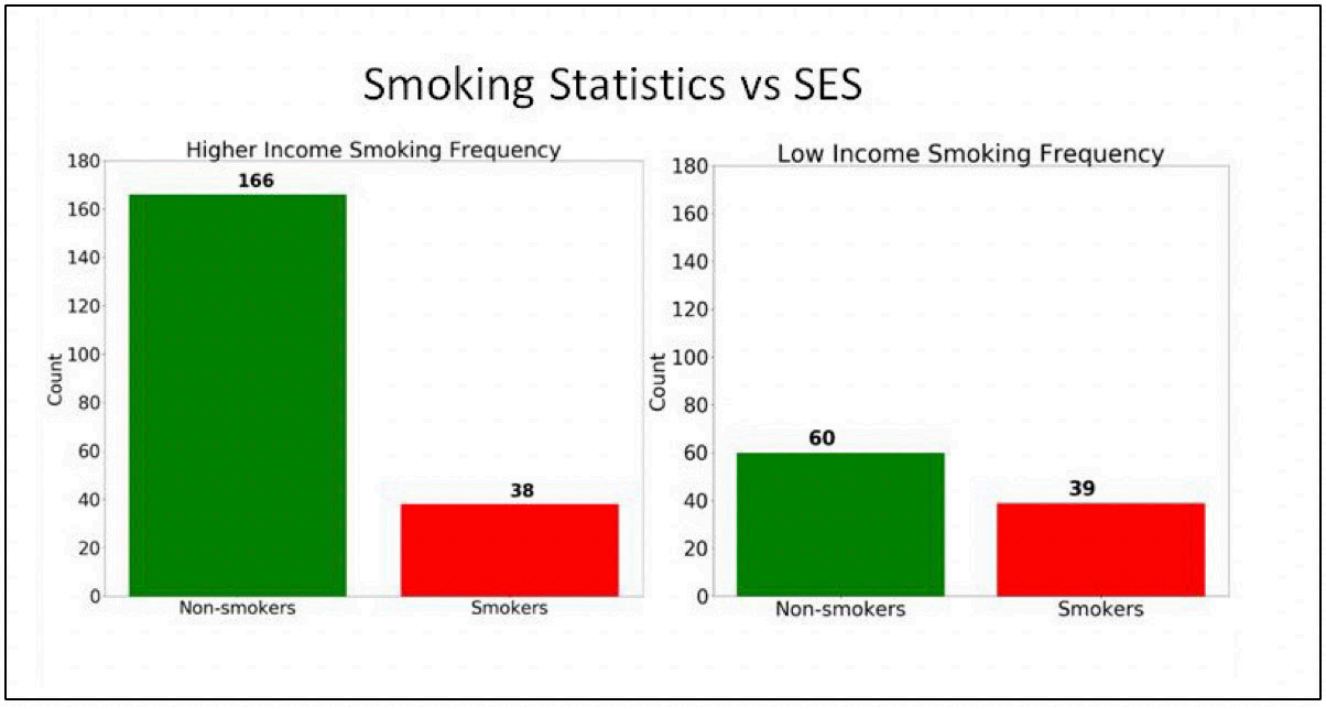


Figure 21 - The number of smokers and non-smokers in each gender group.

Appendix: Fairness and Bias Considerations for Specific ML techniques

For readers interested in more detailed computer-science based exploration of fairness and bias, this appendix serves as an overview of specific ML techniques highlighting fairness and bias considerations for each approach.

TECHNIQUES DISCUSSED IN THIS APPENDIX

- Regression Analysis
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K Nearest Neighbors (k-NN)
- Receiver Operating Characteristic (ROC) curves
- K Means Clustering
- Hierarchical clustering
- Density Based Clustering (DBSCAN)
- Support Vector Machines
- Classification and Regression Trees (CART)
- Naïve Bayes Classifiers
- Random Forests
- Artificial Neural Networks (ANN)

Regression Analysis

Fairness and bias considerations: There is significant risk of bias and unfairness when regression analysis is used in the analysis of socially relevant data. Although the formulae provided in this section provide unique solutions for the estimators given a particular form of model, the choice of a model has a first order impact on the conclusion drawn in a study. The analyst must exercise vigilance to ensure that the variables included in a regression analysis, the variables excluded from a regression analysis, and the form of the model are not reinforcing some preconception of the phenomena. Methods like "best subsets" regression can be used as a countermeasure to make results more dependent on explicitly selected statistical criteria and less dependent on unstated preferences of the data analyst.

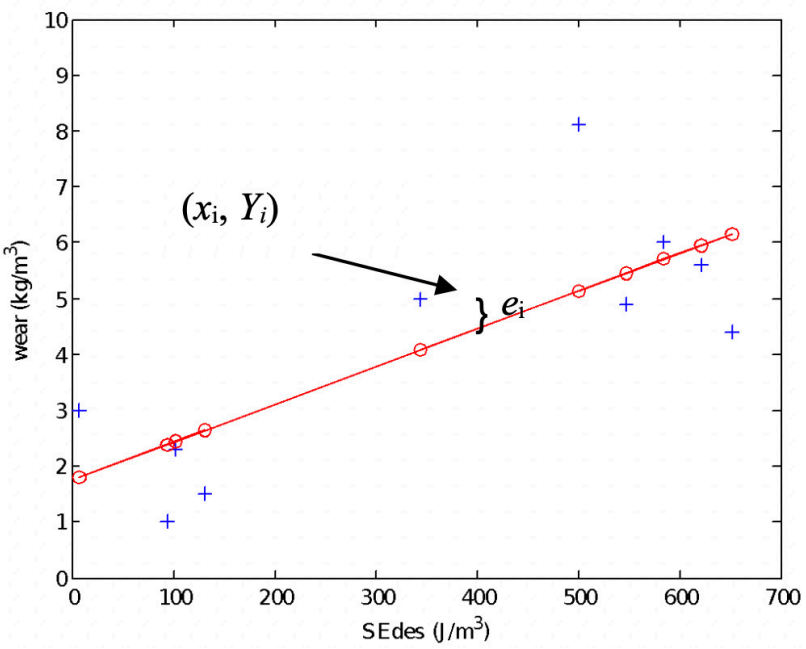
Overview of technique: Regression analysis is a statistical technique for describing and exploring data. Its purpose is to model the effect of continuous valued independent variables on a continuous valued dependent variable. While regression analysis is not itself a Machine Learning algorithm, it is an essential ingredient in many Machine Learning procedures, so it is useful to review it here.

To begin this introduction, it is most convenient to start with "simple linear regression" in which we have just one independent variable and we seek to examine a linear relationship with the dependent variable. To formalize the concept, we use a regression equation

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$$

where Y is the dependent variable and x is the independent variable. In this equation, and are the slope and intercept of a proposed line that is meant to approximate the relationship between x and Y. The term e represents the error or the deviation between the proposed line and each data point.

Figure A1



The Figure A1 illustrates an example of simple linear regression. The blue '+' symbols represent data. The values plotted on the abscissa (aka x-axis) are the values of an independent variable. The values plotted on the ordinate (aka y-axis) are the values of the dependent variable. The red line is the regression model. The red circles represent predictions of the regression model and the vertical distance from each blue '+' symbol and each red circle represents error.

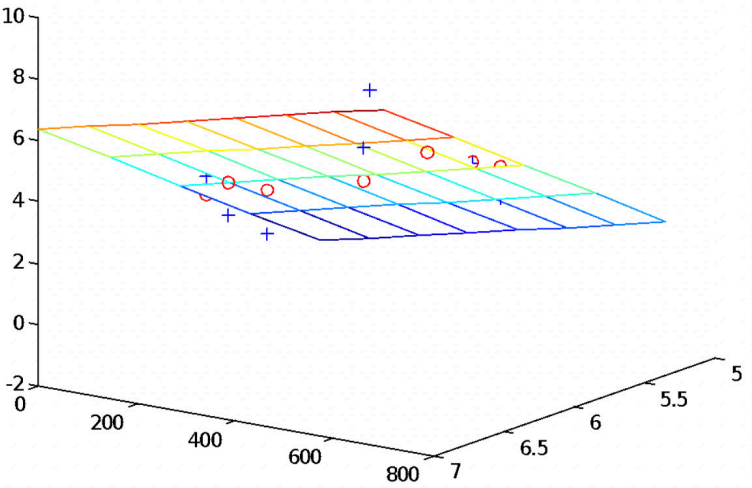
A key fact to understand about simple linear regression is that there is a unique solution that minimizes the sum squared error. In simple linear regression, the solution is a pair of values $\hat{\beta}_0$ and $\hat{\beta}_1$ and this pair of values makes the following sum as small as possible: $\sum_{i=1}^{n} \varepsilon_i^2$

The concept of "simple" linear regression with one independent variable extends naturally to multiple independent variables. The same procedure is sometimes called "multiple regression" to emphasize the contrast with "simple" linear regression. The linear regression equation is:

$$Y = X\hat{\beta} + \varepsilon$$

The sort of analysis using this equation as a model is usually called just "linear regression" because the analysis still assumes a linear superposition of multiple effects. As a result, the system of equations is linear in the model parameters b. When this linear form is used, we can conceptualize the regression model as a plane that is optimally fit to the data. If there are two independent variables, the plane is visualized as a three-dimensional space as shown in Figure A2.

Figure A2



If there are more than two independent variables, then we can conceptualize the model as a hyperplane in four or more dimensions, but this is beyond the capacity of most people to visualize, so we rely on the mechanisms of matrix algebra to manage the complexity of the operations. The equation is in a very compact matrix form and, to avoid confusion with scalar equations, it is useful to unpack the notation. The data in this model is arranged into structures (Figure A3).

where $\hat{\beta}$ is the vector containing the model parameters on the data in X and Y. The "hat" over the symbol is placed there to emphasize that what we computed is an estimate. In practice, for large systems of equations, the matrix inverse would not be computed but rather an alternative algorithm such as Gaussian Elimination or QR decomposition would be employed to make the process more efficient and numerically stable.

$$Y = X\hat{\beta} + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & | & | & & | \\ 1 & | & | & & | \\ \vdots & | & | & \cdots & | \\ 1 & | & | & & | \\ \text{vector of data \#1} & | & \text{vector of data \# 2} & & \text{vector of data \# } k \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Figure A3

If we include a constant term in the regression then there is conventionally a set of 1's in the left-hand column of the model matrix X. In statistics, the constant term is denoted $\hat{\beta}_0$ and the other coefficients go up from there $\hat{\beta}_1$, $\hat{\beta}_2$ and so on up to $\hat{\beta}_k$. The length of vectors *Y* and e is n which is the number of data points. The length of $\hat{\beta}$ is p=k+1 where *p* is the number of "predictors" and *k* is the number of "independent variables" in the regression equation. Consistency of the matrix equation requires a "model matrix" *X* that is *n* by *p=k+1*. You might say one of the "predictors" is a degenerate sort related to predicting the value of the dependent variable given all the independent variables take the value of zero.

There is a closed-form solution to the problem of minimizing sum squared error in solution of a system of linear equations. A least-squares fit is provided by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Principal Component Analysis (PCA)

Fairness and Bias considerations: The results of Principal Components Analysis (PCA) are combinations of factors that enable an outcome to be explained with a small number of combinations rather than a long list of separate factors. PCA is now a common procedure used in where the vector containing the model parameters on the data in *X* and *Y*. The "hat" over the symbol is placed there to emphasize that what we computed is an estimate. In practice, for large systems of equations, the matrix inverse would not be computed but rather an alternative algorithm such as Gaussian Elimination or QR decomposition would be employed to make the process more efficient and numerically stable. early stages of machine learning projects to accomplish "dimensionality reduction." It has been found that PCA sometimes exhibits different reconstruction error rates when applied to different sub-populations. Samadi et al (2018)[34] propose a procedure for conducting PCA that equalizes error rates of PCA across

34. Samadi, Samira, Uthaipon Tao Tantipongpipat, Jamie H. Morgenstern, Mohit Singh and Santosh S. Vempala. "The Price of Fair PCA: One Extra Dimension." In proceedings of 32nd International Conference on Neural Information Processing Systems 32, Montreal, 2018. https://arxiv.org/pdf/1811.00103.pdf

the relevant populations. In addition, the results of PCA are a function of not only the data but also of some implementation decisions made by the analyst. A PCA using the covariance matrix directly can depend on the units in which the analyst expresses the variables. Normalization of data is often used to address this situation and there are a number of judgements to make such as whether to normalize by the standard deviation or the range. For this reason, analysts using PCA (and also consumers of the analysis) should be on guard for ways that implementation decisions affect the outcomes.

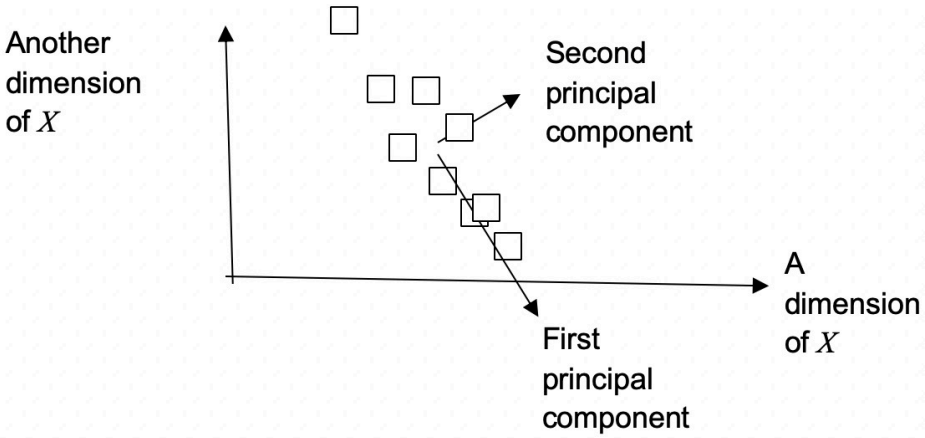


Figure A4

Overview of technique: Principal Component Analysis (PCA) is a statistical technique for data analysis and dimensionality reduction. PCA identifies linear combinations of independent variables that explain the maximum amount of variability in the data with as few linear combinations as possible.[35, 36] PCA employs orthogonal transformations to convert a set of data into a set of values of linearly uncorrelated variables called principal components. While PCA is not itself a machine learning algorithm, it is an essential ingredient in many machine learning procedures.

If our data set is collected into a matrix *X*, then we may define its sample covariance matrix as *K*. The first principal component will be the eigenvector of *K* corresponding to the largest eigenvalue. Similarly, the second principal component will be the eigenvector of *K* corresponding to the second largest eigenvalue. In practice, it is often the case that data can be described using a model that includes a subset of the principal components and that such models still explain the majority of the variance in the data set.

Linear Discriminant Analysis (LDA)

Fairness and bias considerations: Linear Discriminant Analysis (LDA) classifies objects into different groups based on multiple measurements. The performance of LDA is affected strongly by the degree to which the required assumptions hold. LDA assumes similar patterns of variation and correlation among the variables for both classes of objects being classified. In classification between just two alternatives, if this assumption were violated, then, subsequent to training of the classifier, the group exhibiting higher variability would experience higher rates of misclassification. This is a subtle but potentially significant mechanism by which bias and unfairness can find its way into an LDA classifier.

Overview of technique: Linear Discriminant Analysis (LDA) is a statistical technique for data analysis and parametric categorization. LDA is closely related to regression analysis as it forms a model of a dependent variable as a linear combination of independent variables in a data set.[37] However, discriminant analysis has continuous independent variables and a categorical dependent variable (that is, the class label). LDA can be understood most easily for the case of discrimination into just two classes. LDA defines a hyperplane in the space of the vectors of independent variables *x*. Any vectors lying on one side of

35. Karl Pearson. "On Lines and Planes of Closest Fit to Systems of Points in Space". *Philosophical Magazine* 2, no 11 (1901): 559–572, 498–520.in Space". Philosophical Magazine. 2 (11): 559–572., and 498–520.

36. Harold Hotelling. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24 (1933): 417–441.

37. Geoffrey Mclachlan. *Discriminant Analysis and Statistical Pattern Recognition*. (Boston: Wiley, 1992).

the plane are categorized as being in one class and any vectors lying on the other side of the hyperplane are categorized as being in the other class. The hyperplane can be understood as a plane which contains a point at the midpoint of the line between the mean of the two classes and is perpendicular to the inverse of the covariance matrix times the vector difference of the mean of the two classes. In the case of multiple classes, the boundaries are the convex subsets of the hyperplanes that connect at the intersections among them. The LDA approach can be considered as a Bayesian optimal classification when the independent variables are normally distributed and the classes all have the same covariance matrix.

Quadratic Discriminant Analysis

Fairness and bias considerations: Quadratic Discriminant Analysis (QDA), like LDA, classifies objects into different groups based on multiple measurements. The technique allows for more flexibility as compared to LDA so that some statistical assumptions can be relaxed. On the other hand, providing more flexibility to the data analyst can enable overfitting. When a model has a larger number of parameters, a wider range of conclusions can be supported. It may be possible for data analysts to search for models that fit their preconceptions. This confirmation bias could lead to unfairness when a QDA classifier is employed.

Overview of technique: Quadratic Discriminant Analysis (QDA) is a statistical technique for data analysis and parametric categorization. QDA is closely related to LDA in that they both assume normal distribution of the independent variables used for classification. However, unlike LDA, QDA does not require the assumption that the covariance of each of the classes is identical. In this case, the optimal rule for deciding membership between two classes involves a likelihood ratio test. In the two-class case, the results of this test can be interpreted geometrically as a quadratic surface in the space of independent variables.

k Nearest Neighbors (k-NN):

Fairness and bias considerations: A drawback of "majority voting" classification schemes like k-NN can be observed when the class distribution is not uniform. The class that is more frequently represented in the training set tends to dominate the prediction of the new examples. This tendency to classify new entities more often in the better-represented class is a potential way that unfair outcomes can arise from the analysis.

Overview of technique: The k Nearest Neighbors algorithm is among the simplest non-parametric categorization methods. The k-NN algorithm requires a set of training data that are labeled entities with multidimensional features. In the classification phase, the training data are

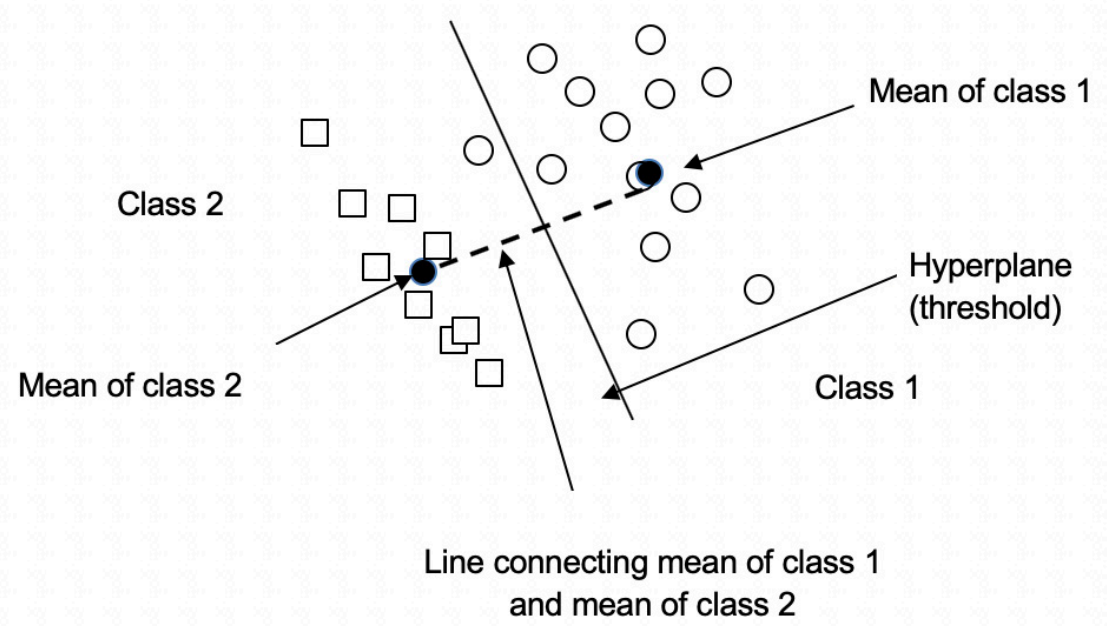


Figure A5

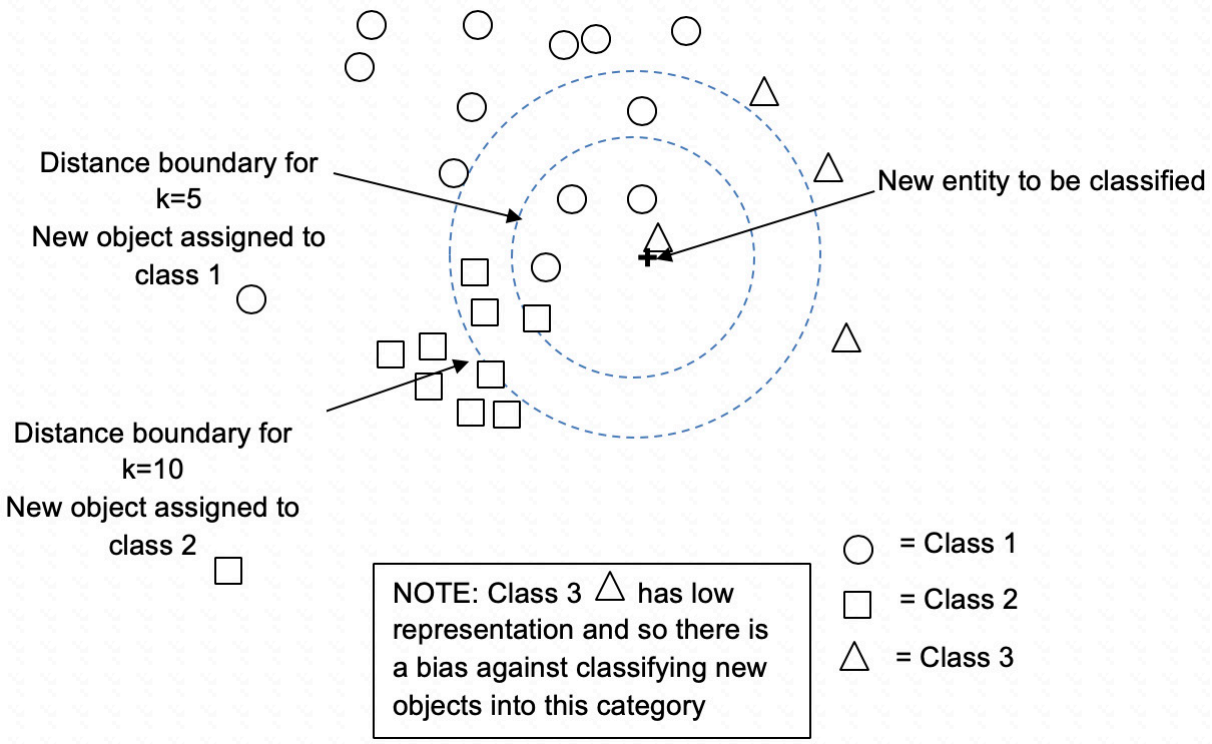


Figure A6

used by computing the distances to all the elements of the training set. When k-NN is used for classification, the k nearest neighbors then "vote" on membership of the entity. Alternately, the k-NN algorithm can be used to assign a continuous value (rather than a class label), in which case a regression procedure is applied to the nearest neighbors' values. The class chosen to assign is whichever class has the largest number of the k nearest neighbors. A common choice of the distance metric is the Euclidean distance; however, given that choice, the classification scheme can be very sensitive to the scaling of the axes or the units in which the features are measured. Alternative choices of the distance are sometimes formulated to account for the covariance structure in the feature variables. The parameter k must be chosen by the analyst/programmer.

Receiver Operating Characteristic (ROC) curves

Fairness and bias considerations: A Receiver Operating Characteristic (ROC) curve is a widely used tool for assessing the performance of ML techniques. The ROC curve makes it clear how tradeoffs between different types of mistakes are necessary in any realistic application. For example, in judging a person's credit, we can make the mistake of giving a loan to a person who is at high risk of not paying it back and we can sometimes make the mistake of refusing a loan to a person who would have paid it back. By enabling a visualization of such trade-offs, ROC curves have become a valued tool. However, the usual ways that ROC curves are used will tend to place emphasis on accuracy of an ML technique rather than on equitable treatment of protected groups.

Overview of technique: Having just reviewed several classification algorithms it is useful to describe a common tool for evaluating their performance – a Receiver Operating Characteristic (aka ROC) curve. The terminology of ROC curves is related to its historical origins in radar systems. The receiver (the "R" in ROC) collects electromagnetic radiation reflected from targets. When the receiver detects an actual target, that is referred to as a true positive. When the receiver fails to detect an actual target, that is referred to as a false negative. When the receiver signals detection of a target when there is actually no target present, that is referred to as a false positive. On an ROC curve, the abscissa plots the false positive rate, which can usually be adjusted using a parameter

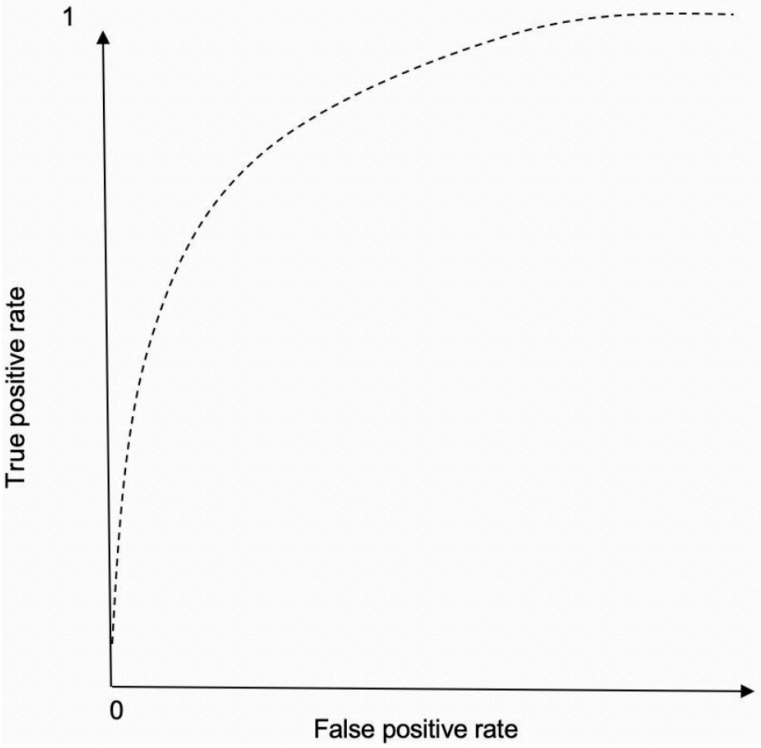


Figure A7

of the receiver such as a detection threshold. The ordinate of an ROC curve plots the true positive rate.

The corner values of an ROC curve characterize some extreme settings that are never actually used. The upper right corner is at 1,1 because the detection threshold of any receiver could be set so that it alarms constantly. It would never fail to detect a target, but it would also fail to discriminate at all. The lower left corner is at 0,0 because the detection threshold of any receiver could be set so that it never alarms. It would never detect a target, but it would also never create false positives.

In some cases, it is useful to have a single number that summarizes the discrimination ability of a receiver across the whole range of detection thresholds. For that purpose, many people compute the Area Under the Curve (AUC). A receiver with no discrimination capability at all would have an ROC that is a straight line from corner to corner and it would have an AUC of 0.5. An ideal receiver could have the true positive rate rise very quickly with the detection threshold and it could approach an AUC of 1.0.

k Means Clustering

Fairness and bias considerations: A significant challenge in k means clustering is that the results may depend strongly on how the clusters are initialized. This creates an opportunity for the results to be affected by the prejudices of the analyst. When the clusters reinforce the previously held views of the people reviewing the data, the results tend to be accepted. When the clusters contradict the previously held views of the people reviewing the data, the analysis can simply be repeated with a different initial set, with a different value of the hyperparameter k, or both. Through such an iterative process, a confirmation bias can emerge.

Overview of technique: The k means clustering algorithm is a cousin of the k-NN classification algorithm. Its purpose is to partition a set of data (numbering more than k) into a modest number (k) of classes. Unlike k-NN, no labeled training set is needed. The parameter k must be chosen by the analyst/programmer. The method requires some form of initialization of the clusters — for example, choosing k members of the set at random. After initialization, refinement of the clusters proceeds by computing the mean values of all the clusters in space of the feature vectors, adding one new observation by assigning it to the cluster with the smallest distance to the mean of that cluster, and repeating. This results in a partitioning of the feature space into k convex regions with linear boundar-

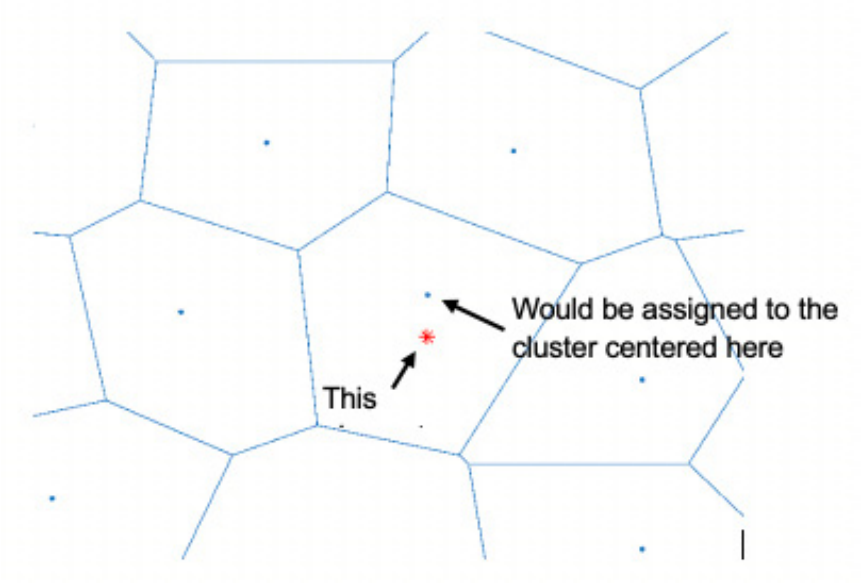


Figure A8

Hierarchical Clustering

Fairness and bias considerations: A major advantage of hierarchical clustering is that the results tend to be interpretable by humans, which is a useful hedge against bias and unfairness.

Overview of technique: A set of methods in Machine Learning in which new clusters are formed from existing clusters to form a hierarchy. The evolution of clusters can be accomplished either by agglomeration or by subdivision of existing clusters. The resulting groups of clusters can be shaped in complex patterns. The relationships among clusters in the hierarchy are frequently represented by a dendrograph (see below). A major disadvantage is that the standard algorithm for hierarchical agglomerative clustering has a time complexity of $O(n^3)$ and requires $O(n^2)$ memory.

Density Based Clustering (DBSCAN)

Fairness and bias considerations: Practitioners sometimes train classifiers like DBSCAN in the presence of fairness objectives or constraints (e.g., demographic parity). However, there are concerns that ML systems trained with a fairness constraint may not generalize well. After the training is complete and the classifier is used on a new set of data, sometimes the fairness guarantees are no longer provided. It is an ongoing area of research to improve the generalization performance of classifiers in the presence of fairness constraints.[38]

Overview of technique: A non-parametric clustering method, DBSCAN is one of the most widely used algorithms in Machine Learning. Clusters are defined as areas of higher density in comparison with the rest of the dataset. Objects that do not belong to the high-density clusters are identified as outliers. A major advantage of DBSCAN are its time and memory complexity. Under reasonable conditions, an overall average runtime complexity of O(n log n) is observed and memory requirements of O(n) can be attained.

Support Vector Machines

Fairness and bias considerations: Support vector machines have been a central tool in research to promote fair classification. Zafar et al. have proposed a novel measure of decision boundary fairness which they call "decision boundary covariance." The authors launching this new approach employed SVMs in their earliest implementations because of the advantage SVMs afford due to their simple boundary structure.[39]

Overview of technique: Support vector machines (SVMs) are algorithms frequently used for classification. They

38. Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. "Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints", *csLG* (June 2018): 1-27. arXiv:1807.00028

39. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi, "Fairness constraints: Mechanisms for fair classification." In *20th International Conference on Artificial Intelligence and Statistics PMLR* 54 (2017):962-970. http://proceedings.mlr.press/v54/zafar17a.html

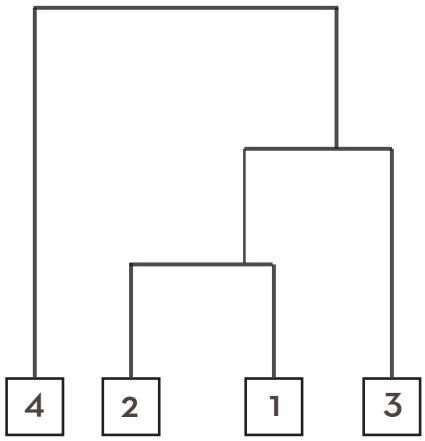


Figure A9

bear some similarity to k-means clustering in that they use hyperplanes to define class membership. Unlike k-means clustering, SVMs require labeled training data. Another difference with k-means clustering is that the hyperplane is not equidistant from the means of the clusters but instead is located to provide the largest possible gap between pairs of clusters (they seek a maximum-margin hyperplane). Solving for this optimal hyperplane can be done by gradient descent methods and, for large data sets, usually involves a technique called the "kernel trick." Many implementations of the "kernel trick" are closely related to Principal Components Analysis (PCA). An advantage of SVM's is that they provide good accuracy even with a small set of training data. A disadvantage of SVMs is that their performance is often sensitive to outliers in the training set. Many researchers have proposed more robust variants of SVM procedures so that the classification based on contaminated data can still provide information similar to that based on uncontaminated data.

Classification and Regression Tree (CART)

Fairness and bias considerations: Classification and Regression Trees offer some compelling advantages that enable robust and transparent decisions. When the number of cues and branches are kept low, decisions are easier to understand and yet the accuracy of the resulting decisions can still be quite high.

Overview of technique: A classification and regression tree (CART) analysis is a form of supervised learning. Regression analysis is applied to training data and the results are used to form a decision tree. Observations about an item are used along with the decision tree in order to draw conclusions about the item's class or else the probability of membership in a class. One of the most famous examples of CART was its application to classifying heart attack patients into two groups.[40] The low-risk group was defined as those who will survive 30 days. The high-risk group was defined as those who will not survive 30 days. After examining 19 variables, including age and blood pressure, the classification tree (Figure A11) was produced.

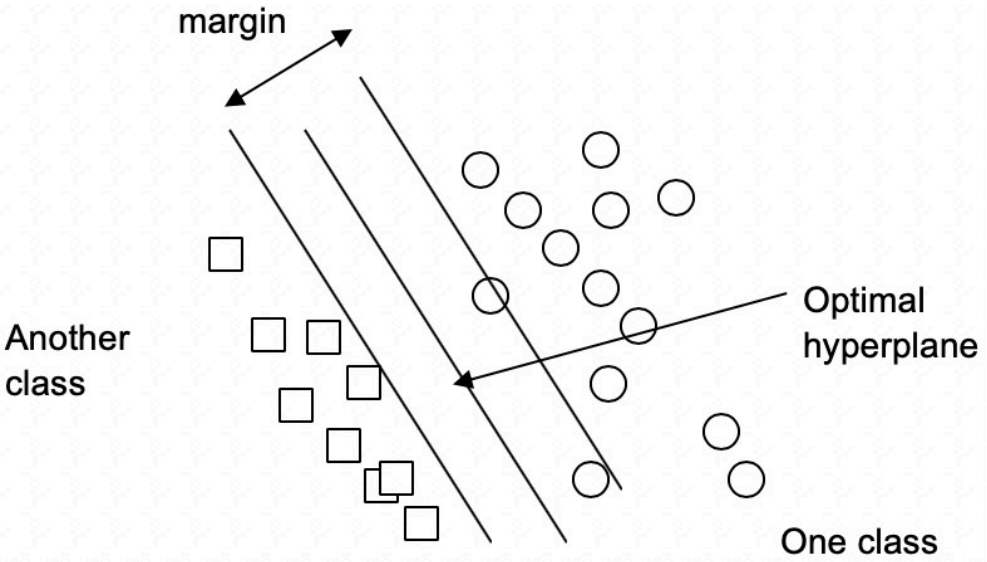


Figure A10

40. Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. (Monterey: Wadsworth & Brooks/Cole Advanced Books & Software, 1984).

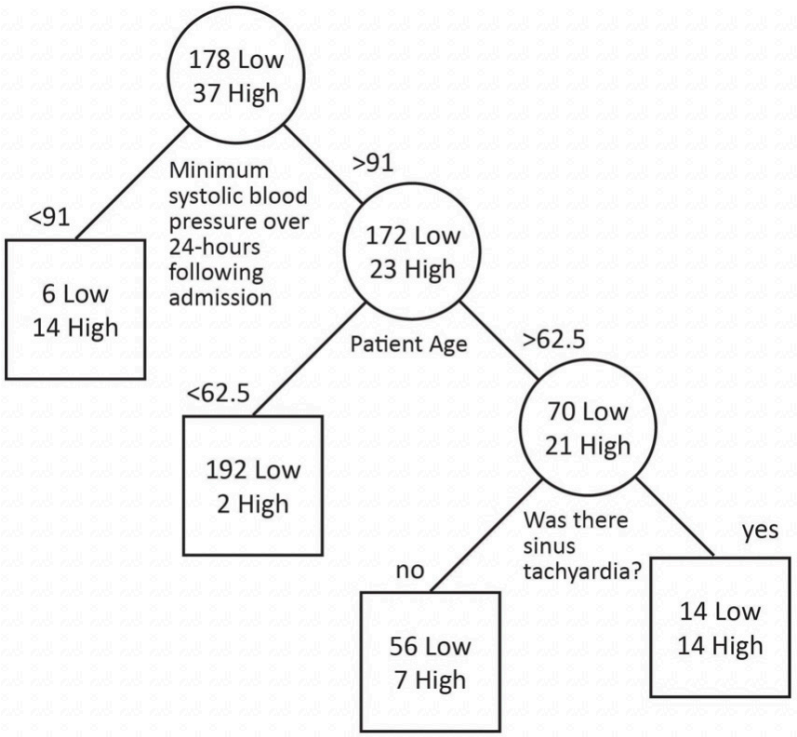


Figure A11

One advantage of CART is that it is readily applicable to large data sets. Another advantage is that it mirrors some human procedures for classification and therefore the results are more easily interpreted. A disadvantage (shared by most tree-based methods) is the level of vigilance required during the training process. A small change in the training data can result in a large change in the predictions of CART. Therefore, careful validation of CART decision trees is essential. Once that vigilance is in place, CART and related decision tree methods are among the most effective procedures available to practitioners.[41,42]

Naïve Bayes Classifier

Fairness and bias considerations: Naïve Bayes classifiers can bring about unfair classification outcomes. Kamishima et al. showed that NB classifiers would assign female data entries to the class "low income" even when the sensitive attribute of gender was removed.[43] Most of the remedies to such prejudicial outcomes require some form of probabilistic discriminative modeling, which is a feature NB classifiers lack.

Overview of technique: A classifier based on a conditional probability model with an assumption of independence among the features. The models are not necessarily Bayesian in the strict sense, but the updated rule is based on Bayes' Law, so the name has persisted. Given a class C_k and a set of features x_1, x_2, through x_n, Bayes Law states:

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k)P(x_1, x_2, \ldots, x_n|C_k)}{P(x_1, x_2, \ldots, x_n)}$$

If we assume that the features are independent, which is an assumption that would not apply in many applications, then the chain rule for probabilistically independent events can be used to implement the conditioning.

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{\prod_{i=1}^{n} P(x_i)}$$

41. Laura Martignon, Konstantinos V. Katsikopoulos, and Jan K. Woike. "Categorization with Limited Resources: A Family of Simple Heuristics." *Journal of Mathematical Psychology* 52, no 6 (2008): 352-361.

42. Nathaniel Phillips, Hansjörg Neth, Jan Woike, and Wolfgang Gaissmaier. "FFTrees : A toolbox to create, visualize, and evaluate fast-and-fru gal decision trees." *Judgment and Decision Making*, 12, no. 4 (2017): 344-368.

43. Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. "Fairness-aware Learning through Regularization Approach." In *2011 11th IEEE International Conference on Data Mining Workshops* (New York: IEEE, 2009): 643-650. http://dx.doi.org/10.1109/ICDMW.2011.83

In this case, the naïve Bayes classifier will be computationally efficient even if there are a large number of features. The computations for the updating step grow only linearly with the number of features. In practice, naïve Bayes has demonstrated good performance even if the training sets are small. The accuracy of classification is good enough in many contexts even when the independence assumption is not very well satisfied, however the accuracy is usually not as good as that of random forests.[44]

Random Forest

Fairness and bias considerations: Random forests can provide some of the best performance in decision making and work well in the presence of uncertainty and variability. However, a critical drawback is that it can be hard to explain why a random forest leads to a particular decision in any particular case. This can be a serious drawback in cases where individual-level explanations are needed for transparency or accountability.

Overview of technique: An ensemble method for classification in which the training process involves constructing a multitude of decision trees and each tree casts a unit vote for the most popular class.[45] The predominant training algorithm for random forests is bootstrap aggregating, or bagging. During training the training data are sampled at random with replacement (thus, "bootstrap" is the "b" in bagging). Additional trees are constructed to fit the bootstrap sample and added to the forest. The differences in trees can also be brought about through various other techniques such as the 'Random Subspace method'[46] (aka "feature bagging"). Here, a subset of the input variables (features) is selected for decision-making at each node. Additionally, all these methods have to deal with aggregation of the outputs of multiple trees which is central to the procedures' effectiveness. Some of the most common approaches to aggregation are voting (commonly applied for classification) or averaging (commonly applied in random forest regression). A significant advantage of random forests is that they benefit from the properties of ensemble methods that overall outputs are superior to the best individual predictors in the ensemble as long as some mild conditions are met regarding diversity and accuracy of the ensemble.

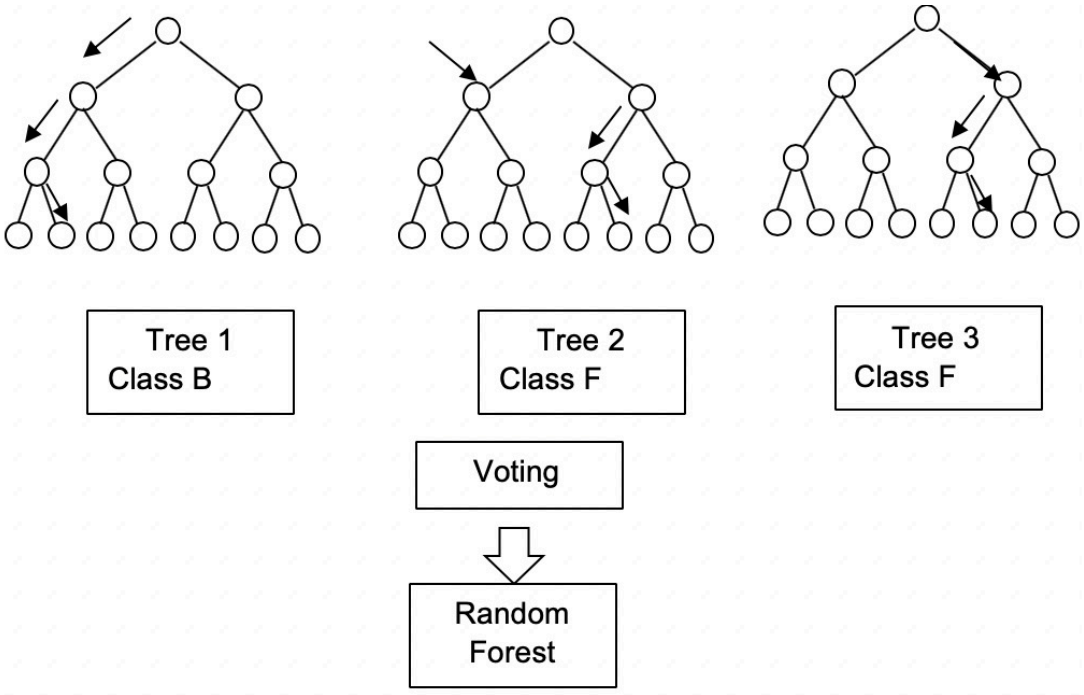


Figure A12

44. Rich Caruana and Alexandru Niculescu-Mizil, "An empirical comparison of supervised learning algorithms." In proceedings in 23rd International Conference on Machine Learning. (New York: Association for Computing Machinery, 2006): 161-168. https://doi.org/10.1145/1143844.1143865

45. Breiman, L., 2001, "Random Forests," *Machine Learning* 45, (2001): 5-32.

46. Tim Kam Ho, "The random subspace method for constructing decision forests" In IEEE Trans.on Pattern Analysis and Machine Intelligence 20, no 8 (New York: IEEE, 1998): 832–844. https://doi.org/10.1109/34.709601

Artificial Neural Network (ANN)

Fairness and bias considerations: A common critique of ANNs is that the requirements for training can be too demanding for many real-world applications. ANNs can converge to many different solutions because many local minima exist and the training procedures may get stuck in local minima. Importantly, the outputs of ANNs can be very difficult or sometimes impossible to interpret. Therefore, great vigilance must be exercised in applying ANNs to interventions in the developing world so that instances of unfair outcomes can be detected and mitigated in a timely fashion.

Overview of technique: This is a family of computational models comprising interconnected systems of simple units or nodes that are often called artificial neurons because they are loosely based on the operation of neurons in animals. The signal into the artificial neurons are continuous real values. The output of each artificial neuron is a (usually) non-linear function of the sum of its inputs. The connections between artificial neurons are often called edges and these are assigned weights whose values change during the training / learning process. ANNs can be used for supervised or unsupervised learning. ANNs have demonstrated significant flexibility in applications with unstructured data and complex, non-linear relationships.

Exploring Fairness in Machine Learning for International Development

MIT D-Lab | CITE
Massachusetts Institute of Technology

January 2020