Final Report: Visualizing Fairness in COVID-19 Resource Allocation Model

Gurpreet Kaur Khalsa, Mudit Mangal and Orissa Rose

May 8, 2023 | INFO 247

Project Links: https://egaleco.herokuapp.com/scrolly + Product Overview Video (YouTube)

Project Goals

Machine Learning (ML) algorithms increasingly dictate opportunities and outcomes for individuals and groups across economic, social, political, and legal contexts. This is especially true in healthcare, where algorithms are being introduced to support practitioners and professionals as they predict and diagnose disease,[1] allocate medical resources and manage patient health.[2] While the use of ML has contributed to important advancements in the field, some applications have created disparate impacts that replicate discriminatory societal and institutional inequities along race, gender, age and ability categories.[3]

As acknowledgement of the impacts of ML bias grows, data scientists are expected to evaluate their own models for bias and implement mitigation strategies before going to market. Accordingly, numerous "Fairness Toolkits" have emerged to help data scientists evaluate their models for undue bias and proactively reduce algorithmic harms. These toolkits rely on "fairness metrics" that consider model outcomes across data subgroups, such as TPR Parity, FPR Parity, TNR Parity, FNR Parity, PPV Parity, FOR Parity, FDR Parity, NPV Parity, Demographic Parity, Accuracy Parity, and many more. A glossary of these terms is available in the appendix.
The 2022 study, *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits* found that ML professionals use existing fairness toolkits inconsistently and encounter shortcomings that limit their usefulness.[4] That study and our own competitive analysis of 10 prominent fairness toolkits revealed these key problems:

- Fairness tools have a steep learning curve. Many are code-only interfaces which require high technical acumen from users and that they be familiar with fairness auditing and fairness metrics;
- Fairness tools lack guidance about how to choose the best of many fairness metrics for their particular context; and
- Fairness tools oversimplify model bias nuances by providing users a singular fairness metric when in reality several would provide a more holistic view into the fairness implications of their model.

Motivated to address this problem space, our Capstone project team built Egaleco, a web-based tool that acts as "training wheels" for data scientists seeking to evaluate their healthcare ML models for bias. Users are able to run a fairness assessment on their own test data, or use the COVID-19 mortality prediction dataset we enable for demo purposes. Within the tool, we use animation and progressive disclosure to re-imagine decision tree logic that is central to many

[1] (Ullah et al., 2020)
[2] (Kumar et al., 2022)
[3] (Obermeyer et al., 2019; Straw et al., 2022; Ross, 2023)
[4] (Deng, et al., 2022)

predictive models, and leverage interactive visualizations to contextualize the meaning of fairness in ways that code-only toolkits cannot. We've created a web page specifically for this InfoViz submission that details the work Gurpreet, Mudit and Orissa did. Specific tasks our visualizations are targeted towards include helping data science users:

- Determine the best fairness metrics to evaluate for their use-case;
- Define what fairness means for their use-case; and
- Contextualize the probable harms of their models and identify which groups need to be focused on to mitigate unfairness.

Related Works

The literature and applied works that informed our project are a combination of: technical machine learning papers discussing strategies for measuring quantitative fairness in models; visualization literature and STS investigations of the ways that people learn about fairness from toolkits and visualizations.

1) *Exploring how machine learning practitioners (try to) use fairness toolkits* (Deng et al 2022).

   This work was a central motivation for our project and very helpful in understanding the problem space. The research reviewed popular fairness toolkits and interviewed data scientists about how they use fairness toolkits to check for bias in their models and products. Key findings include:

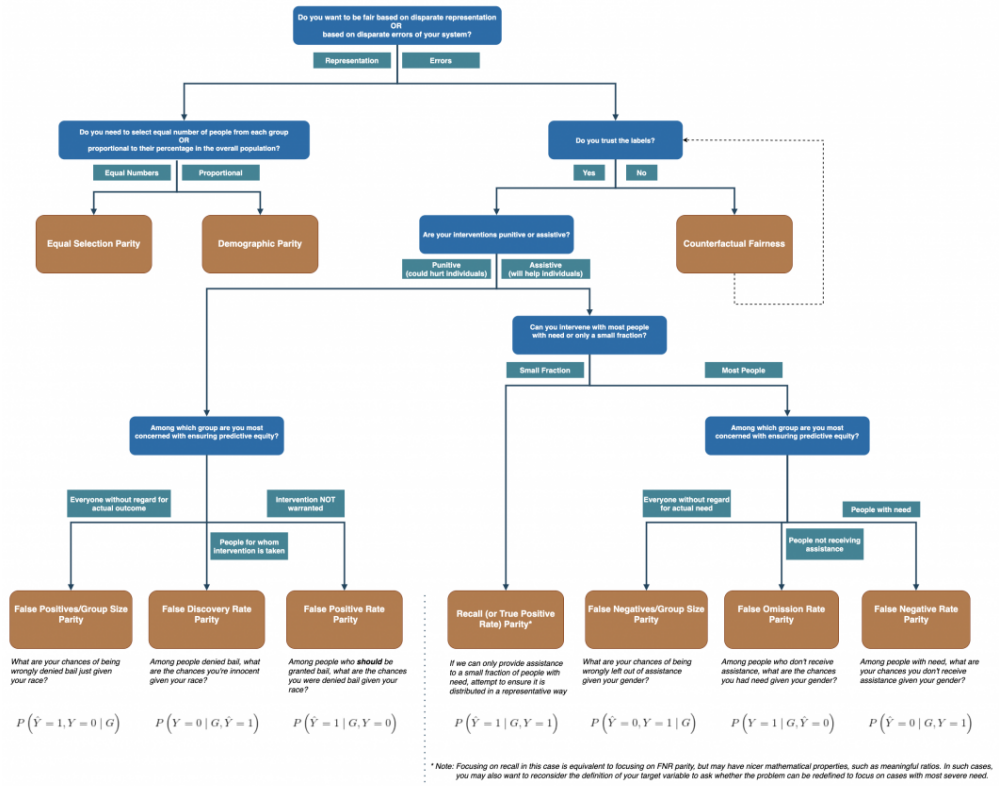
   - Data scientists want fairness toolkits to contextualize the meaning of fairness and present the information in a way that they can easily share with colleagues;
   - Data scientists are pressed for time. They want quick and intuitive tools;
   - Data scientists want to toolkits to be educational and help them learn more about the fairness space; and
   - Data scientists want to be able to run their domain specific algorithms through toolkit assessments.

2) *A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education* (Quedado, et. al., 2022)

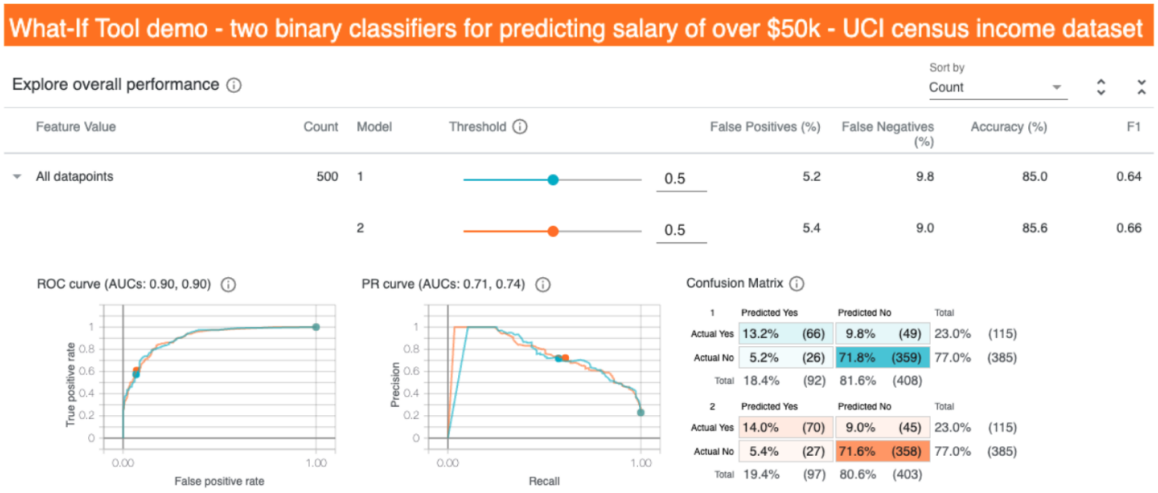
   This paper evaluated the effectiveness of the visualization components of six popular open-source fairness toolkits (Aequitas, Fairlearn, AI Fairness 360, What-If Tool, Dalex, and Responsibly). Of particular interest to us, the study showed that the Aequitas visualizations were favored by ML students for their simplicity, while the What-If Tool felt like information overload for some users. Overall, interactive fairness visualizations

performed the best. This paper was a useful guide in our research and compelled us to design Egaleco's fairness visualizations with an eye towards simplicity, interactivity and animation.

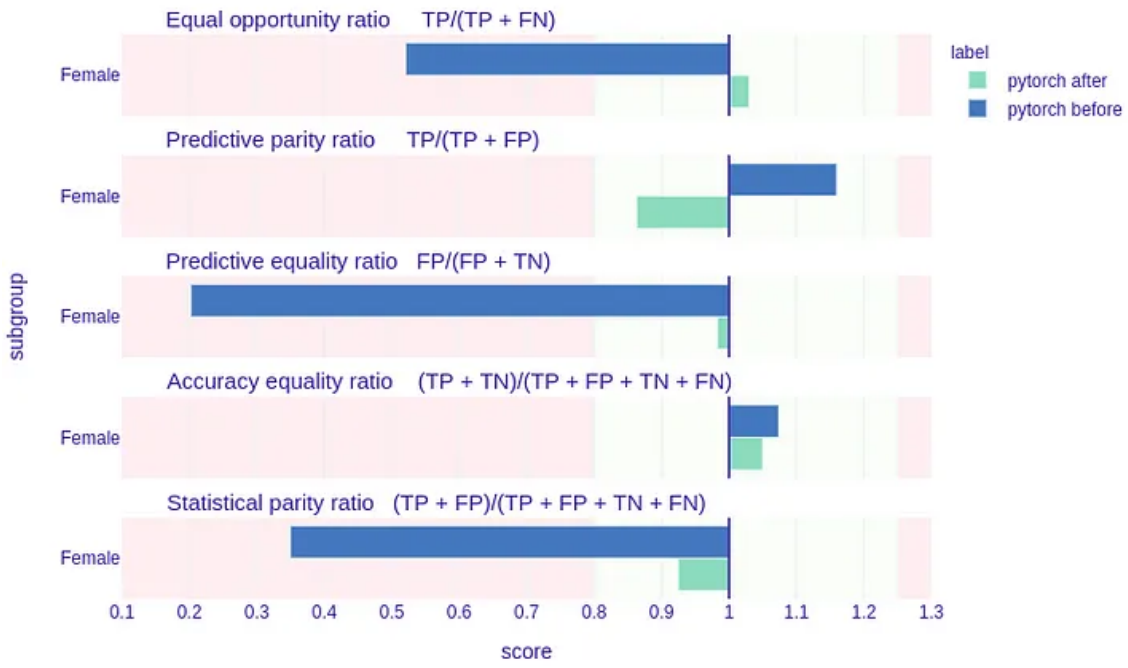
3) Aequitas is an open-source fairness tool that was created for data scientists by academics at the Center for Data Science and Public Policy at University of Chicago. To help users conceptualize the decision tree logic that is central to binary prediction algorithms they developed a "fairness tree" (below). We were inspired by the way they embedded questions to guide users towards the appropriate fairness metric for their use case. We felt that the presentation of the entire tree at once led to cognitive overload that we could lessen through progressive disclosure.

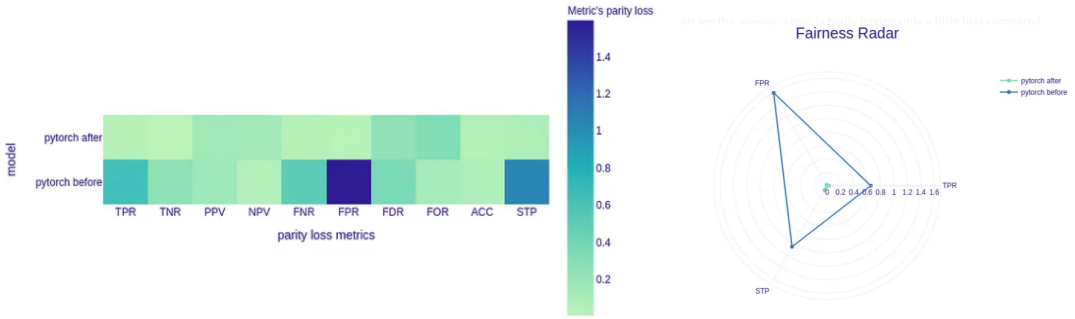


4) What-If is a tool designed by Google to help users "visually probe the behavior of trained machine learning models, with minimal coding". This tool lets users explore the performance of a ML model through different prediction evaluation metrics like the percentage of False Positives, False Negatives and Accuracy. We like how the interactivity allowed users to see the real-time impacts of changing disparity tolerance, however thought the charts were a bit confusing and wished that we could see a more granular analysis of model performance by subgroup.



5) "Visualizing ML model bias with dalex" by Jakub Wiśniewski (2021) is a blog post that discusses ways to use the Dalex python package to create visualizations for model bias. The author runs the package on COMPAS Recidivism Risk assessment data to show the different ways users can create visualizations that communicate fairness metric parity. Several of the visualizations Wiśniewski developed (below) informed our initial design plans. We liked the idea of creating a heatmap to show the trends of different fairness metrics simultaneously, however ultimately abandoned that idea because heatmap adjacencies are meant to show a pattern that wasn't feasible with the fairness metrics.



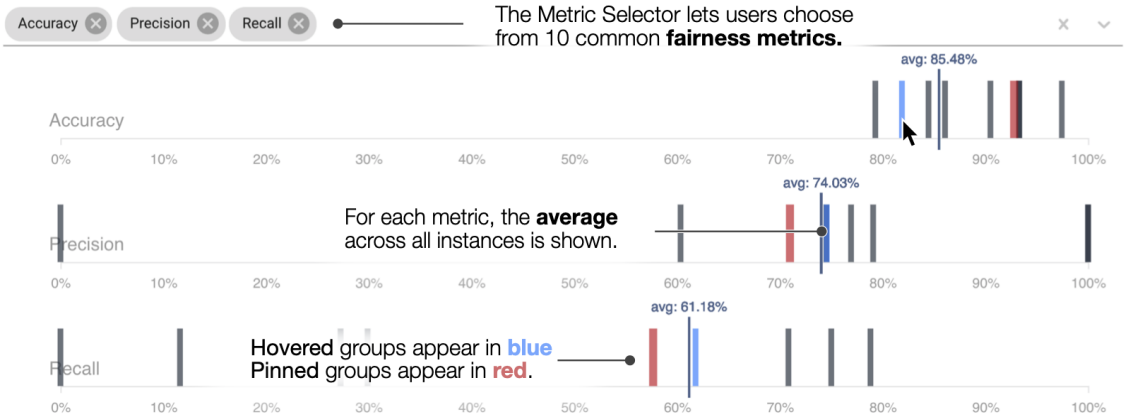


6) *COVID Mortality Prediction with Machine Learning Methods: A Systematic Review and Critical Appraisal* (Bottino et al 2021)

This work surveys 24 papers on the different types of ML-based COVID-19 prognosis systems that were established during the pandemic. Bottino et al. found that the most common ML techniques for COVID mortality prediction are logistic regression, random forest, gradient boosting decision tree, and artificial neural network models. This was a useful piece of background research with many connections to our problem area given that Egaleco is a healthcare focused application, our demo uses a CDC case surveillance dataset and the scrolly portion of our design aims to bring decision tree logic to life.

7) FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning (Cabrera et. al IEEE VAST, 2019)

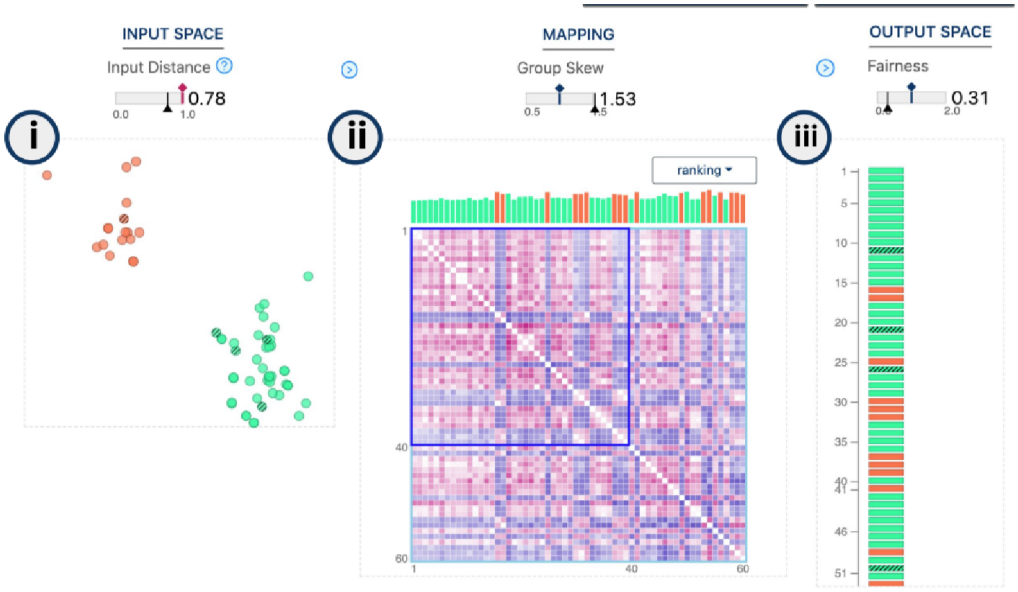
This paper presents FairVis, a visual analytics system that helps users find biases in their model through summaries of model performance by subgroup. It is especially relevant to the development of Egaleco's visualizations because it's considered useful in helping data scientists understand intersectional bias and explore the varied definitions of fairness. FairVis' interactivity allows users to audit for known biases (based on the groups a user selects) and explore unknown biases (based on the subgroups this tool recommends through clustering and similarity techniques).



Similar to FairVis' Subgroup Overview (pictured), we allow users to explore performance of many subgroups using different metrics. We iterated on this by clearly displaying the reference group so that users can easily make sense of the disparity between groups; recommending which metrics to prioritize; and presenting more metrics.

8) FairSight: Visual Analytics for Fairness in Decision Making (Ahn et. al, 2019)

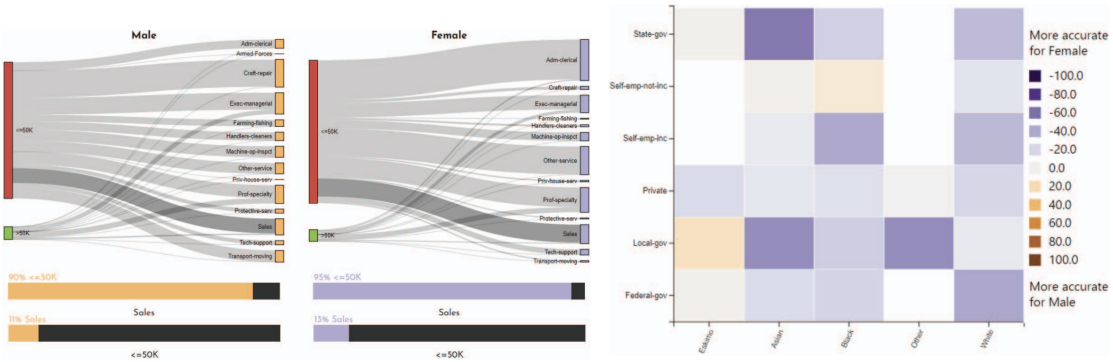
This paper presents "FairDM" a workflow framework for unbiased decision making and "FairSight" a visual analytic that ranks individuals based on protected class membership and summarizes different notions of fairness. The paper suggests many interesting ways to represent measures of fairness including a 2d plot, a color coded matrix, and a ranked list plot but ultimately these didn't work for our decision tree logic model since that produces binary classification outcomes (not ranked outcomes). Finally, this work discussed that even with the best design, "human scrutiny through the interactive visualization is still required". This informed our choice to make all of our visualizations interactive for the user.



9) How Do Algorithmic Fairness Metrics Align with Human Judgement? A Mixed-Initiative System for Contextualized Fairness Assessment (Constantin et al, 2021)

This paper presents the FairAlign system for improving cognition of fairness assessments through interactive visualizations. Their toolkit is aimed at two different types of users evaluating the fairness of classification models – laypeople and data scientists. The laypeople analyze interactive visualizations then provide qualitative annotations about the fairness of a particular model. The data scientists then analyze the laypeople's feedback alongside predefined fairness metrics. The goal of this study was understanding the best ways to contextualize metrics and fairness so that creators

of fairness tools can close the gap between human judgment and automated fairness evaluation. The paper includes numerous examples of visualizations used for algorithmic fairness assessments (below). Sankey diagrams are used to compare distribution of outcomes over categorical features. It shows how the predicted label distributes over the distinct subgroups spanned by a particular protected attribute (like male or female). This does not target a particular fairness metric and is more part of the data exploration phase, which is out of scope for the InfoViz project, but is included in another part of the Egaleco tool. The heatmap matrix is used to compare accuracy differences between two groups. It incorporates fairness metrics like accuracy differences over two protected groups to highlight intersectionality but doesn't compare differences between categories of a single protected class – which is what we aim to do with our intersectionality dot plot.



Visualization Usability Testing and Results

For our first round of usability testing, we recruited three data scientists to participate in pre and post visualization interaction questioning, task observation, semi-structured interviews, a demographic survey, and a Likert-style evaluation. They provided rich feedback which we summarize in the following sections and responded to in subsequent visualization redesigns.

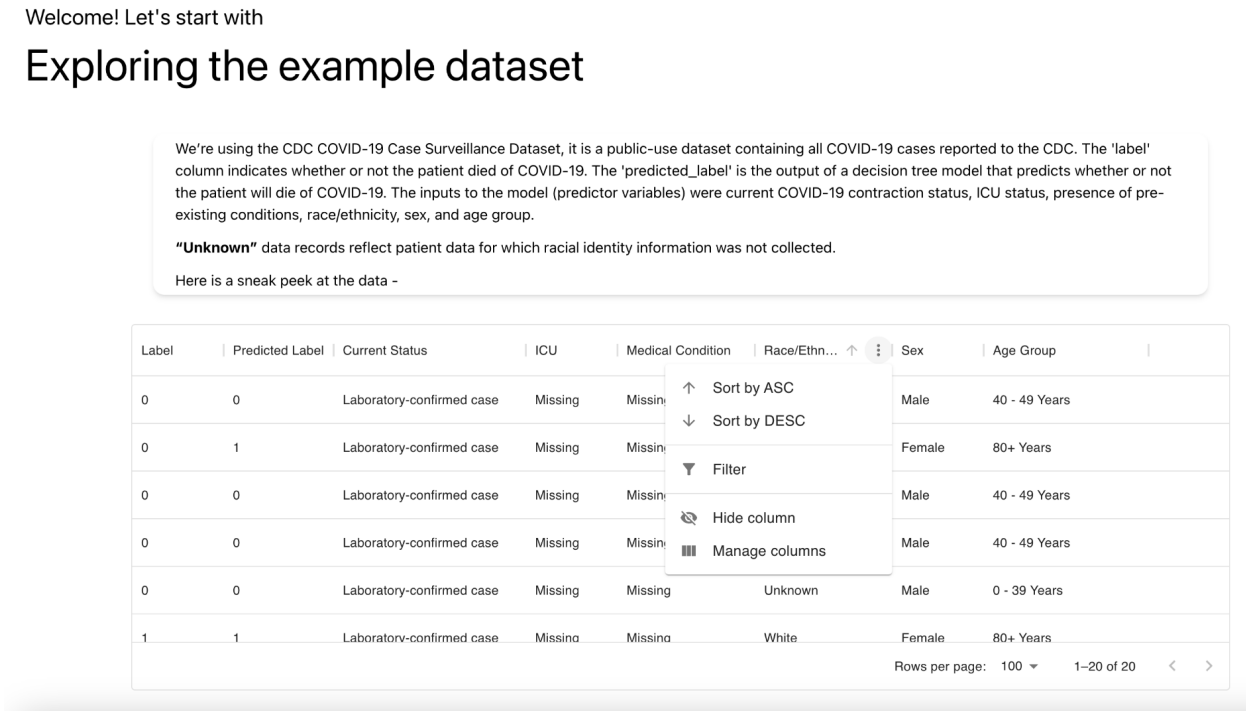
After revisions based on feedback from the first round of usability testing, we solicited feedback from a new batch of prospective users. That feedback led us to these final designs and helped us identify interface bugs and narrative misspellings.

A key takeaway from all of our usability testing was that fairness is a complex and nebulous topic even for people with expertise in the field. This aligns with what our larger Egaleco research team learned during interviews with data scientists. Many of them have a preferred set of reference groups and fairness metrics that they repeatedly use. Our hope is that Egaleco will make it easy for them to explore new approaches to fairness because we believe that diversifying evaluation methods and perspectives is critical to understanding the disparate

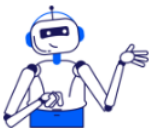
impacts of algorithmic systems. We think this InfoViz implementation makes the process seem less daunting and gives prospective Egaleco users a sense of how the tool can help their work.

Visualization #1: Dynamic Metric Selection

Since our target audience is data scientists that expect to see the data they are working with, we chose to begin the visualization with a view of the demo dataset being used. Initially, it was a static datatable. In response to user feedback, we embedded an interactive datatable that lets users sort and filter data columns as desired. This gives the user a chance to explore the category labels and sample entries that will be evaluated and visualized. Other changes we made in response to usability testing include: the addition of an introductory narrative that explains the scenario more clearly, and a disclaimer on the meaning of the “unknown” race category.



Upon scrolling down the page, the user makes contact with the Egaleco bot. Users requested that we make the process very clear, so we added explicit details that the bot will ask them three questions and update the ranking in order to identify the appropriate fairness metrics for their use case. We chose the bot iconography because Human-Centered AI best practices warn against using a design that would lead users to anthropomorphize the algorithmic system.



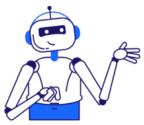
Auditing for fairness can be hard. There is a plethora of fairness metrics you can use in an audit, and many are mutually exclusive.

To help you identify the most appropriate metrics for your use case, we will ask you 3 questions and the metrics on the right will get prioritized with each choice you make!

False Negative Rate Parity (FNR)
False Discovery Rate Parity (FDR)
Demographic Parity
Equalized Odds
True Positive Rate Parity (TPR)
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity

At the beginning of the scrolly, we show users a list of all of the possible fairness metrics that the Egaleco assessment can evaluate. Some usability testing participants wanted us to provide definitions for each of these metrics at this stage in the visualization, but that would have resulted in cognitive overload and long, dense text blocks. To ensure that we are presenting users with the definitions that matter to them, we provide a curated list of definitions at the end of this process. The list covers only the specific metrics that are determined to be appropriate for their use case.

The three questions that we ask users are our reformulation of the decision tree questions that Aequitas developed (see image in the Related Works section). We chose the multiple choice and progressive disclosure format of the questions because these design choices make the user an active part of the process, which imbues them with a sense of agency. Foundational research and interviews conducted with data scientists who focus on ML fairness experts as part of our broader capstone effort revealed that there is potential for fairness work to be sidelined because of a lack of clarity and ownership. We believe involving the user in question answering from the start will deepen their sense of authority and commitment to fairness work.



Which best describes the way your model is defining fairness?

○ **A: Equal Representation**
The probability of the positive predicted outcome within each group should be equal between groups—for example, 30% of group 1 and 30% of group 2 are positive outcomes.

Show Help Show example

◉ **B: Equal Errors**
The model should perform equally well, and shouldn't harm one group more than the other. We'll narrow down how best to define performance next.

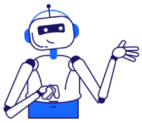
Show Help Show example

False Negative Rate Parity (FNR)
False Discovery Rate Parity (FDR)
Demographic Parity
Equalized Odds
True Positive Rate Parity (TPR)
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity

In addition to the animation of the metrics, we included “show help” and “show example” buttons that reveal definitions and contextual examples so that uncertain users won't have to leave the interface to google a concept or term.

As the user answers the bot’s questions and continues their scroll downwards, the list of possible metrics transforms. The metrics that are still available for use are highlighted in light blue and move to the top of the list and, while those that are not viable move to the bottom. We chose to animate the changing of metrics instead of having the metric disappear once it’s no longer appropriate because we want the user to develop a sense of how the decision tree logic works and how each answer they give has an impact on the fairness assessment.

In response to usability testing feedback, we enabled reverse scrolling, so the user can scroll upwards if they want to review what triggered a given reordering. In our first interaction of the visualization, the metric shuffle was the only visual change that happened and it was triggered exclusively by a downward scroll. User feedback prompted us to introduce a second visual transformation, so that now the color of the metrics change as soon as a radio button is selected, and then the shuffle is triggered by a downward scroll.



How would you characterize the availablility of resources available to people who need them?

◉ **A: Limited Resources**
The intervention based on the model's predictions is resource-constrained, and you can only give benefits to some of the people who should get them.

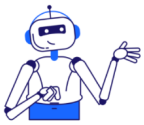
Show example

○ **B: Abundant Resources**
There are sufficient resources to offer benefits to all people who should get them.

Show example

False Negative Rate Parity (FNR)
Equalized Odds
True Positive Rate Parity (TPR)
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity
Demographic Parity
False Discovery Rate Parity (FDR)

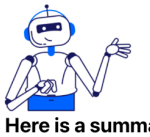
In addition to the animation of the metrics, we included “show help” and “show example” buttons that reveal definitions and contextual examples so that uncertain users won't have to leave the interface to google a concept or term.



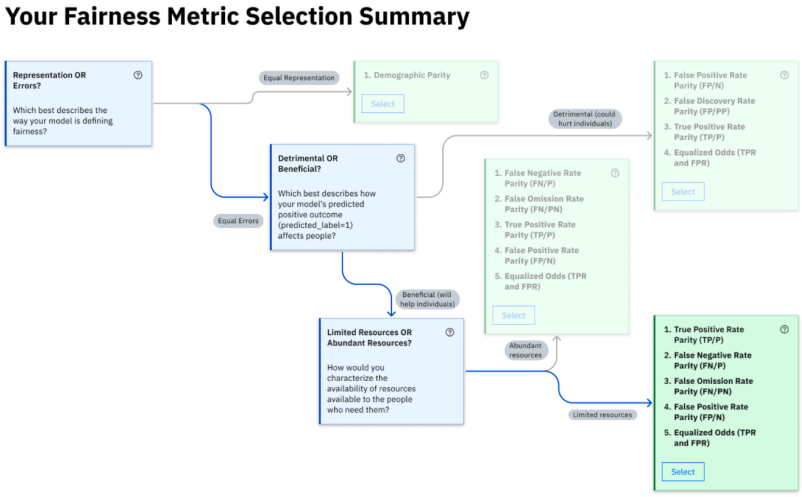
Based on the information you shared, here are the fairness metrics that matter the most for your use case.

True Positive Rate Parity (TPR)
False Negative Rate Parity (FNR)
Equalized Odds
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity
Demographic Parity
False Discovery Rate Parity (FDR)

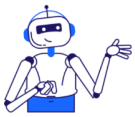
At the conclusion of the question and answer interaction, we display a flowchart that explains how their responses dictated the fairness metrics shown to them. We provide this to the user as a comprehension tool, a sort of map showing them their journey to arrive at these metrics. These elements were implemented in response to user requests for a summary of what led to the final recommendation.



Here is a summary of your interactions



We leverage the transition between our interactive features to educate the users about two key themes in ML fairness assessments – protected classes and reference groups. We use this imagery of people to emphasize that there are humans impacted by these choices.

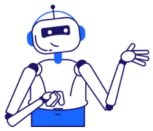


Protected Classes

Now that we have finalized the fairness metrics let's familiarize ourselves with the concept of Protected Classes. Protected Classes are personal attributes like sex, age, race, national origin, and so on. The law prohibits discrimination on the basis of those characteristics. For the COVID-19 resource allocation model, we will treat Race and Sex as protected classes.

Protected Classes for COVID-19 resource allocation model are

Race and **Sex**



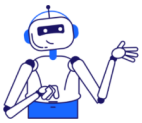
Reference Groups for Protected Classes

A Reference group is a group that you use as the baseline for comparison when assessing the fairness of your model's classifications. Usually, groups that are historically advantaged or highly represented in the dataset are selected as reference groups because the model tends to perform well on them. Once the reference group is selected, fairness is viewed as how the model performs for other groups as compared to the reference group.

Reference Groups for Protected Classes Race and Sex are

White and **Male**

In response to user confusion over the significance of particular fairness metrics, we provide metric definitions, illustrative charts and implemented a dynamic confusion matrix that appears before displaying charts. We show metric name, definition and guidance on when to use it. To drive cognition with visual aids, we display bar charts that show raw values for protected class attributes in the given metric.

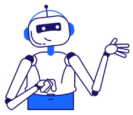


Let's visualize these metrics!

False Negative Rate Parity (FNR)
Equalized Odds
True Positive Rate Parity (TPR)
False Positive Rate Parity (FPR)
False Omission Rate Parity (FOR)
Accuracy Parity
Demographic Parity
False Discovery Rate Parity (FDR)

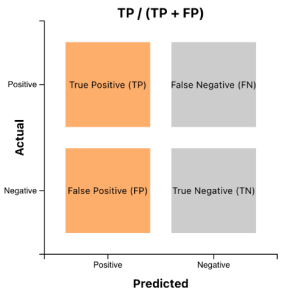
A confusion matrix is a common data science tool that summarizes model performance through the rate of true positives, false positives, true negatives, and false negatives. Our

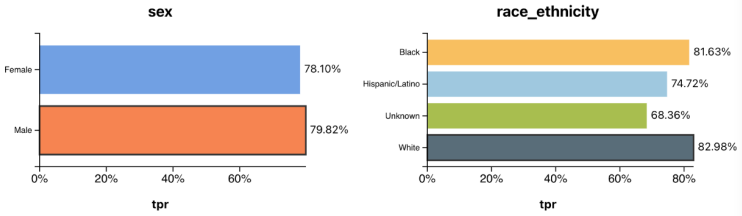
confusion matrix visualization displays calculation formulas and highlights relevant metric cells to explore each one individually and compare them against each other. While they may seem new to an unacquainted user, confusion matrices are commonly used in data science thus a useful visual for our target audience. All together, the sample data, asking of questions, ranking of metrics, and presentation of key concepts and term definitions **ensure we satisfy the first goal of our project – to help data scientists determine the best fairness metrics to evaluate for their use-case**.



True Positive Rate (TPR) and TPR Parity

True Positive Rate (TPR) measures the probability that subjects in the positive class (P) have positive predictions. TPR Parity requires these probabilities be equal between groups. Let's look at the TPR values for each group in the bar charts on the right to get a sense of the disparities.

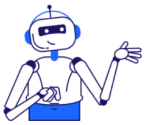




Visualization #2: Model Fairness Disparity Across Races

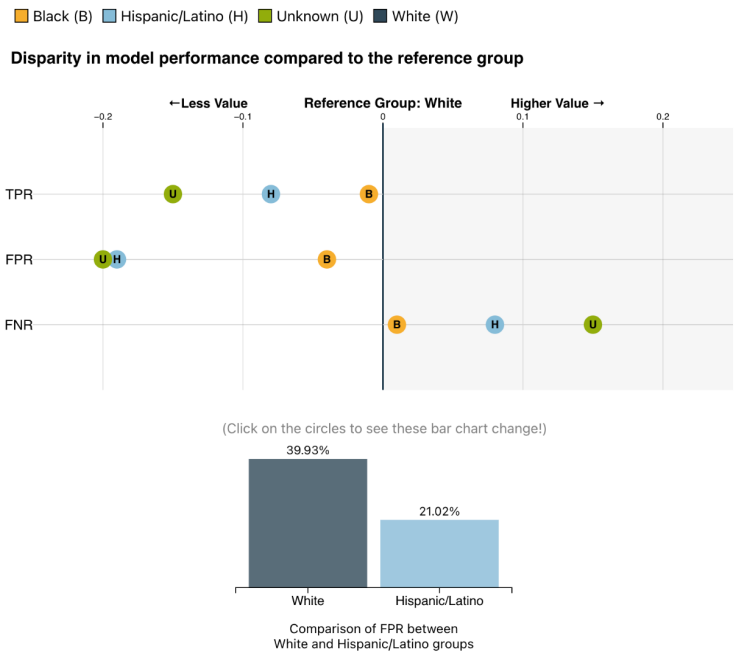
We were inspired to create a visualization that displayed several fairness metrics at once because we hadn't seen it done in other tools. We think this is an important part of contextualizing fairness because it helps the user decide which metric is the most revealing of model bias. **By presenting TPR, FNR and FOR simultaneously we enable comparisons and provoke the user to define what fairness means for their use-case — which is the second goal of this project.** Being able to see a raw disparity value upon clicking, the user can explore the data as much as they need to make a decision.

A core element of understanding the fairness implications of a model is the reference group. Since we defined the concept in the prior scrolly interaction, the user is ready to understand it in context here and note patterns in model performance. The purpose of a reference group is to serve as a benchmark, which is why we made it the central axis in the dot plot and a constant bar in the accompanying chart.



Disparity across Races

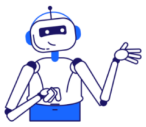
This chart shows the difference in model performance across racial groups and across fairness metrics. Move your cursor over the colored circles to explore the disparity values between a given racial group and the White reference group. Click on the circles to see the metric values change in the bar chart.



Our first version of this disparity dot plot did not include the comparison bar chart. The second version of the chart had the bar charts above the dot plot but no descriptive caption and the y axis range changed according to the group selection. In response to user feedback, for our final version we moved the bar chart below the dot plot, and removed the y axis entirely since the goal was for the user to compare the two bars and we provide the raw value as a bar label. We also renamed the X axis to be more clear, and changed the position of the instructions because it wasn't immediately intuitive to all users. Finally, we added relevant abbreviations for the race groups into the legend. All of this was done as an alternative to adding more details into the tooltip and helped us avoid chart clutter.

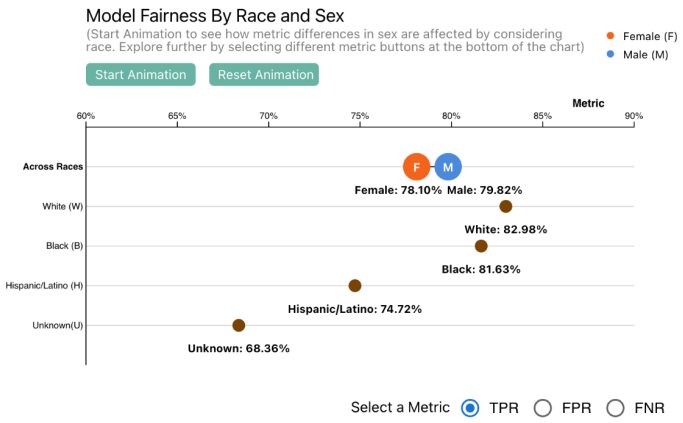
Visualization #3: Effect of Intersectionality on Model Performance

A truism of data work is that a high level interpretation of trends can be disproven when we look at the data more granularly. This holds true of fairness metrics when you consider intersectionality and was the core motivation for this visualization. To establish a baseline, we show users how a given fairness metric performs when looking at overall performance for race without respect to sex, and overall performance for sex across all race groups. The first view tells one story about the fairness of the COVID-19 resource allocation model – that it performs better for men overall and better for white individuals overall.

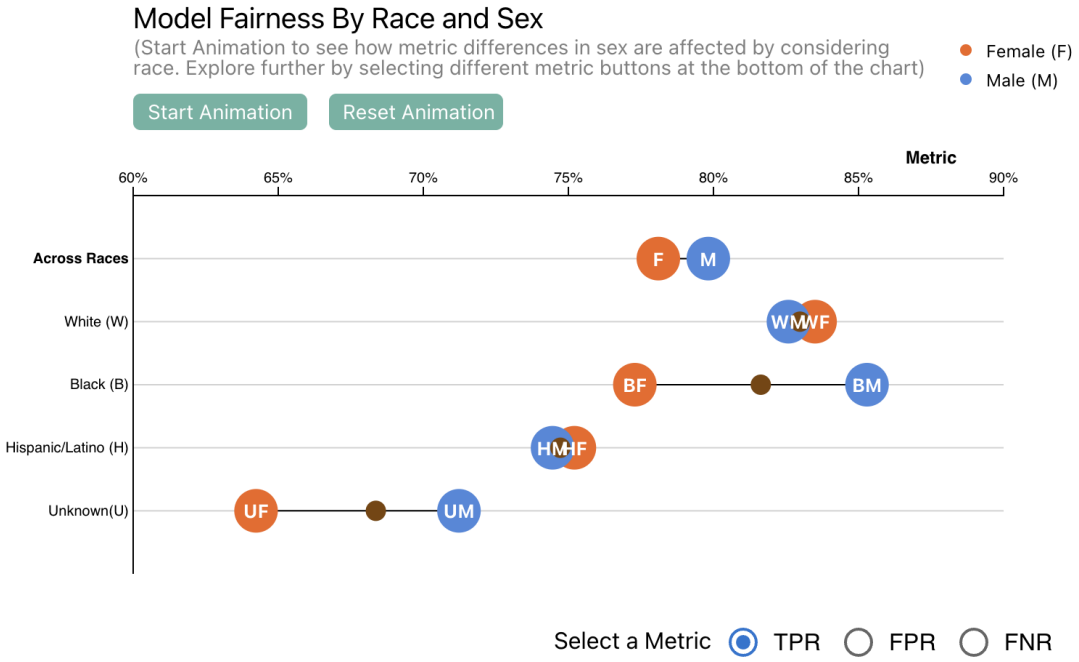


Disparity across Race and Sex

This chart demonstrates how model performance varies when you consider race and sex simultaneously. Interact with the visualization to explore which groups are most impacted by an intersectional lens.



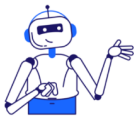
However, when the viewer explores the data more by activating the animation, that story changes.



To support interpretation, and in response to usability testing and user requests, we added instructions about how to activate the animation and created two obvious trigger buttons that say “start animation” and “reset animation”. We also inserted letters denoting the sex and race of a given mark.

The interactivity of selecting a single metric and exploring its performance across different race and sex groups fulfills our third project objective – to help users contextualize the probable harms of their model and identify which groups need to be focused on to mitigate unfairness. For example, if the user assessed the bias of their model based on the overall view for TPR (first image), they might assume that the model performs roughly as accurately for females as males. However, when they view the assessment broken down for race *and* sex (second image), they see that the model performs much worse for females who are Black or race Unknown. By showing the impact of intersectionality, users get a clear understanding of which group needs to be the focus of their bias mitigation efforts.

In the final round of usability testing our users interpreted the chart correctly but said they wanted us to provide language about "next steps". They expressed uncertainty like, "okay so I get it's unfair for women and unknown race groups but what am I supposed to do?". Advising users on specific bias mitigation strategies is out of scope for this project and is not something that can be visualized, however we did add guidance about mitigation steps to take.



So what's next?

If you identified any unfairness in your model today, it's important to take steps to address and mitigate those biases before you deploy your model.

Bias mitigation can be done at a few stages in the model development process. There are **pre-processing, in-processing, and post-processing** bias mitigation algorithms.

If the training data is able to be modified, pre-processing algorithms can be applied to the training data to reduce bias by revising the data. When the model is being trained, in-processing algorithms that penalize bias and prioritize fairness can be used.

Lastly, once the model has produced outputs, post-processing algorithms can be used to revise the model's predictions to prioritize fairness.

Finally, we strove to use color palettes that do not reinforce discrimination or stereotypes. Accordingly, we chose this earth-tone palette for racial groups and orange, brown and blue for sex.

Data and Modeling

- CDC COVID Dataset: The dataset includes demographic information related to race, sex and age group apart from information on presence of underlying comorbidity/disease, ICU admission status etc.

  https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf

- Data Cleaning and ML model implementation on this dataset were done in conjunction with other team members of our Capstone project. Some data cleaning and data pre-processing steps are adapted from this Kaggle notebook. Steps included:

  - Removal of outliers and imputation of missing entries;

    - Removal of observations with death status as “missing” or “unknown” as they were irrelevant for our use case.
    - Creation of a less granular Age group variable which combined age groups in the “0-19 years” category
    - Removal of race groups like Native Hawaiian which constituted less than 0.5% of the data. Other minority groups which were aggregated in the category “Multiple/Other” and groups like “Asian” had a ~3% proportion and were also removed. In the absence of sufficient data points, our basic ML model lacked the complexity to be able to capture the patterns for these data points well. We capture removal of such race groups which are not represented well in the data and are not considered in our fairness analysis, recognizing that might lead to possible “erasure harms” and note this down as a limitation of our work.
- Creation of a classification model that predicts the likelihood of an individual dying from COVID-19. Our python notebook with all the steps is here. Some important considerations:
    - We use target variable (dependent variable) as “death_yn” and the features (model inputs) as “current_status”, “hosp_yn”, “icu_yn”, “medcond_yn”, “age_grp”, “race” and “sex_grp”.
    - Various ML models are implemented, and we choose the decision tree model as it gives the highest overall model accuracy.

Tools Used

- Figma
- Observable (D3.js)
- React
- React-scrollama library

Conclusion & Future Work

We are pleased with the extent to which we were able to achieve our project goals during the semester timeline. We feel confident that this work helps data scientists understand fairness in a way that existing fairness visualizations don’t. However, we plan to continue developing the functionality of our visualizations. In particular, current users only have the option of visualizing fairness with our demo dataset, and with regards to sex and race. In future iterations, we hope to allow users to explore fairness metric visualizations with the protected class variable of their choosing (e.g. age, national origin, etc). Additionally, we hope to fine tune the model so that it performs accurately even for groups that are a small portion of the dataset. This will allow us to include minority groups as a part of the fairness assessment instead of having to remove them.

Project Contributions By Teammate

Project Component	Sub Component	Gurpreet K.	Mudit M.	Orissa R.
Data Preparation	Data Sourcing	0%	100%	0%
	Data Preprocessing	40%	60%	0%
Visualizations	ScrollyTelling	60%	5%	35%
	Disparity Chart	30%	60%	20%
	Intersectionality Chart	15%	80%	15%
	Website embedding	100%	0%	0%
Design	Website Text Writeup	45%	10%	45%
	Website Layout Design	40%	10%	50%
User Testing and Reporting	User Testing	10%	10%	80%
	Report Writing	5%	5%	90%

Assets uploaded separately

- Project thumbnail image for project web pages (here)
- Software note: Sharing the code for our platform is not feasible given our larger goals for this product.

References

Ahn, Y., & Lin, Y. (2020). FairSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 1086-1095. 10.1109/TVCG.2019.2934262

Bottino, F., Tagliente, E., Pasquini, L., Napoli, A. D., Lucignani, M., Figà-Talamanca, L., & Napolitano, A. (2021). COVID mortality prediction with machine learning methods: A systematic review and critical appraisal. *Journal of Personalized Medicine, 11*(9), 893. 10.3390/jpm11090893

Cabrera, A. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (Oct 1, 2019). FAIRVIS: Visual analytics for discovering intersectional bias in machine learning. Paper presented at the 46-56. 10.1109/VAST47406.2019.8986948 https://ieeexplore.ieee.org/document/8986948

Constantin, R., Duck, M., Alexandrov, A., Matosevic, P., Keidar, D., & El-Assady, M. (Jan 1, 2022). How do algorithmic fairness metrics align with human judgement? A mixed-initiative system for contextualized fairness assessment. Paper presented at

the 10.1109/TREX57753.2022.00005
https://search.proquest.com/docview/2754957098

Quedado, Jay, Mashhadi, Afra and Zolyomi,Annuska. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 37, 1–7. https://doi.org/10.1145/3491101.3503568

Wiśniewski, J. (2021,). Visualizing ML model bias with dalex.
https://medium.com/responsibleml/visualize-ml-model-bias-with-dalex-b63f182cd649

Appendix

Egaleco Fairness Metrics Glossary

Demographic Parity

- **Synonyms: Disparate Impact, Statistical Parity**
- **Definition**: Equal Selection Rate (Predicted Positives/Total Predictions) between two groups
- **When to Use**: When you care about the percentage of data points that are classified as positive, independent of ground truth. Don't use it when you know that positive rates justifiably differ among groups. E.g. when screening for Glaucoma (primarily affects the elderly), or when predicting likelihood of breast cancer (more likely in women).
- **Example Application:** *Does each group have equal opportunity of achieving a favorable outcome given their gender? An AI tool that screens for the flu and predicts that 2 of the 8 men (25%) will test positive and 3 of the 12 women (25%) will test positive is achieving demographic parity.*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want the probability of your predicted outcome to be equal across groups.

True Positive Rate (TP/P)

- **Synonyms: Sensitivity, Recall**
- **Definition**: Equal probabilities by subgroup for subjects in the positive class (P) to have positive predictions.
- **When to Use**: When the resources are limited, it's important to ensure that people who truly need the resource (P) are getting it (TP) through the model prediction, and that this rate is the same across groups.
- **Example Application:** Is your likelihood of being given a limited resource like a ventilator, given that you actually need the ventilator, dependent on your race? Black individuals who need a ventilator should be equally as likely to receive a ventilator as white individuals who need a ventilator.
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, and resources are limited so you're only able to provide the intervention to a small fraction of the people who need the intervention.

False Positive Rate (FP/N)

- **Synonyms: False Alarm Rate**
- **Definition**: Equal probabilities by subgroup for subjects in the negative class (N) to have positive predictions (FP).
- **When to Use**: When people who don't need a punitive intervention are being subjected to such an intervention, it can cause harm to such individuals, so it's important to make

sure that this rate is not higher for one group than it is for another group. Also, when people who don't need an assistive intervention are mistakenly given some assistive resources, it can lead to wastage of those resources which can be undesirable in some contexts, so it's important to make sure that this rate is not higher for one group than it is for another group.

- **Example Application:** *Among the people who shouldn't have been charged higher premiums (N), what are the chances that they were incorrectly charged higher premiums (FP) given their age?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors adversely harm people, you have a punitive intervention, and you're most concerned about treating those who should not receive the intervention fairly by group.

False Negative Rate (FN/P)

- **Synonyms: Miss Rate**
- **Definition**: Equal probabilities by subgroup for subjects in the positive class (P) to have negative predictions (FN).
- **When to Use**: When someone who is in need of an assistive intervention but does not receive the intervention because of the model's prediction (FN), this can have severe consequences in the healthcare context, and so it's important to ensure that this rate is not different by subgroup.
- **Example Application:** *What are your chances of being wrongly left out of assistive resources like enrollment in a healthcare management program given your race?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, you're able to intervene with most of the people with need, and you're most concerned about treating those with need fairly by group.

Equalized Odds (TPR and FPR)

- **Synonyms: Equality of Odds**
- **Definition**: Is satisfied if both True Positive Rate Parity and False Positive Rate Parity are satisfied.
- **When to Use**: When both allocating resources fairly by group (TPR) and preserving resources fairly by group (FPR) is important.
- **Example Application:** *Are the chances that an individual is moved into the intensive care unit when they are in need of it independent of their race, and are the chances that an individual is moved into the intensive care unit when they are not in need of it also independent of their race?*
- **When it's suggested:** Egaleco suggests this metric whenever FPR Parity or TPR Parity is suggested - equalized odds takes each of these metrics a step further and enforces a stricter definition of fairness because it requires parity for both FPR and TPR.

False Omission Rate (FN/PN)

- **Synonyms:**
- **Definition**: Equal percentage of data points by subgroup that are incorrectly classified as negative (FN) out of all data points classified as negative (PN).
- **When to Use**: Among people who are not being given the resource (PN), what fraction are incorrectly predicted to not need that resource (FN)? Focuses on people not receiving assistance.
- **Example Application:** *Among people who don't receive additional health services to prevent a stroke, what are the chances that they actually had a stroke and were in need of those additional health services? Is this rate the same for men as it is for women?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors give opportunity to one group over another, you have an assistive intervention, you're able to intervene with most of the people with need, and you're most concerned about treating those who don't receive assistance fairly by group.

False Discovery Rate (FP/PP)

- **Synonyms:**
- **Definition**: Equal percentage of data points by subgroup that are incorrectly classified as positive (FP) out of all data points classified as positive (PP).
- **When to Use**: When the intervention is punitive, it's important to ensure that the likelihood of incorrectly receiving the punitive intervention (FP) is the same for different groups.
- **Example Application:** *Among the people who are charged higher health insurance premiums, what are the chances they actually had higher health costs and should have been charged higher premiums, given their race?*
- **When it's suggested:** Egaleco suggests this metric when you indicate that you want to be fair based on how the model's errors adversely harm people, you have a punitive intervention, and you're most concerned about treating those who receive the intervention fairly by group.

True Negative Rate (TN/N)

- **Synonyms: Specificity, Selectivity**
- **Definition**: Equal probabilities by subgroup for subjects in the negative class (N) to have negative predictions (PN).
- **When to Use**: When the resources are limited, it's important to ensure that people who truly don't need the resource (N) are not receiving the resource (TN) through the model prediction, and that this rate is the same across groups.
- **Example Application:** *Among healthy people, is the likelihood that they are correctly identified as not having a condition the same given their race?*

- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but it can still be useful to explore if the benefit of a true negative is higher than the cost of a false positive.

Accuracy

- **Synonyms: Error Rate**
- **Definition**: Percent of correctly predicted data points out of all data points.
- **When to Use**: When you want to check if a model performance metric that accounts for all types of errors is equal between groups.
- **Example Application:** *If we have a model that predicts the likelihood of not arriving at a doctor's appointment, is the accuracy the same for older individuals as it is for younger individuals?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but it can be useful to explore alongside the other metrics Egaleco recommended for your use case, given its ease of interpretability.

Positive Predictive Value Parity (TP/PP)

- **Synonyms: Predictive Rate Parity, Precision**
- **Definition**: Checks if the Positive Predictive Value (True Positives divided by Predicted Positives) is equal between subgroups.
- **When to Use**: When you want to equalize the chance of success, given a positive prediction (success in this case is defined as correctly predicting someone as positive when they are indeed positive).
- **Example Application:** *Of the individuals the model predicted as being at high risk for heart disease (Predicted Positive), what percent of them actually had heart disease (True Positive)? Is this percent the same for white individuals as it is for black individuals?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but given that precision (which is the same as PPV Parity) is frequently used as a traditional model performance metric this metric may be useful to explore alongside the other metrics Egaleco recommended for your use case, given its ease of interpretability.

Negative Predictive Value Parity (TN/PN)

- **Synonyms:**
- **Definition**: Checks if the Negative Predictive Value (True Negatives divided by Predicted Negatives) is equal between subgroups.
- **When to Use**: When you want to equalize the chance of success, given a negative prediction (success in this case is defined as correctly predicting someone as negative when they are indeed negative).
- **Example Application:** *Of the individuals that the model predicted as not needing to be moved into an intensive care unit (Predicted Negative), what percent of them actually*

didn't need to be moved into an intensive care unit (True Negative)? Is this percentage the same for males as it is for females?

- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but if you are particularly concerned with ensuring that when the model makes a negative prediction the model is correct, this metric may be useful to explore alongside the other metrics Egaleco recommended for your use case.

Predictive Value Parity

- **Synonyms:**
- **Definition**: Is satisfied if both Positive Predictive Value Parity and Negative Predictive Value Parity are satisfied.
- **When to Use**: When equalizing the chance of success is important for both positive and negative predictions (correctly predicting someone as positive when they are indeed positive, and correctly predicting someone as negative when they are indeed negative).
- **Example Application:** *Say we have a model that predicts the severity of someone's condition in the emergency room. Is the positive predictive value (True Positives divided by Predicted Positives) the same for men as it is for women, and is the negative predictive value (True Negatives divided by Predicted Negatives) the same for men as it is for women?*
- **When it's suggested:** Egaleco doesn't suggest this as a top metric for any use cases, but if you choose to explore PPV Parity or NPV Parity, then Predictive Value Parity can also be useful to explore because it combines both of these metrics to enforce a stricter definition of fairness.

References:

1. Fairlearn
2. Aequitas
3. FairMLHealth
4. Verma 2018