CS 6603: AI, Ethics, and Society
Homework Project #5: Fairness and Bias

Readings:

- "Data preprocessing techniques for classification without discrimination" by Kamiran and Calders (2012) [https://link.springer.com/content/pdf/10.1007%2Fs10115-011-0463-8.pdf]
- "A clarification of the nuances in the fairness metric landscape" by Castelnovo, Crupi, Regoli, Penco, and Cosentini. (2022) [https://www.nature.com/articles/s41598-022-07939-1]
- "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,"
    - https://arxiv.org/abs/1810.01943
- The What-If Tool: Code-Free Probing of Machine Learning Models
    - https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html
    - https://pair-code.github.io/what-if-tool/

A predictive learning algorithm predicts an outcome based on learning from previous instances of data. For example: Given an instance of a loan application, predict if an applicant will repay a loan based on prior financial loan application data. The learning algorithm makes these predictions based on a training dataset, where many other instances (other loan applications) and actual outcomes (whether they repaid) are provided. The model is then evaluated on a test dataset. The train and test datasets are generated by randomly splitting the original dataset.

As you have discovered, patterns found by these learning algorithms amplify historical biases. For example, a loan repayment algorithm may favor one age group compared to the other. In turn, the favored age group is presented with better loan repayment terms. Even though the outcome accurately represents the data. There are legal precedents and laws that prohibit making decisions based on an applicant's age, regardless of whether it's a valid prediction based on historical data.

To enable the mitigation of bias, several "fairness metrics" have been proposed. As Castelnovo et al. (2022) points out, fairness metrics are a way to measure and assess fairness. They are often used to mitigate bias in AI models. Recognizing bias is essential because without identifying what's flawed, it's impossible to rectify it. Identifying bias is crucial because without understanding what's flawed, you can't repair it.

In this assignment, we will look at the impact of computing and applying fairness metrics to "fix" data used to train algorithms that learn from credit-based data sets.

Step 1 – Dataset Selection
Select one of the following datasets:

- German Credit Data Set – https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
- Taiwan Credit Data Set – https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
- Portuguese Bank Marketing Data Set - https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Step 2 – Explore the Dataset
Using the dataset selected in Step 1, answer the following:

1. Which dataset did you select?
2. How many observations are in the dataset?
3. How many variables are in the dataset?
4. How many variables are associated with a legally recognized protected class under U.S. law? List these variables and any associated legal precedents/laws.

Step 3 – Defining Creditworthiness and Preparing the Dataset
For this step, you will specify an outcome variable, protected attributes, and split the dataset into training and testing sets based on your selected dataset,

Step 3.1: Dataset Outcome Variable for Approving or Denying a Loan
Table 1 below lists the outcome variable for each dataset. The outcome variable determines who is approved or denied a loan.

Dataset	Outcome Variable
German Credit Data Set	class
Taiwan Credit Data Set	y
Portuguese Bank Marketing Data Set	y

Table 1

In your report, note the outcome variable for your dataset.

Step 3.2: Formula to Determine Creditworthiness
Over the next few steps, we will come up with our own method of approving or denying a loan. This method will then be compared to the outcome variable from the dataset.

Derive a formula to score each customer based on whether they are an *Excellent Credit Risk (i.e. likely to pay back a loan)* versus *Bad Credit Risk (i.e. likely to default on loan)*. The formula should output a range of scores from 0 to 100.

To compute creditworthiness, you can apply any algorithm or set of calculations on the variables from the dataset that makes sense to you. Feel free to experiment: implement an ML algorithm to predict creditworthiness, create a mathematical formula, or come up with an algorithm that randomizes the credit score.

Note: While your formula is allowed to be open-ended, it is illegal to approve or deny credit based on an individual's protected class status. Thus, your formula should not make use of a member's age, gender, national original, race, sexuality, etc. to determine creditworthiness.

In your report, provide the formula for creditworthiness.

Step 3.3: Select Protected Class Attribute
Select a protected class attribute from the dataset – i.e. choose an attribute on which bias can occur. This attribute will be used in subsequent steps to test for bias.

Provide the protected class attribute in your report.

Step 3.4: Define Privileged and Unprivileged Groups
Using members of the protected class, define a privileged and an unprivileged group. This can be done by choosing a subset of protected attribute values which are considered unprivileged from a fairness perspective (i.e., your unprivileged group would be your historically disadvantaged group of interest).

Example: If age is selected as the protected class attribute, Older (age >= 40) would be the unprivileged group and Younger (age < 40) the privileged group.

Report the privileged and unprivileged group.

Step 3.5: Split Dataset into Train and Test Sets
Randomly split your original dataset into equally-size training and testing sets.

Report how many members of your protected class are in the privileged and unprivileged groups.

Step 4 – Maximizing Profit
In this step, you will graph and compute a default threshold that maximizes profit.

Step 4.1: Plot a Histogram for Creditworthiness
Using a histogram, graph the data associated with *Creditworthiness* (where creditworthiness is on the X-axis and the number of associated customers with that creditworthiness is on the Y-axis)

Step 4.2: Compute a Threshold for Loan Approval that Maximizes Profit
Recall in Step 3.2 the formula created to determine creditworthiness from a scale of 0 to 100. For Step 4.2, compute a threshold (i.e., a number between 0 and 100) for approving a loan (based on credit risk) that *tries* to maximize **profit**.

For this project, **profit** for each record in the dataset is calculated based on the following criteria:

- If the original dataset approves a loan and you also approve the loan, the profit earned is +10.
- If the original dataset approves a loan and you deny the loan, the profit lost is -5.
- If the original dataset denies a loan and you approve the loan, the profit lost is -3.
- If the original dataset denies a loan and you also deny the loan, no profit is earned (profit=0).

This can be summarized in table 1 below:

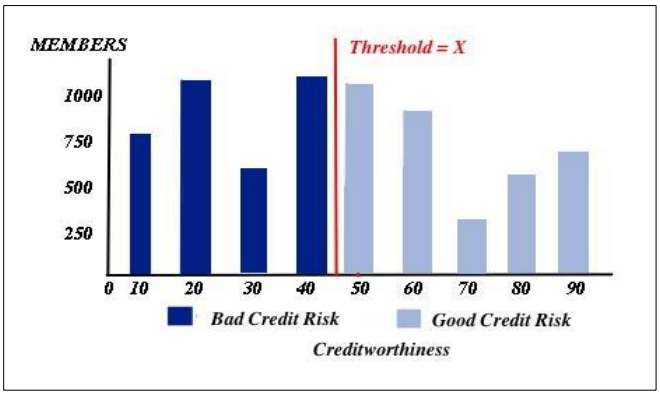
Dataset Approval Status (Outcome variable from Step 3.1)	**Creditworthiness Approval Status** (X >= Creditworthiness)	**Profit**
Approved	Approved	+10
Approved	Denied	-5
Denied	Approved	-3
Denied	Denied	0

Table 2

Report the threshold X that maximizes profit in your report.

Note: An example for calculating profit is provided in the FAQs.

Step 4.3 Plot your threshold on the Histogram from Step 4.1
On your histogram from Step 4.1, plot a line for the threshold selected. An example is shown below for illustration purposes only:



Provided for illustrative purposes only.

Step 4.4: Compute Favorable verses Unfavorable outcomes for protected class subgroups
Create a table documenting how many members in the privileged and unprivileged groups received Favorable (i.e. Approved) versus Unfavorable (i.e. Declined) outcomes based on the threshold value selected.

Step 5: Fairness Metrics

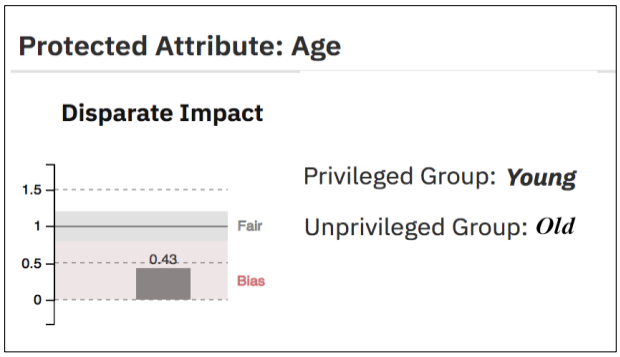
Step 5.1: Fairness Metric Selection
Based on your protected class attribute and the privileged and unprivileged groups, select ***two*** different fairness metrics (as defined in either the AI Fairness 360 Toolkit or What-If Tool).

Compute the differences between privileged and unprivileged groups in your training set.

Note: You can code up your own mathematical formulations, modify open-source code that wasn't developed for this course, or use the python functions provided by the Toolkit(s) directly.

For example, if using Disparate Impact as the fairness metric, the ratio of the rate of favorable outcome for the unprivileged group versus the privileged group would be computed. The ideal value for the Disparate Impact metric is 1.0. A value < 1 implies higher benefit for the privileged group and a value >1 implies a higher benefit for the unprivileged group.

Step 5.2: Plot Fairness Metrics
Graph the result for both fairness metrics (indicating the fair/bias thresholds).



Note: Provided for illustrative purposes only

Step 5.3: Document your Fairness Metrics
In a single table, include the following in your report:

1. The two fairness metrics selected.
2. The computed differences between the privileged and unprivileged groups for each fairness metric.
3. The acceptable ranges for the metric selected.
4. Indication of whether the fairness metric indicates bias for each metric.

Step 5.4: Analyze Bias with the Fairness Metrics Selected
In your report, discuss if the differences for the fairness metrics indicate bias for or against the privileged or unprivileged groups. The discussion should be a minimum of one paragraph reflecting on the results produced and reasoning.

***Note**: Step 6 and Step 7 must be completed, even if Step 5 did not indicate bias.*

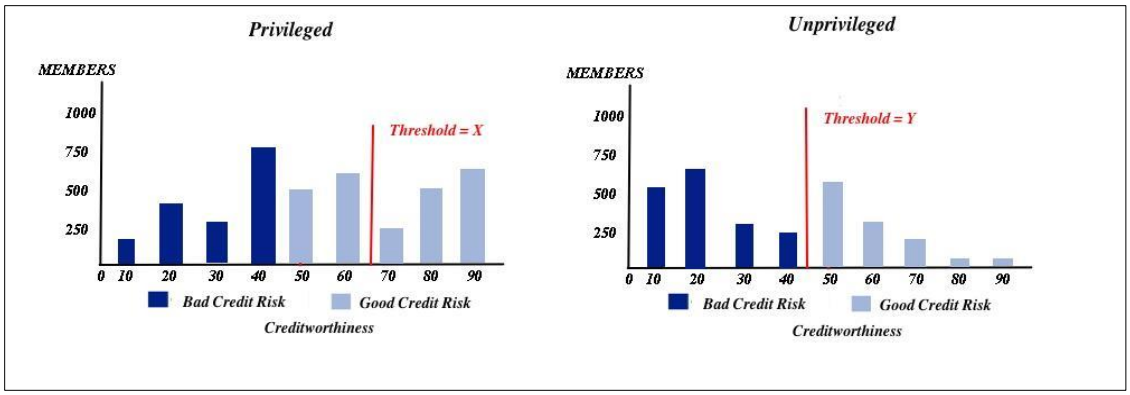
Step 6 – Mitigate bias in the training set
In this step, you will try to mitigate bias in the training dataset by selecting a different creditworthiness formula or threshold for Privileged and Unprivileged groups.

Step 6.1: Select Different Thresholds for Privileged and Unprivileged Groups
Define different threshold values for approving a loan even if they are considered a bad credit risk (Recall, in Step 4, we assumed that a good credit risk is associated with a creditworthiness score >=X).

Step 6.2: Find Threshold Values that Minimize Bias for Privileged and Unprivileged Groups
Select one of the two fairness metrics used in Step 5.

For the privileged and unprivileged group, find threshold values that minimize the differences in approval rates between the two groups based on the fairness metric selected while also maximizing profit based on the profit formula in Step 4.

Step 6.3: Plot Histograms
Graph the histograms associated with *Good Credit Risk* versus *Bad Credit Risk* as a function of your protected class attribute. Highlight the thresholds for the privileged and unprivileged groups on each graph.



Note: Provided for illustrative purposes only.

Step 6.4: Bias Mitigation Results
Answer the following in your report:

1. What are the threshold values for the privileged and unprivileged groups?
2. What is the profit based on your threshold values?

Step 6.5: Document your results
Create a table documenting how many members in the privileged and unprivileged groups received Favorable (i.e., Approved) versus Unfavorable (i.e. Declined) outcomes based on the new threshold values.

Step 7 – Post Bias Mitigation Analysis
Discuss the following questions in 2-3 paragraphs:

1. For each of the fairness metrics selected in Step 5, discuss if there were any differences in the outcomes for the privileged versus unprivileged group?
2. Was the mitigation step in Step 6 effective and for whom? Did any group receive an advantage? Was any group disadvantaged by the mitigation step?
3. Identify any issues that would arise if this method was used to mitigate bias. Justify your reasoning.

Step 8: Submission
Turn in a report (in PDF format) documenting your outputs for each question. The report should follow the JDF format. You can find a link to the JDF template here: JDF Templates.

We suggest using the Microsoft Word template for proper formatting and styling. Reports that are not neat and well organized will receive up to a 10% deduction.

The file name for submission is GTuserName_Assignment_5, (ex. gBurdell3_Assignment_5.pdf). Deductions will be made if your file name is not submitted correctly.

All charts, graphs, and tables should be generated in Python or Excel, or any other suitable software application.