Cs 6603: AI, Ethics, and Society
Final Project

The goal of the final project is to explore bias mitigation techniques. You will train a model on two different datasets: one without any bias mitigation and one with a bias mitigation technique applied. Fairness metrics will be computed on both datasets to see if the bias mitigation was effective.

Group Requirements

For the final project, you may work in "groups" of 1 to 4 individuals.

Teams of Size 1:
If working individually, you are still required to fill out the Final Project Group Members Survey.

For groups of 1, Step 5 (Analysis) is not required.

Teams of Size 2 to 4 Individuals:
If working in a group, ONE member of the group must submit the Final Project Group Members Survey. The names and GT usernames of all students within the group are required to complete the survey.

When a member of your group submits the final project, that submission will show up for ALL members of the group. *Note: The last version of the project submitted before the due date by any member of the team will be the version graded for all team members.*

For groups of 2 - 4 members, Step 5 (Analysis) is required.

If you are interested in joining a group, please see the Final Project Teammate Search thread on Ed Discussion.

Step 1: Dataset Selection

For this project, you may select 1) any dataset from the UCI machine learning repository - https://archive.ics.uci.edu/ml/index.php, 2) any dataset from Kaggle - https://www.kaggle.com/datasets, or 3) any dataset openly provided by an organization, preferably non-profit, that could benefit from this analysis. The dataset selected must satisfy the following criteria:

- Must have a sample size of at least **500 observations**
- Must have at least **two variables** belonging to a legally recognized protected class
- Must have at least **two dependent variables** (outcome variables) that could result in favorable or unfavorable outcomes.
  Note 1: Use your subjective opinion based on the discussions we've had in class.
  Note 2: You may also derive a new dependent variable if it comes directly from the dataset itself (i.e. you create the variable approved using the independent variables credit_score and good_payment_history "approved = credit_score > 600 and good_payment_history").
- Must be related to one of the regulated domains: *Credit, Education, Employment,* or *Housing and 'Public Accommodation'*

  Note: Loosely, any dataset that could have potential bias in outcomes based on protected class membership is acceptable. Don't be biased by how the dataset is labeled/organized –you can think creatively about how to structure the dataset so it's compliant to project requirements.

- You are **not** allowed use any dataset used in any of our previous assignments, case studies, written critiques, or exercises. For a list of datasets NOT allowed in the final project, please see the table below (or refer to the FAQs).

Dataset Name	URL
Mental Health in Tech Survey: Survey on Mental Health	https://osmihelp.org/research/
in the Tech Workplace in 2019	https://health.data.ny.gov/dataset/Hospital-Inpatient-Discharges-SPARCS-De-Identified/22g3-z7e7
Information on deaths that occur in custody or during the process of arrest in California	https://openjustice.doj.ca.gov/data
toxicity-per-attribute (unintended-ml-bias-analysis)	https://github.com/conversationai/unintended-ml-bias-analysis/tree/master/data
Google News dataset	https://code.google.com/archive/p/word2vec/
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
Compas (ProPublica recidivism)	https://github.com/propublica/compas-analysis
Adult census income	https://archive.ics.uci.edu/ml/datasets/adult
UCI Census Income Dataset	http://archive.ics.uci.edu/ml/datasets/Census+Income
Compas dataset	https://www.kaggle.com/datasets/danofer/compass
IBM HR Analytics Employee Attrition & Performance	https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Provide answers to the following questions in the final project report:

1. Which dataset was selected? Please include the link to the dataset.
2. Which regulated domain does the dataset belong to?
3. How many observations are in the dataset?
4. How many variables are in the dataset?
5. Which variables were selected as the dependent/outcome variables?
6. How many and which variables in the dataset are associated with a legally recognized protected class?
7. Which legal precedence/law (as discussed in the lectures) does each protected class fall under?

Step 2: Explore the Dataset

1. Identify the members/subgroups associated with the protected class variables in the dataset.

   In your report, in table format document the protected classes and their corresponding subgroups.

2. Discretize the subgroups from Step 2.1 into discrete numerical values.

   Provide these mappings in table format in your report.

3. Select two protected classes from the dataset. You will use these protected classes in the rest of your project.

4. For each protected class, create frequency tables documenting the number of members in each subgroup associated with each dependent variable selected in Step 1.5.

   Include the frequency tables in your report (1 for each protected class/dependent variable combination).

5. For each protected class, create a bar chart the graphs the frequency of each subgroup as a function of the dependent variables identified in Step 1.5.

   Include the bar charts in your report (1 for each protected class/dependent variable combination).

Step 3: Fairness Metric Selection and Mitigating Bias with PreProcessing Algorithms
For the next set of questions, you are allowed to code up your own mathematical formulations, modify open-source code that wasn't developed for this course, or modify code found from the AI Fairness 360 Open Source Toolkit (https://aif360.mybluemix.net/) or the What-If Tool (https://pair-code.github.io/what-if-tool/) to work with your dataset.
Note: Others have found it easier to create their own formulas based on the fairness definitions found in the class lectures or on the toolkit website rather than modifying the code in the AI Fairness or What-If Tool packages.

1. In your report, identify the privileged and unprivileged groups for each protected class in your dataset.

2. Select two fairness metric algorithms. For each protected class in the dataset, compute the two fairness metrics selected for the privileged and unprivileged groups as a function the two dependent/outcome variables identified in Step 1.5. Include your results in a table format.

   *Note: Your table will have 8 values ([2 fairness metric algorithms] * [2 protected classes] * [2 outcome variables]).*

3. Select a pre-processing bias mitigation algorithm to transform the original dataset (e.g. Reweighting, Disparate Impact Remover, etc.) as a function of one of your dependent variables.

4. Using the two fairness metrics identified in 3.2, compute fairness metrics on the transformed dataset. Include your results in a table format.

   Note: In Step 3.4, you will calculate values for the same 2 fairness metric algorithms from Step 3.2. You will have 8 values again.

Don't forget to include the following in your report:

- The code for Step 3 (please include in your. ipynb file)

Step 4: Mitigating Bias
Like Step 3, Step 4 allows you to code up your own algorithm, modify open-source code that wasn't developed for this course, or modify code found from the AI Fairness 360 Open Source Toolkit to work with your dataset (https://github.com/IBM/AIF360/tree/master/examples). For example, code for training a classifier based on a credit scoring example can be found here:
https://github.com/IBM/AIF360/blob/master/examples/demo_reweighing_preproc.ipynb**.**

Note: Others have found it easier to create their own algorithm rather than modifying the code in the AI Fairness package.

Recall that in Step 3.3, the transformed dataset was created by applying a **bias mitigation** algorithm to the original dataset. The goal of Step 4 is to see if the bias mitigation in Step 3.3 is effective.

Original Dataset:

1. Randomly split the original dataset into training and testing datasets.
2. Train a classifier using the original dataset; select one of your dependent variables as the output label to train your classifier.
3. Using the testing dataset, compute the same fairness metrics selected in Step 3.2 for the privileged and unprivileged members of the two protected classes identified in Step 2.3.

Transformed Dataset:

4. Randomly split your transformed dataset (from Step 3.3) into training and testing datasets
5. Train a classifier using the transformed training dataset; select one of your dependent variables as the output label to train your classifier.
6. Using the testing dataset, compute the same fairness metrics selected in Step 3.2 for the privileged and unprivileged members of the two protected classes identified in Step 2.3.

For each fairness metric, in table format, indicate if there were any differences in the outcomes for privileged versus unprivileged group.

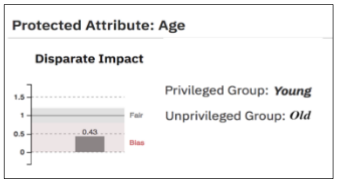
In table format, answer the following question: Was there a positive change, negative change, or no change for each fairness metric after transforming the dataset? Also include an answer to the following: Was there a positive change, negative change, or no change on each fairness metric after training the classifier - with respect to the original testing dataset and the transformed testing dataset? [*Note: The FAQs provide an example of what this table looks like.*]

For Step 4, don't forget to include the following in your report:

- The code for Step 4 (please include in your .ipynb file)

Step 5: Analysis

- If you are an individual (team of 1), *Provide the following in the final project report*:
  - Step 5: I am a team of one
- If you are a team > 1, *Provide the following in the final project report*:
  - List the members of your project team
  - Graph the results from applying the two fairness metrics on your privileged/unprivileged groups as derived from Step 3.2, 3.4, and 4.5
  - Explain which fairness metric (if any) is best and provide a justification for your answer.
  - Each team member must provide a separate answer to the following questions in at least a one-paragraph response (this is to be included in the submitted group report).
    - Did any of these approaches seem to work to mitigate bias (or increase fairness)? Explain your reasoning.
    - Did any group receive a positive advantage?
    - Were any group disadvantaged by these approaches?
    - What issues would arise if you used these methods to mitigate bias?



Step 6: Submission

Turn in a final PDF report in JDF format with all information requested. The report should be called *'GTuserName_Final_Project_Report.pdf.'* Also submit all code for the assignment in a **single** Jupyter notebook titled *'GTuserName_Final_Project.ipynb'*. GTuserName is the Georgia Tech username of the group member that submitted the report.

Reports that are not neat and well organized may receive up to a 10-point deduction. All charts, graphs, and tables should be generated in Python or Excel, or any other suitable software application.

Please note that, when submitting a Jupyter notebook (.ipynb) for the assignment submission, you need to make sure you have **clearly printed/displayed all outputs necessary** to receive full credit (as you would for a PDF submission) before submitting. This means that the Jupyter notebooks must be run before your submission. Credit would not be awarded for just submitting the code in the notebook and not displaying the output.