AI, Ethics, and Society

Final Project

In this assignment, you can work independently or in teams of (no more than) 4 students. **If you are working on the project as a team**, only one person needs to submit this assignment. Make sure to coordinate who is submitting it, however. If you choose to work in a group, there is one additional set of questions for the team provided at the end.

Please **read** the following instructions **very carefully** before starting your project.

- Submit the Python Notebook named '***GTuserName*_Final_Project**'.ipynb.
  - The report titled '***GTuserName*_Final_Project_Report**'.pdf in JDF format is required in addition to submitting '*GTuserName*_Final_Project.ipynb'. When submitting the *.ipynb file make sure you run the code and the outputs for steps 3 and 4 are clearly labeled and visible per project requirements. The 'GTuserName' is the user name of the submitter.
    - The PDF document (report) should be submitted in JDF format in addition to *.ipynb file.
    - The*.ipynb file for Steps 3 and 4 is Required, but all other steps it is optional and can be submitted using a single *ipynb file, and this will be in addition to your PDF final report file.
- Do **not** submit python files (.py) for your code.
- You **can** derive a **new dependent variable** if it comes directly using inputs from the dataset itself - i.e., it can't be imaginary data.
- You should **not** 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 link below:**
  - **https://docs.google.com/spreadsheets/d/12u_VryQu3aR4-sFkPEo1Sg3x92IhbpxQLjmN5KiDs_o/edit?usp=sharing**
- You should only use **Python** as your programming language (not R or any other language)
- When submitting the final project code (for Steps 3 and 4 or for the whole project) as *ipynb file it should be **a single *.ipynb** file (for example, you should not submit separate files for each step).
- During submission of the final project in **canvas**, all students are required to include in the **comments section** their **team members**, or **for a one-member team**, include "Team of one." **Please note:** To type in the comments section in canvas, you may either create a file (*.Doc or *.PDF) with a list of members (as an attachment file) or directly add team members in the comments section.
- **Not** following the above instructions will result in the **deduction of points.**

Step 1: You may select 1) any dataset from the 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. Your selection must be based on the following characteristics [*Note: on Kaggle – many of the datasets provide links to the original dataset such that you do not have to set up a new Kaggle profile. Kaggle was acquired by Google in 2017*]

1. Must have a sample size of at least **500 observations**
2. Must have at least **two variables** belonging to a legally recognized protected class
3. Must have at least **two dependent variables** (outcome variables) that could result in favorable or unfavorable outcomes [*Note: Use your subjective opinion based on the discussions we've had in class*]
4. 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. Also, 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 the requirements]

Answer the following questions in the final project report:

- Which dataset did you select?
- Which regulated domain does your dataset belong to?
- How many observations are in the dataset?
- How many variables are in the dataset?
- Which variables did you select as your dependent variables?
- How many and which variables in the dataset are associated with a legally recognized protected class? Which legal precedence/law (as discussed in the lectures) does each protected class fall under?

Step 2:

1) Identify the members associated with your protected class variables and group together into a subset of membership categories as appropriate
2) Discretize the values associated with your dependent variables into discrete categories/numerical values as appropriate
3) Compute the frequency of each membership category associated with each of your protected class variables from Step 2.1
4) Create a histogram for each protected class variable that graphs the frequency values of its membership categories as a function of the dependent variables

Provide the following in the final project report:

- Table documenting the relationship between members and membership categories for each protected class variable (from Step 2.1)
- Table documenting the relationship between values and discrete categories/numerical values associated with your dependent variables (from Step 2.2)
- Table providing the computed frequency values for the membership categories each protected class variable (from Step 2.3)
- Histograms derived from Step 2.4

Step 3: 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) Based on your dataset, identify the privileged/unprivileged groups associated with each of your protected class variables
2) For each protected class variable, select two fairness metrics and compute the fairness metrics associated with your privileged/unprivileged groups as a function of each of your two dependent variables. You may choose any reasonable threshold in order to generate a baseline for comparison using the fairness metrics.
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) Use the two fairness metrics identified in 3.2 and compute the fairness metrics on the transformed dataset

Provide the following in the final project report:

- Provide the resulting code (can be as an additional .ipynb file if submitting a PDF)
- Provide a table documenting the protected class variable selected, the privileged/unprivileged groups/values, the pre-processing bias mitigation function selected, and the fairness metrics/resulting values computed in Step 3.2 and Step 3.4

Step 4: There are two options for Step 4 – Choose one to complete for the final project. Specify which option you are choosing in your final report.

Option A: For the next set of questions, you are allowed 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.

1) Randomly split your original dataset into training and testing datasets
2) Randomly split your transformed dataset into training and testing datasets (from Step 3.3)
3) Train a classifier using the original training dataset from Step 4.1; select one of your dependent variables as the output label to train your classifier.
4) Train a classifier using the transformed training dataset from Step 4.2; select one of your dependent variables as the output label to train your classifier.
5) Select the privileged/unprivileged groups associated with one of your protected class variables (from Step 3.1); Use the two fairness metrics identified in Step 3.2 and and compute the fairness metrics for the classifier output associated with the original testing dataset and the transformed testing dataset
6) For each fairness metric, in table format, indicate if there were any differences in the outcomes for the privileged versus unprivileged group. Was there a positive change, negative change, or no change on that fairness metric after transforming the dataset (from Step 3.4)? Was there a positive change, negative change, or no change on that fairness metric after training the classifier - with respect to the original testing dataset and the transformed testing dataset? [*Note: Use your subjective opinion*]

Provide the following in the final project report:

- Provide the resulting code (can be as an additional .ipynb file if submitting a PDF)
- Document 1) the privileged/unprivileged groups, 2) the dependent variable, 3) the quantitative results from applying the two fairness metrics on the classifier output associated with the original and transformed dataset, 4) a table documenting whether there was positive, negative, or no change in each of the fairness metrics after transforming the dataset, after training the classifier on the original dataset, and after training the classifier on the transformed dataset.

Option B: For the next set of questions, you are to design your own bias mitigation algorithm

1) Design your own bias mitigation algorithm (must be different than ones already represented in the aif360.algorithms.preprocessing class) to transform your original dataset [*Note: Provide sufficient comments in your code so that the algorithm/math can be deciphered*]
2) Randomly split your original dataset into training and testing datasets
3) Apply your bias mitigation algorithm to your original testing dataset in order to produce a transformed testing dataset; select one of your protected class variables and one of your dependent variables as the output label
4) Randomly split your transformed dataset into training and testing datasets
5) Select the privileged/unprivileged groups associated with your protected class variables; Use the two fairness metrics identified in Step 3.2 and compute the fairness metrics on the original testing dataset and the transformed testing dataset
6) For each fairness metric, in table format, discuss if there were any differences in the outcomes for the privileged versus unprivileged group. Was there a positive change, negative change, or no change on that fairness metric after transforming the dataset earlier in Step 3.4? Was there a positive change, negative change, or no change on that fairness metric after applying your bias mitigation algorithm - with respect to the original testing dataset and the transformed testing dataset? [*Note: Use your subjective opinion*]

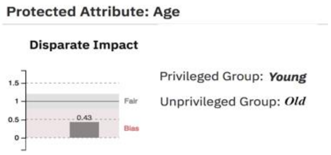
Provide the following in the final project report:

- Provide the resulting code (can be as an additional .ipynb file if submitting a PDF)

- Document 1) the privileged/unprivileged groups, 2) the dependent variable, 3) the quantitative results from applying the two fairness metrics associated with the original testing dataset and transformed testing dataset after bias mitigation, 4) a table documenting whether there was positive, negative, or no change in each of the fairness metrics on the transformed dataset from Step 3.4, the original testing dataset (Step 4.5), and on the transformed dataset after applying your bias mitigation algorithm (Step 4.5)

Step 5:

- 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 no-more than a one-paragraph response (this is to be included in the submitted group report, with a reference to the student author). Note: If a group member fails to provide a response, the team is free to indicate that in the final report submission with *No Response* and a reference to the student author.
    - Did any of these approaches seem to work to mitigate bias (or increase fairness)? Explain your reasoning. Did any group receive a positive advantage? Was any group disadvantaged by these approaches? What issues would arise if you used these methods to mitigate bias?



Step 6: - Turn in a final PDF report in JDF format with all information requested plus the code associated with the final project as a jupyter notebook. Reports that are not neat and well organized will 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, else appropriate points will be deducted, which could be the maximum.

Please review the Canvas instructions for guidance on submission. If you are working on the project as a team, only one person must submit this assignment. 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 for receiving 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. Note that you need to submit a final report as a PDF in JDF format in addition to the jupyter notebook.