Exploring Bias in Datasets

Group 5

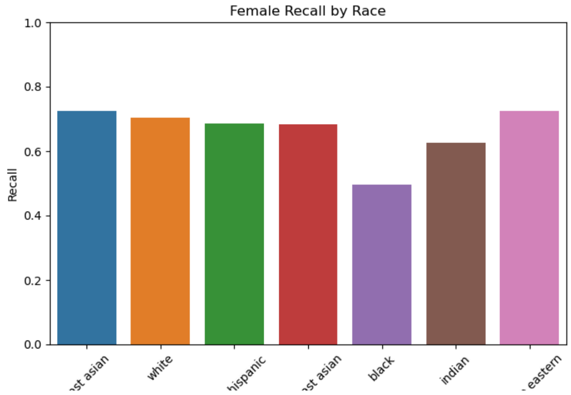
Introduction:

For this group project which covered exploring bias within data, our group discussed at length several points discussed in class. The definition of bias, according to Oxford Languages, is a systematic distortion of a statistical result due to a factor not allowed for in its derivation. Bias within data can come to exist for several reasons, as we have learned in our readings for this course up to this point. The largest reason being issues with representation of certain groups found throughout the dataset. For example, if one group is represented at a greater rate, or overrepresented, compared to another group, a model that is trained on that data may have an easier time correctly classifying that group into its correct category, such as gender or race. At the surface level, this may seem harmless. “Oops, our model incorrectly classified an African American female as male, but we can always tune parameters and increase model accuracy.” Unfortunately, this isn’t always the case. If a model *consistently* incorrectly classifies certain groups, it can learn to accept these biases and even exploit them to obtain a higher accuracy. When models do this, results can demonstrate the underlying biases which are not beneficial for research meant to be published. Additionally, these types of bias are hard to catch as overall accuracies or proportions can hide these biases sometimes. Its best ethical practice is to examine how much each group is represented within the training data prior to any type of statistical modeling. What can be worse is when these biases go unnoticed and these types of biased models are utilized for impactful decision-making. The situation then becomes drastic as people’s livelihoods can be affected by a model which is knowingly making decisions off of existing bias. For this project, we also would like to address the question as to why this happens. The best answer we have found from our reading is that machine learning models do exactly what is asked of them to the best of their ability. In turn, this means if bias exists within the training data, a machine learning model will find it, exploit it, and improve its accuracy one way or another.

Research: When discussing what we should do we went through different data sets, this one FairFace data set, resume dataset, and insurance claims datasets. We went through a vote and went with the FairFace face attribute data set because it went along with readings that we did in the *The Alignment Problem* book. We chose this dataset because it does real world implications of racial and gender bias in law enforcement and national security in using facial recognition software. From the readings we decided not to use any Neural Network modeling, due to time constraints and to complex, and just do KNN and Decision Tree analysis for the data. The known bias in the data is overrepresentation of white male faces, underrepresentation of women, people of color, and non western faces, and High error rates for women and people of color. For the analysis the issue arose of how to load the images which are in jpg to jupyter notebooks. Embedding is changing the images into numbers that represent certain pixels from the photo, any feature analysis would become redundant. It is the same as passing the photos in a convolutional neural network where it embeds in the last layer. The embeddings are labeled as features to differentiate between the columns for the analysis.

Methods:

The FairFace dataset is a dataset that came from kaggle. It is a face attribute dataset for balanced race, gender, and age. We uploaded this dataset and conducted an initial exploratory data analysis to assess its raw structure and contents. After the initial review, we cleaned the dataset by removing the 'service_test' column which essentially denoted whether the image was used in the original FairFace experiment. After this was done, race and gender text formatting was standardized across both training and testing sets. For the last step of the data cleaning, it was verified that no missing/null values or duplicate records exist. Our next step was to check the normalized class imbalance for race, gender, and their intersections. To look at how race and gender were distributed in the dataset, we created a table that showed the proportion of males and females within each race group. From this visual, we could clearly see which gender was not represented as well as the other across each race within the data. Then, for a more visually appealing graphic, we decided to make a bar chart to visually compare gender differences across different races. It was clear to the group at this point that indeed bias existed within our chosen data. As we learned in the readings, any machine learning model that is trained on data with representation issues is prone to the many inherent biases that are present in our world. To exemplify intersectional relationships, a new column was created which combined the race and gender attributes into one 'race_gender' which allowed for the analysis of patterns across different demographics. We looked at how the data was spread out across gender, race, and combinations of both to see which groups were underrepresented. Then, we used a pretrained convolutional neural network for image classification, ResNet-18 model to turn each face image into a set of numerical features by removing its final classification layer. These feature vectors were combined with race and gender labels to create clean datasets that could be used for machine learning models such as decision trees and KNN.



For the decision tree the methods used were using the embeddings for the model then having two separate models, one for gender and one for race. Using only the training embedding to do the analysis for the time allotted was not enough to do the test embeddings as well. Split the data into 80% training and 20% test size with a random state of 42 for each model. The only cleaning I had to do was for the y value for race, I had to recode the races into numeric form to use all in the model. Then did a full evaluation of each model with finding accuracy, precision, recall, and f1 scores. 5 fold cross validation was used to make sure the model is accurate as it can be for modeling. For visualizations made decision trees, confusion matrices, then have the results into their own specific dataframe for better analysis. The final method used was a correlation on the embedding to see if there was any specific feature(vector) that was being used for the model more than others. To make some of the code with multiclass scoring, AI was used to help make the code blocks and debug code using ChatGPT and Claude.

Findings:

The dataset showed some clear imbalances across gender, race, and age. There's about 5,000 more male than female images, which could influence how well models perform for each gender. We also saw race distribution was uneven, with white individuals being the most represented and Middle Eastern being the least represented groups. The age distribution showed that people between 20-39 made up about half of this dataset. When looking at the combinations of race and gender, some groups were noticeably underrepresented, most especially Middle Eastern Females.

Decision Tree Model

When predicting gender, the decision tree model only had an accuracy score of 65.2% of predicting the gender correctly. However the precision, recall, and f1-score showed the imbalance of the genders in the training sets. When classifying males, the model had a more accurate precision of 68% but when it came to classifying female counterparts, the model only achieved 62% because the representation is imbalanced with males having 9244 and females having 8105 in the training set. From the confusion matrix, it was clear how much of an impact having an imbalanced training set can have on false positives of 3126 and false negatives of 3126. False positives occur when the model incorrectly classifies an image to a certain gender and a false negative occurs when the model incorrectly labels an image as outside of a certain gender. When it came to classifying race, after examining the target and displaying the value counts, i5 was apparent that there was a very big underrepresentation of people of color. The largest group, people of white ethnicity, had the largest available data with 16527 entries while people of the middle eastern group only had 9216 entries. Getting a little more in depth, the accuracy score of the model for this small group was very low at 21.4% and the precision, recall, and f1-score seemed to increase for ethnicities with more values. The model had a hard time classifying race because it was trying to determine race by skin color. This is backed up with many values being misclassified, for example someone Indian had been classified as African American or even southeast Asian. This is also backed up in the confusion matrix where the prediction misclassified middle eastern more as Black and southeast Asian because it was so underrepresented in the data. The same thing goes for African Americans where they misclassified as either east asian and middle eastern. Final analysis was used in making a correlation matrix for each model to see if any pairs of features (vectors) were used more than the other and the results were the same for both with the top pair of feature 241 and 233 having the highest correlation of 67%. If we had the time to decode the embeddings of what this feature was, we could see what the model was looking at for the predictions. The patterns remained consistent after preprocessing the data and embedding extraction, indicating that the imbalance comes from the original dataset rather than the data preparation steps.

KNN and Logistic Regression Analysis

In this part of the analysis, we focused on checking whether the gender prediction performance actually changes across different demographic groups. We broke the results down by race and then by race combined with gender to see if bias becomes more visible at those levels.

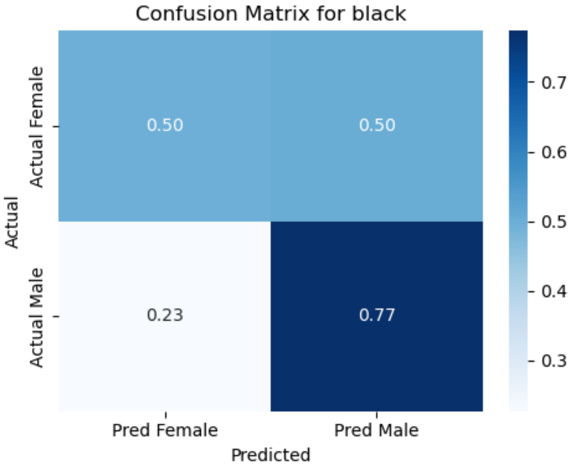
We decided to compare Logistic Regression and KNN because they behave differently. Logistic Regression is more structured and linear in how it separates classes, while KNN depends on distance and local neighborhood patterns. We wanted to see if these differences in modeling style would affect fairness outcomes.

Logistic Regression achieved 77.1% overall accuracy, while KNN achieved 75.1%. On the surface, both models look decent. Logistic Regression showed fairly balanced precision and recall across genders. KNN, however, leaned more toward males, showing higher recall for males and lower recall for females. That was the first sign that performance was not entirely symmetric.When we broke accuracy down by race, the disparities became more obvious.

- For Logistic Regression, the Black subgroup had the lowest accuracy at 0.6928, while the Middle Eastern subgroup had the highest at 0.8404. That gives a gap of 0.1476.
- For KNN, it was very similar. Black individuals had the lowest accuracy at 0.6671, while Middle Eastern individuals reached about 0.8263, giving a gap of 0.1592.

Looking at both models side by side, it is clear that Black individuals were predicted less accurately compared to other racial groups. A gap of about 15 to 16 percent is not small. That level of difference means the model is more reliable for some groups than others.The differences became even stronger when I looked at intersectional results. When race and gender were combined, the gaps widened.

Logistic Regression had an intersectional gap of 0.2293. KNN had a much larger gap of 0.391. Under KNN, females had a lower accuracy rate with black females rate sitting below 50%, while some male subgroups were above 80%. That is almost a 40 percent difference, which is quite significant. It shows that bias becomes more severe when overlapping identities are considered.



One important thing to mention is that Black males and females were not underrepresented in the dataset. Their sample counts were similar to other groups. So this cannot simply be explained by lack of data. It suggests that something deeper is happening in how the embedding space separates facial features across demographics.

Implications:

Because some race and gender groups are underrepresented, machine learning models trained on this data may perform better for some groups than for others. This could lead to bias predictions, especially for the underrepresented groups. This dataset in particular showed some clear imbalances across gender, race, and age. There were more male images than female images and far more lighter individuals than those with darker skin tones which would influence how well the model performs. The imbalances found in this dataset connect closely to the types of bias discussed in *Understanding data bias,* bias is not only caused by unequal sample sizes but also by missing or incomplete representation of certain groups which can lead to unfair modeling predictions. This idea is also reflected in *The Alignment Problem,* which explained how poor representation and lack of transparency in machine learning systems can result in real world harm if these issues are not addressed. In addition the *Datasheets for datasets* framework stresses the importance of clearly documenting dataset limitations so users understand potential risks and biases before using the data. All of these readings support the need to recognize and address demographic imbalances early in the modeling process to promote fairness and accountability and to avoid situations where people are deemed high risk for loans or high risk to be repeat offenders based on biased models from underrepresentation.

Conclusion:

To conclude this project report and demonstrate what we have learned. It is understood that bias exists inherently throughout our world, including in datasets used to train machine learning models. These models can indeed be a helpful tool in exposing these biases especially to professionals of a specified field. However, these models should never be trained on biased data with the intent of interfering or making decisions because the biases that exist will continue to be proliferated by model performance. Additionally, these decisions will negatively affect the groups of people who receive the brunt end of these biases.