Dependable AI

Assignment Report

Roll Number	B21AI055, B21CS092
Name	Drithi Davuluri, G Mukund
Assignment Number	04
Assignment Title	Bias Detection

Objective

The CelebA dataset is a widely used face attribute dataset that serves as a benchmark in various computer vision tasks. Like many real-world datasets, CelebA is not immune to biases. In this report, we aim to detect and mitigate biases present in the dataset, focusing on gender bias.

Data Processing

Dataset Overview

The CelebA dataset consists of celebrity images with annotations for various attributes such as 'Male', 'Young', and more. The dataset contains 202,599 images with 40 binary attribute annotations for each image.

Sample Images

To gain an initial understanding of the dataset, we sampled four images and visualized them. The images show varying attributes such as gender ('Male'), age ('Young'), and more.

Bias Detection

- *Imbalance Ratio*

  The first step in bias detection was to calculate the imbalance ratio between the underrepresented and overrepresented groups. In this case, the 'Male' attribute was considered underrepresented.

  Result: Imbalance Ratio: 0.71 : Indicates a substantial imbalance in the 'Male' and 'Female' categories, with 'Male' being underrepresented.

- *Chi-Squared Test*

  To statistically validate the gender distribution imbalance, we conducted a Chi-Squared test. The null hypothesis was that the gender distribution is equal.

  Chi-Squared Test Statistic: 5615.92
  P-value: 0.0000

  A high Chi-Squared value suggests a significant difference between observed and expected frequencies, indicating a biased gender distribution.

- *Mean Differences*

  We calculated the mean differences between the two gender groups for each attribute to identify which attributes are biased towards one gender. Positive values indicate bias towards 'Male', while negative values indicate bias towards 'Female'.

- *Variance Ratios*

  The variance ratio provides insight into the variability of each attribute between the underrepresented ('Male') and overrepresented ('Female') groups.

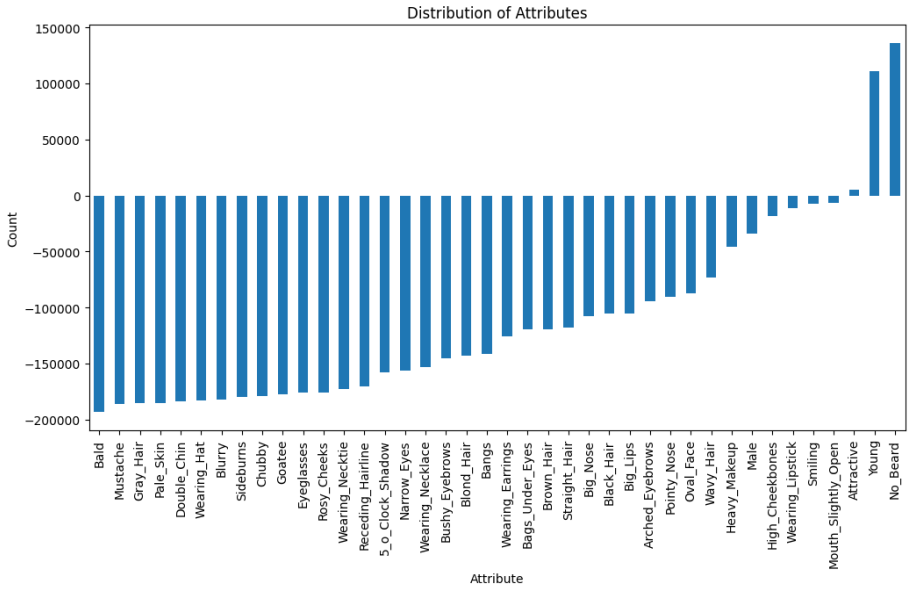
- *Correlations with Gender*

  Positive values indicate a positive linear relationship with 'Male', while negative values indicate a negative linear relationship.

Visualizing Detected Biases

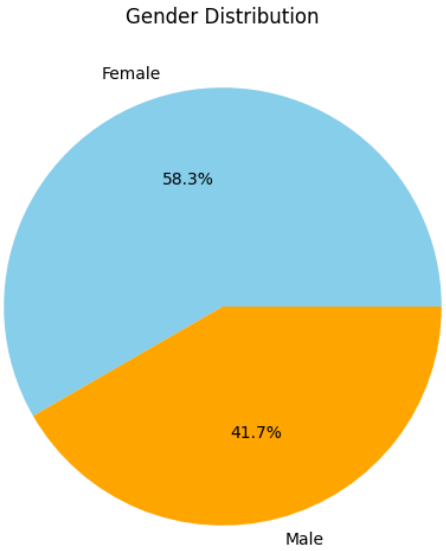
Visualize Attribute Distribution

We visualized the distribution of each attribute to gain an understanding of the dataset's composition. This helps us identify which attributes are more prevalent and which are less represented, thereby highlighting potential biases. The bar chart provides a clear visual representation, with attributes sorted by their counts.



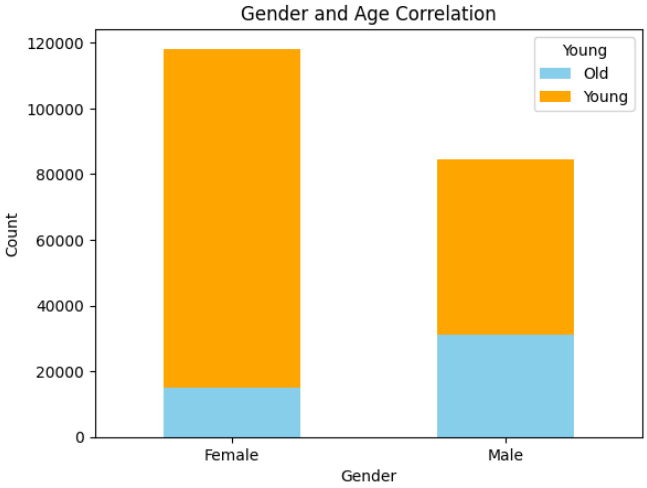
Visualize Gender Distribution

Understanding the gender distribution within the dataset is crucial as it helps us assess whether there is an imbalance between male and female samples. The pie chart clearly shows the proportion of male and female samples, providing a quick overview of the dataset's gender distribution.



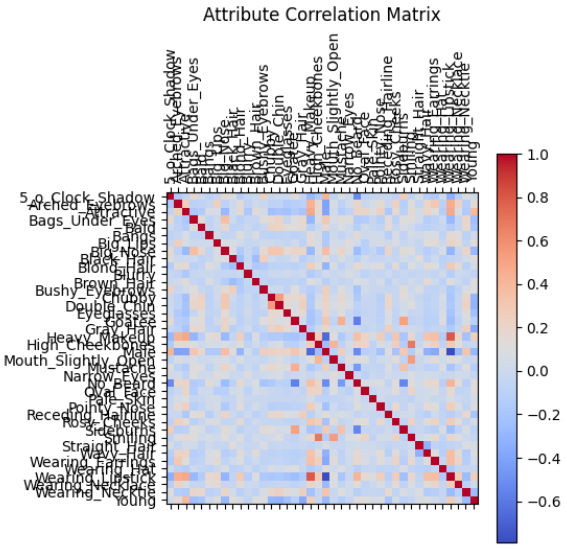
Visualize Gender and Age Correlation

Analyzing the correlation between gender and age is important to identify any potential biases related to age representation across genders. The stacked bar chart presents the count of young and old individuals for each gender, allowing us to observe if there are any significant differences in age distribution between males and females.



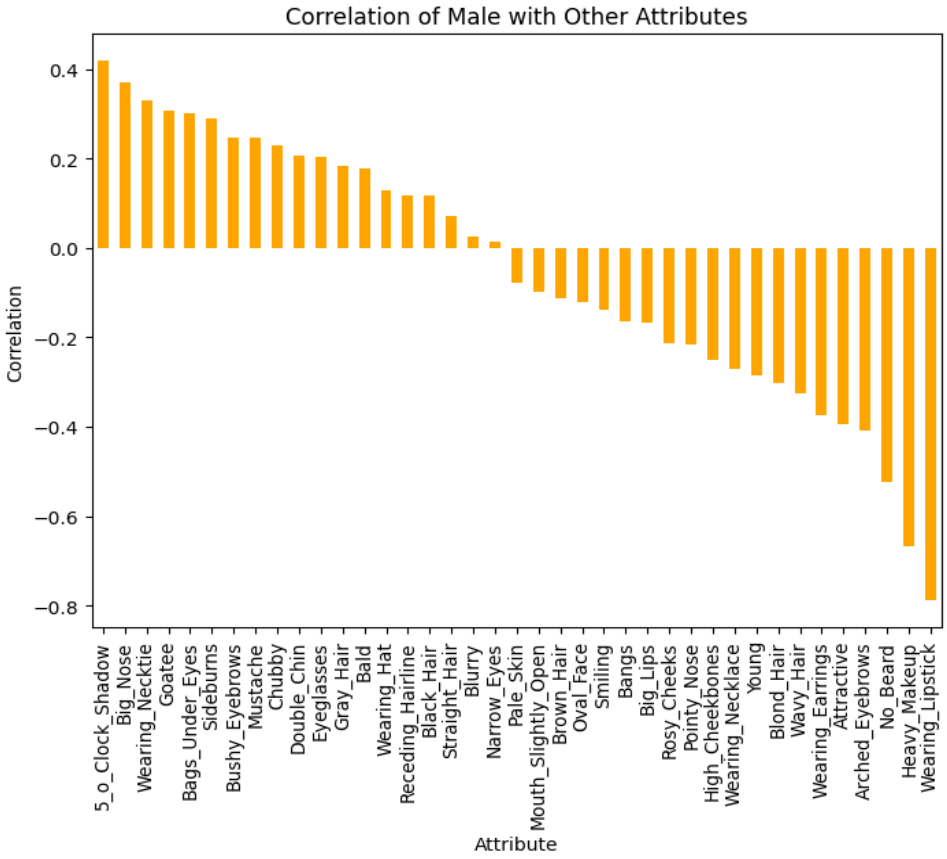
Visualize Attribute Correlation Matrix

The attribute correlation matrix provides insights into how attributes correlate with each other. This helps us understand the relationships between different attributes and identify any dependencies or patterns in the data. The heatmap visualization offers an intuitive way to identify strong correlations (both positive and negative) between attributes.



Visualize Gender and Other Attribute Correlation

To understand the relationship between gender and other attributes, we computed the correlation of each attribute with gender. This bar chart highlights the attributes that are most positively or negatively correlated with gender. By examining this correlation, we can identify attributes that might be influenced by gender biases and require further investigation.



These visualizations collectively provide a comprehensive view of the dataset's biases and characteristics, enabling us to make informed decisions on bias detection and mitigation strategies.

Bias Mitigation

To mitigate the bias detected in the dataset, we employed an oversampling technique to balance the gender distribution. The imbalance was evident from the original data distribution, where the male count was significantly lower than the female count.

Oversampling to Balance Gender Distribution

We first separated the male and female data from the dataset based on the 'Male' attribute. The original data distribution showed 84,434 male samples and 118,165 female samples, indicating a gender imbalance.

To address this imbalance, we applied the oversampling technique to the male data. We resampled the male data with replacement to match the number of female samples, resulting in a balanced dataset where both genders have an equal representation.

Balanced Data Distribution

After oversampling the male data, we combined the resampled male data with the original female data to create a balanced dataset. The balanced data distribution confirms that both male and female counts are now equal, with each gender having 118,165 samples.

By balancing the dataset in this manner, we aim to reduce the bias associated with gender representation, thereby creating a more equitable dataset for subsequent analysis and modeling.

```
Original Data Distribution:
Male Count: 84434
Female Count: 118165

Balanced Data Distribution:
Male Count: 118165
Female Count: 118165
```

Quantitative and Qualitative Analyzing New Mitigated Dataset

After mitigating the bias in the dataset through oversampling, we performed a comprehensive analysis to evaluate the effectiveness of the mitigation strategy. This involved calculating various statistical metrics to understand the distribution, variance, and correlation of attributes with respect to gender in the balanced dataset.

- *Imbalance Ratio*
  The imbalance ratio, calculated as the ratio of underrepresented to overrepresented samples, serves as a quantitative measure of dataset balance. In our balanced dataset, the imbalance ratio is 1.00, indicating that both male and female samples are now equally represented.

- *Chi-Squared Test*
  The Chi-Squared Test was conducted to statistically assess the independence between the gender attribute and the overall dataset. A Chi-Squared Test Statistic of 0.00 with a p-value of 1.0000 after resampling confirms that the gender distribution is now statistically equal, supporting the effectiveness of our oversampling strategy.

- *Mean Differences between Groups*
  We calculated the mean differences for each attribute between male and female groups to understand the extent to which each attribute varies between genders. The

mean differences after resampling reveal changes in the attribute distributions, which can be useful for understanding the residual biases in the dataset.

- *Variance Ratios between Groups*
  Variance ratios were computed to analyze the variability of each attribute within male and female groups. The variance ratios after resampling provide insights into how the dispersion of attributes has changed post-mitigation, helping us to understand if the oversampling has led to more balanced variability across genders.

- *Pearson Correlation between Gender and Attributes*
  We also calculated Pearson correlation coefficients to quantify the linear relationship between gender and each attribute in the balanced dataset. The correlation values after resampling indicate the strength and direction of the association between gender and individual attributes.

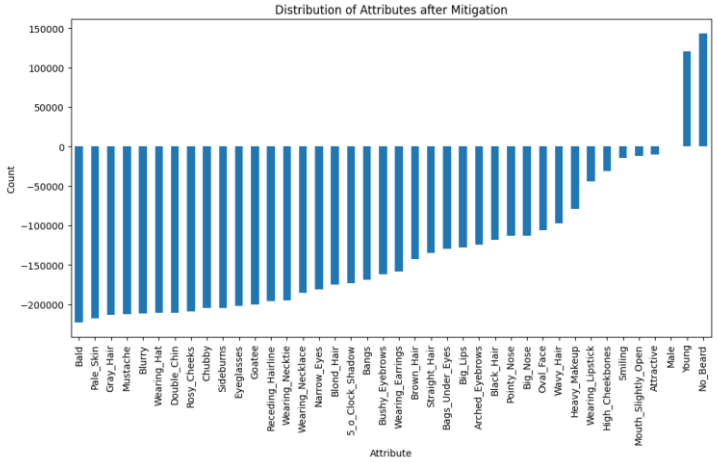
The comprehensive analysis of the mitigated dataset reveals promising results in terms of bias reduction and dataset balance. The imbalance ratio, Chi-Squared Test, mean differences, variance ratios, and Pearson correlation coefficients collectively indicate that the oversampling strategy has effectively mitigated the bias, resulting in a more equitable dataset for subsequent analysis and modeling.

Visualizing Detected Biases

Following the mitigation of bias through oversampling, we conducted visual analyses to gain insights into the distribution and correlation of attributes and gender within the balanced dataset.

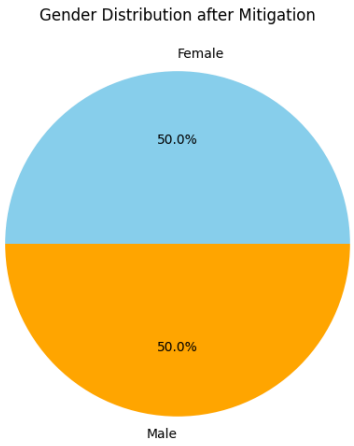
Attribute Distribution after Mitigation

The bar chart represents the distribution of individual attributes in the balanced dataset. It helps us visualize the count of each attribute, allowing us to observe the relative prevalence of different facial features and characteristics.



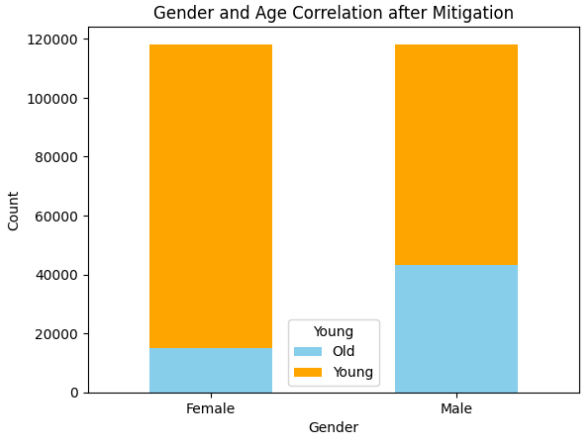
Gender Distribution after Mitigation

The pie chart displays the gender distribution within the balanced dataset, presenting the proportion of male and female samples. This visualization confirms the equal representation of genders, which aligns with the quantitative metrics calculated earlier.



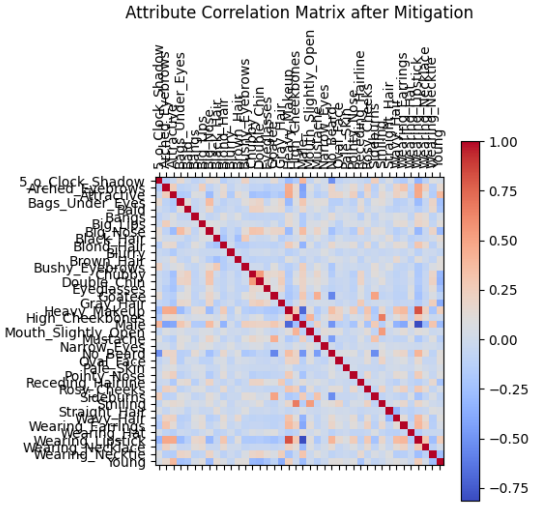
Gender and Age Correlation after Mitigation

The stacked bar chart illustrates the correlation between gender and age groups (Young or Old) in the balanced dataset. It enables us to understand how age is distributed across genders, providing insights into potential correlations or biases related to age and gender.



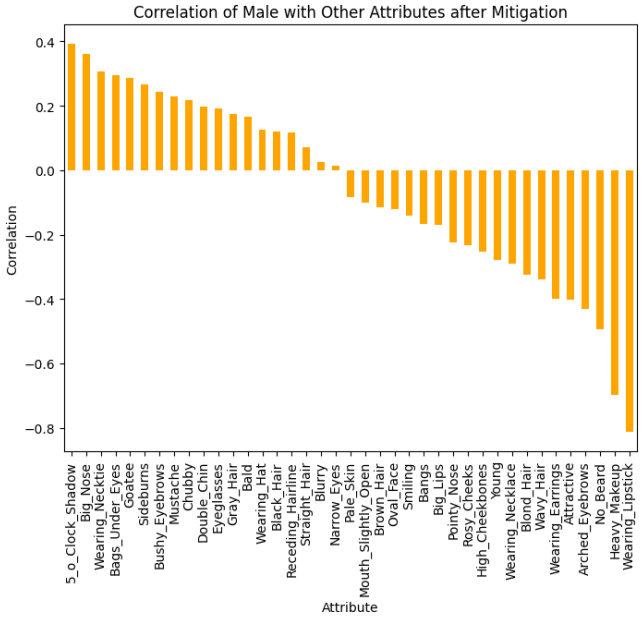
Attribute Correlation Matrix after Mitigation

The heatmap showcases the correlation matrix of attributes after mitigation. It offers a comprehensive view of how each attribute is correlated with others, aiding in identifying potential relationships or dependencies between facial features and characteristics.



Gender and Attribute Correlation after Mitigation

The bar chart depicts the correlation coefficients between gender and other attributes in the balanced dataset. It allows us to identify which attributes are most positively or negatively correlated with gender, highlighting the attributes that may still exhibit bias or disparity between male and female samples.



The visual analyses of attribute and gender distributions, as well as correlations, provide a holistic understanding of the dataset after bias mitigation. The visualizations corroborate the quantitative findings, confirming that the oversampling strategy has effectively balanced the dataset and reduced biases. This visually balanced and statistically equitable dataset serves as a solid foundation for subsequent analyses and modeling, ensuring fair and unbiased outcomes.

SMOTE Analysis and Model Evaluation

To further enhance the dataset's balance and evaluate the performance of the bias mitigation strategy, we employed the Synthetic Minority Over-sampling Technique (SMOTE) followed by a Random Forest Classifier evaluation.

SMOTE Data Resampling

We applied SMOTE to the training data to address any remaining class imbalance by oversampling the minority class (underrepresented gender) to match the majority class. This technique generates synthetic samples by interpolating between existing samples, thereby creating a more balanced training set.

The resampled training data consists of an equal number of male and female samples, ensuring that the classifier is trained on a dataset without class imbalance.

```
SMOTE Data Distribution:
Male Count: 94750
Female Count: 94750
```

Random Forest Classifier Evaluation

We utilized a Random Forest Classifier to evaluate the performance of the model on the resampled data. Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions, making it robust and accurate for classification tasks.

```
Accuracy: 0.9493716413489612

Classification Report:
              precision    recall  f1-score   support

          -1       0.97      0.93      0.95     23415
           1       0.93      0.97      0.95     23851

    accuracy                           0.95     47266
   macro avg       0.95      0.95      0.95     47266
weighted avg       0.95      0.95      0.95     47266
```

Conclusion

We successfully analyzed the CelebA dataset for gender bias and detected a significant imbalance through various statistical measures. To mitigate this bias, we employed an oversampling technique that balanced the gender distribution, resulting in equal representation of male and female samples. We validated the effectiveness of our bias mitigation strategy through comprehensive quantitative and qualitative analyses, including the use of SMOTE for

further enhancing dataset balance and a Random Forest Classifier for evaluating model performance. The results demonstrate that our approach has effectively reduced the gender bias in the CelebA dataset, leading to a more equitable and representative dataset for subsequent analyses and modeling tasks. Hence we were able to highlight the importance of addressing dataset biases and provide a systematic approach that can be applied to other datasets and domains to promote fairness and inclusivity in machine learning.