Data Appendix

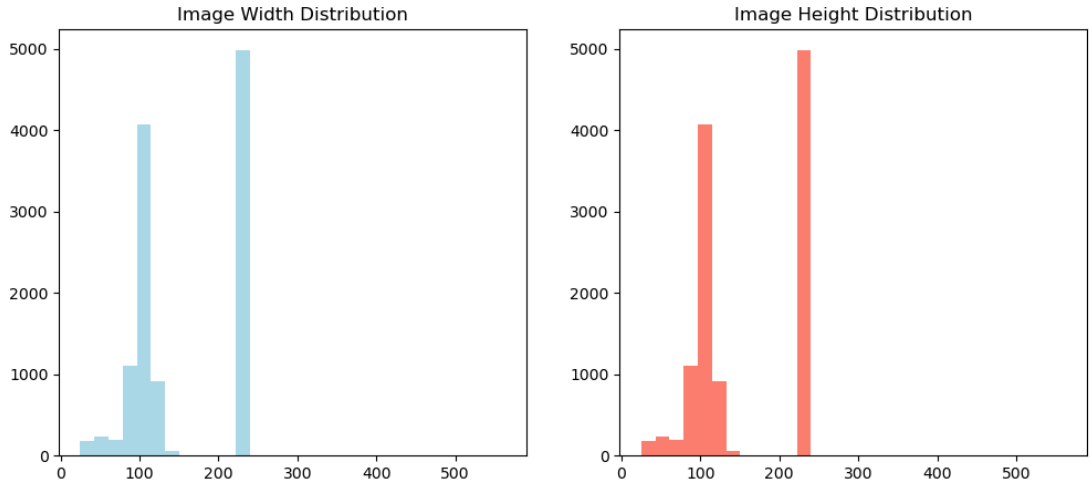
Analysis Data File: Face Mask Detection

Unit of Observation:
Each row in the dataset corresponds to an image of a person's face (either wearing a mask or not) along with the associated binary label for mask present or not (0 or 1 respectively). Furthermore, the units for our results (accuracy, f1-score, positive predictive value, miss-rate, auc-roc) are in the decimal range (0.00 to 1.00) corresponding to (0.00% to 100%).

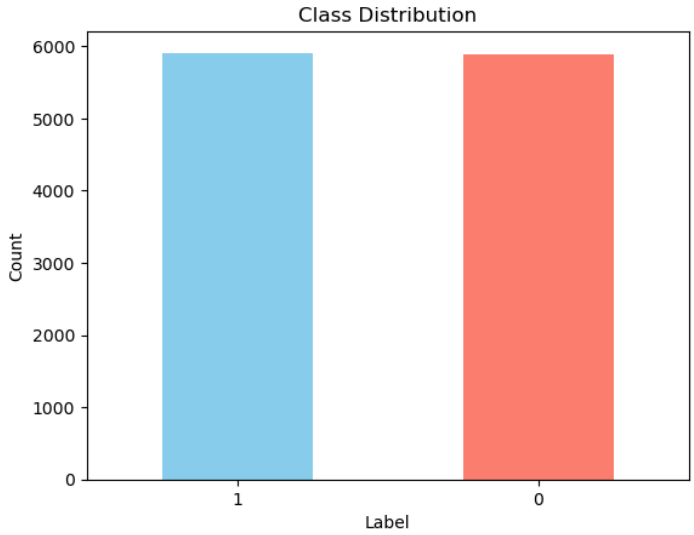
Overview of Analysis Data File:
The data analysis for this project (data exploration & results) is all located within the Jupyter file (script) itself, as no new dataset was generated.

Variables in the Analysis Data File:

1. **Image**
    - **Definition:** Different sizes of images in pixels of individuals either wearing a face mask or not.
    - **Type:** Qualitative
    - **Summary Statistics (units: Count):**
        - Count: 11792
        - Image Width Distribution: Depicted in the histogram below
        - Image Height Distribution: Depicted in the histogram below
    - **Visualization:** Histogram illustrating the distribution of width and height for all images within the dataset.



2. **Label**
    - **Definition:** Binary label indicating whether the image is of an individual wearing a mask or not (0 & 1 respectively)
    - **Type:** Quantitative
    - **Summary Statistics (units: Count):**
        - Count: 11792 (5896 ‘0’ Labels & 5896 ‘1’ Labels)
        - Class Label Distribution: Depicted in the histogram below.
    - **Visualization:** Histogram showing the distribution of binary class labels (mask present (0) and no mask present (1) within the dataset.



3. **Results**
    - **Definition:** Computed Confusion Matrix, F-1 Score, Positive Predictive Value, Miss Rate, AUC-ROC.
    - **Type:** Quantitative
    - **Summary Statistics(Units: 0.00-1.00 - decimal number where 1 = 100%):**
        - Confusion Matrix: Depicted in the visualization below.
        - F-1 Score: 1.00
        - Positive Predictive Value: 1.00
        - Miss Rate: 0.00
        - AUC-ROC: 0.00
    - **Visualization:** Computed Confusion Matrix

```
Confusion Matrix:
 [[591   0]
 [  0 589]]
```

Data Cleaning and Preprocessing:
The dataset was obtained from HuggingFace. From there, the dataset was split into training, validation, and testing sets (80%, 10%, and 10% respectively). All images were resized to 150x150 pixels and then normalized (between 0 and 1) to ensure consistency. Next, the training dataset's images were iterated using a batch size of 32, and a pretrained ResNet50 model was utilized to train the model. After freezing the base model layers and training for ten epochs, the base model layers were unfrozen, and the model was trained on another 5 epochs to gain our results.

Visualization and Summary:
Various plots, including histograms and bar graphs, were generated to understand the distribution among the images (height and width) and class labels across the dataset. These visualizations are available in the Data Appendix for further reference.