Akram Mahmoud

Nnamdi Okeke

11/8/24

Assignment 3.1 Data Summary Section Draft

Smart Mirror Melanoma/Skin Cancer Detection

1. What variables are present in your dataset and what are their datatypes?

   - duplicate_names is an integer type ,unique_patients is also an integer ,patients_with_changes_same_site is an integer, unique_patients_with_changes is also an integer

2. What issues were present in your dataset and what steps were taken to handle them (missing data, regularization, etc.)? What is your best guess for the source of these data issues?

   - We encountered duplicate values in our dataset, so we removed the duplicates by using the duplicate_df command to extract the duplicate values and subtract the extra values.

3. How are your variables related to your project goal? Do you see any patterns in the data that would suggest that they are/are not going to be useful in your machine learning model(s)? Do you need to transform or create new variables in order to reach your project goal?

- We believe the df_unique_ages_per_patient data along with patients_with_changes_same_site capture the essence of our project goal, because it connects the ages of the patients with whether or not the patients were found to have early signs of skin cancer.

4. How are your variables related to each other? Are there any strong correlations? How might this affect how you build your machine-learning model(s)?

   - The distribution of the number of unique years per patient is fairly consistent throughout the dataset. The dataset mostly consists of 2 unique years per patient, followed by 4 unique years per patient, and finally 3 unique years per patient where they were found to have signs of cancer.