Executive Report

How does the probability of dying from covid vary with the age of the patient?

Given that our focus is on analyzing patients who are positive to COVID-19, we've refined our dataset to exclusively include individuals who tested positive for the virus. Consequently, we've removed the column indicating COVID-19 positive status. Additionally, we have dropped certain columns that lack relevance to our analysis.

[L3.1] Load and analyze the dataset

Quasi-identifiers are attributes in a dataset that, when combined, could potentially identify individuals. In this dataset, attributes such as age, sex, native speaker status, and native Mexican status serve as quasi-identifiers. While these attributes alone may not directly reveal someone's identity, their combination poses a risk of re-identification, raising concerns for individual privacy. Therefore, careful handling of quasi-identifiers is crucial. After applying the k-anonymity algorithm, we found that the dataset did not meet the 2-anonymity requirement due to the quasi-identifiers. To address this, we generalized the 'age' attribute, grouping ages into broader categories to reduce individual identifiability. This anonymization process ensured that each individual's age was less distinguishable, leading to the dataset meeting the 2-anonymity requirement.

Additionally, in the next section, we assess whether the dataset satisfies l-diversity criteria. Upon evaluation, we found that the dataset meets the requirements for 2-diversity, 3-diversity, 4-diversity, and 5-diversity.

The analysis reveals an imbalance in the dataset concerning deceased patients, with a notable prevalence of deceased individuals compared to recovered ones across various demographics. This imbalance may introduce bias into predictive models, favoring the majority class. Specifically, there's a higher average rate of deceased patients among males compared to females. Moreover, the disparity persists across different age groups, especially among older individuals. Regarding native speakers, the count of deceased patients is notably higher than that of recovered ones, particularly pronounced among non-native speakers. Similarly, for native Mexicans, deceased patients outnumber recovered ones, highlighting a potentially high mortality rate within the population. Further analysis by age groups suggests varying impacts of the disease between native and non-native speakers, as well as native and non-native Mexicans, with age playing a significant role in recovery rates.

In the context of discrimination analysis, we consider both protected class and sensitive attributes as factors that impact fairness and equality. We identify 'sex' as the protected class due to the significant difference in mortality rates between genders. Females are designated for protection as they have a lower mortality rate compared to males. Sensitive attributes are personal characteristics that can lead to discrimination if used improperly. In this analysis, attributes such as 'sex', 'native_speaker', and 'native_mexican' are identified as sensitive attributes. These attributes influence the outcome of the analysis in a way that disadvantages one group over another, leading to inequality.

[L3.2] Prepare your dataset and train a model

We have chosen Logistic Regression as the classifier. It is a common choice for binary classification tasks, which seems appropriate for predicting whether a patient is deceased or not based on their age group. For the target column, we have chosen the 'deceased_patient' variable. We trained different models to determine which is better for our task, including the analysis of their performance.

1. Age: The model achieved an accuracy of 0.8214. To enhance prediction accuracy, consideration of additional factors such as the top three diseases affecting older individuals is recommended.
2. Age, pneumonia, hypertension, and diabetes: After plotting a heatmap, we identified pneumonia, hypertension, and diabetes as the key diseases related to age. Integrating these variables, improved accuracy to 0.8697.
3. Age, pneumonia, hypertension, diabetes, and sex: Given the initial observation of an imbalance in the 'sex' variable, we included it to assess its impact. The model's accuracy increased to 0.8737 with the addition of sex as a variable.

[L3.3] Assess your model performance

Based on the results provided in the section above, logistic Regression proved to be a suitable classifier for predicting patient mortality based on age group, with the inclusion of variables such as pneumonia, hypertension, diabetes, and sex contributing to improved accuracy. Using this result, we analyzed the performance of the model.

1. Model Performance Analysis

The Logistic Regression model demonstrated promising performance, achieving an accuracy of 0.8737, with a high precision of 0.8933 and recall of 0.9495. These metrics indicate the model's effectiveness in accurately predicting patient mortality for COVID-19 based on age group and other key variables.

The false positive rate was relatively high at 0.3806, while the false negative rate was low at 0.0505, suggesting the model's potential utility for prevention purposes. Specificity, indicating the model's ability to correctly predict negative instances, was moderate at 0.6194.

2. Feature Importance Analysis

The analysis revealed 'age_group_encoded' as the most crucial feature for prediction, followed by 'pneumonia', 'hypertension', 'diabetes', and 'sex'. These findings highlight the significant role of age in determining patient mortality due to COVID-19, with pneumonia emerging as another influential predictor. The SHAP summary plot further emphasized the importance of pneumonia and age group in the model's predictive capabilities.

[L3.4] Assess the fairness metrics

In this section we computed the performance of the model including some new metrics to evaluate the fairness of the model in terms of their predictions across different demographic groups. The disparate impact ratio indicates that men have approximately 0.9019 times the likelihood of favorable outcomes compared to women. The equal opportunity difference indicates that the true positive rate for men is approximately 0.0184 higher than that for women, indicating a small disparity in favor of men in correctly predicting positive outcomes. The average odds difference indicates that, women experience approximately 0.0824 more false positives or fewer true positives compared to men on average. These results suggest that the model may exhibit some bias in favor of men compared to women

[L3.5] Apply mitigation algorithms to your model

In the context of logistic regression, two common algorithms for mitigating unfairness are: Disparate Impact Remover and Reweighing. Given the observed disparities in false positive and false negative rates in Model 3, Reweighing emerges as a more feasible and effective choice to enhance fairness while maintaining model performance. Therefore, Reweighing is preferred due to its practicality and potential to mitigate disparities effectively.

[L3.6] Assess the effects of the mitigation technique

The reweighing algorithm applied to the classifier had mixed effects on fairness and performance metrics. It resulted in notable improvements in fairness metrics, indicating a reduction in bias. However, some performance metrics, such as accuracy, recall, F1 score, and false negative rate, experienced a small decrease, while others, such as precision and false positive rate, showed improvements. Overall, while the reweighing algorithm positively impacted fairness metrics, it had minor trade-offs in terms of predictive accuracy and other performance measures.

Conclusions

Based on the analysis conducted on this project we conclude that the probability of dying from COVID-19 generally increases with the age of the patient. Our findings consistently demonstrate that older individuals face a significantly higher risk of mortality compared to younger age groups when infected with the virus.

Contributions

The two members of the group have contributed equally to the project, working together in practice sessions, and dividing the work on other occasions, always maintaining good communication to facilitate group work.