**Executive report.**

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

1. Anonymization of the dataset. redactar.

**Identifiers:** These are unique attributes that can be used to identify an individual. In this dataset, there doesn’t seem to be any explicit identifiers like name, social security number, or patient ID.

**Quasi-identifiers:** These are attributes that, in combination, can be used to identify an individual. In this dataset, the following could be considered quasi-identifiers: ‘age’, 'sex', 'native\_speaker’, ‘native\_mexican’.

**Sensitive attributes:** These are sensitive attributes that should be protected. In this dataset, the following could be considered confidential attributes: 'sex', 'native\_speaker’, ‘native\_mexican’.

**Other:** These attributes could also be considered sensitive attributes, however, as far as our study is concerned, they are attributes that are related to a person's state of health, as well as possible habits (such as smoking). These attributes are related to a person's age, so we consider them as "other": 'pneumonia', 'diabetes', 'epoc', 'asma', 'immunosuppression', 'hypertension', 'other\_diseases', 'cardiovascular', 'obesity', 'chronic\_renal\_failure', 'deceased\_patient'.

As we can observe, with the given quasi-identifiers, the dataset is not 2-anonymous. If a dataset does not satisfy 2-anonymity, it means that there are individuals in the dataset who can be uniquely identified based on the quasi-identifiers. Given this situation, we will have to anonymize the dataset. In this case, we could apply generalization to the ‘age’ attribute.

k-anonymity Given the previously established quasi-identifiers, does the dataset satisfy 2-anonymity)? True

L-diversity 2-diversity of the dataset: True, 3-diversity of the dataset: True, 4 l-diversity of the dataset: True, 5-diversity of the dataset: True.

2. Fairness in original dataset. redactar.

As we can observe, regarding "deceased\_patients", we could say that the dataset is imbalanced. The factor that there are significantly more deceased patients than recovered ones, could lead the model to become biased towards the majority class.

deceased\_patients by age - Again, we can observe that regarding "deceased\_patients", we could say that the dataset is imbalanced. This means there are significantly more deceased patients than recovered ones in most age groups and, in this case, especially in the older age groups.

deceased\_patients by sex - The dataset, again, is imbalanced across different genders. This means there are significantly more deceased males than recovered ones, and significantly more recovered females than deceased ones. As usual, there are more males than females in the dataset.

This output suggests that the average rate of deceased patients is slightly higher for the ‘sex’ category Male compared to Female.

deceased\_patients by sex and age.

* For both males and females, there are more deceased patients than recovered in each age group, except for males aged 10-40 and females aged 10-50 where the number of recovered is higher.
* The gap between the number of recovered and deceased patients seems to narrow as age increases, especially for females.

deceased\_patients by native speker - As we can observe, the total count for non-native speakers is significantly higher than that for native speakers. Furtheremore, it’s important to note that the count of deceased patients is significantly higher than the count of recovered patients in both groups. However, the difference is much more pronounced for non-native speakers. This output suggests that the average value of the “deceased\_patient” column is slightly higher for the group where “native\_speaker” is 1.0 compared to the group where “native\_speaker” is 0.0.

deceased\_patients by native speaker and age.

The data suggests that the impact of the disease varies significantly between native and non-native speakers and across different age groups.

* Native Speaker Patients: The graph shows that the highest number of deceased patients is in the age group of 80–90. The number of recovered patients is generally lower than the number of deceased patients across all age groups (as exception in the range of 10-40).
* Non-Native Speaker Patients: The counts of both deceased and recovered patients are significantly higher compared to native speakers. The age group of 60–70 has the highest number of deceased patients, while the age group of 30–40 has the highest number of recovered patients.

deceased\_patients by native mexican.- As we can observe, the total count for non-native mexican is significantly higher than that for native mexican. Additionally, from the numbers, it’s clear that the count of deceased patients is significantly higher than the count of recovered patients in both categories. This could indicate a high mortality rate among the patient population. This ouput suggests that the average value of the “deceased\_patient” column is slightly higher for the group where “native\_mexican” is 0.0 compared to the group where “native\_mexican” is 1.0.

deceased\_patients by native mexican and age.- These graphs provide a comparative view of the recovery and death rates among different age groups for Native and Non-Native Mexican patients.

* Native Mexican Patients by Age: For ages 10-20, 20-30 and 30-40, there are significantly more recovered patients than deceased. However, for ages 50 and above, the number of deceased patients is higher than those who have recovered.
* Non-Native Mexican Patients by Age: All age groups except for 10-20, 20-30 and 30-40 have more deceased patients than recovered. The highest count of both recovered and deceased patients is in the age group of 50-60 and 60-70 respectively.

These observations could suggest that age is a significant factor in the recovery rate of patients.

3. Factors that impact on discrimination. redactar.

Protected class.

Considering the average mortality rates we’ve calculated:

* For ‘sex’, the mortality rates are 0.728652 for group 1 (female) and 0.796536 for group 2 (male).
* For ‘native\_speaker’, the mortality rates are 0.771025 for group 0 (non-native) and 0.791355 for group 1 (native).
* For ‘native\_mexican’, the mortality rates are 0.771438 for group 0 (non-native) and 0.742678 for group 1 (native).

So, if we want to protect the group with the highest disparity in mortality rates, ‘sex’ would be the protected class, as it has the largest difference between the two groups. Specifically, we will protect female as is the group with lower rate.

Sensitive attributes.

As we can observe, the attributes studied above influence the outcome of our analysis in a way that disadvantages one group over another, leading to inequality. For this reason, the sensitive attributes are ‘sex’, ‘native\_speaker’, and ‘native\_mexican’, as they they relate to personal characteristics of individuals that are protected under anti-discrimination laws. We cannot forget that medical data, such as illnesses (diabetes, pneumonia, etc.), are also considered sensitive attributes.

4. Fairness metrics.

5. Algorithms used for training the model.

6. Algorithmic performance expected and obtained.

7. Algorithmic metrics applied.

8. Algorithmic performance after the metrics.

9. Conclusions regarding the goal of your model and the expected utility in a real-case scenario.