MedTech
Mediterranean
Institute of Technology

Classification of Covid patients based on their health situations

CS434 Data Analytics

The Team

Amani Touihri

Software Engineering

Emna Bouzid

Computer Systems Engineering

Nour Bennour

Software Engineering

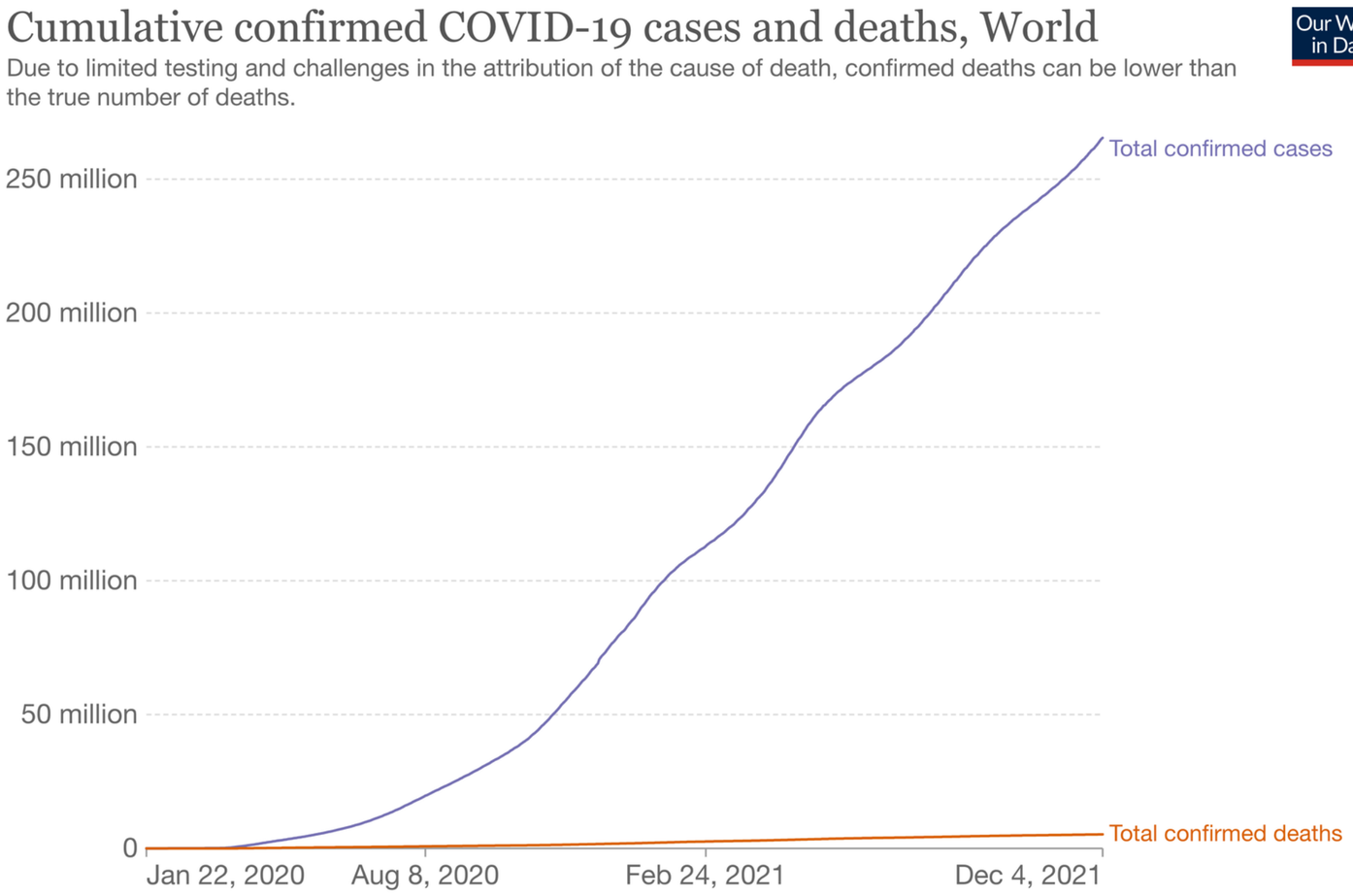
Khalil Chebbi

Software Engineering

Table of content

- Problem
- Data Collection
- Data Cleaning
- Data Visualisation
- Modeling
- Testing
- Challenges
- Future steps

Problem



265M

(4 Dec 2021)

Problem

How to classify COVID patients based on their current health conditions?

Data

A tabular dataset containing information for patients in two American hospitals.

http://ictcf.biocuckoo.cn/Resource.php

Data Collection

Web Scrapping

Data Cleaning

Remove unnecessary data

Example:
°C in Body temperature

Data Cleaning

Remove unnecessary data

Remove outliers

Example: Body temperature < 30

Data Cleaning

Remove unnecessary data

Remove outliers

Fill blank data

Example: replace by the mean of body temperature for negative cases and for positive cases

Data Cleaning

Remove unnecessary data

Remove outliers

Fill blank data

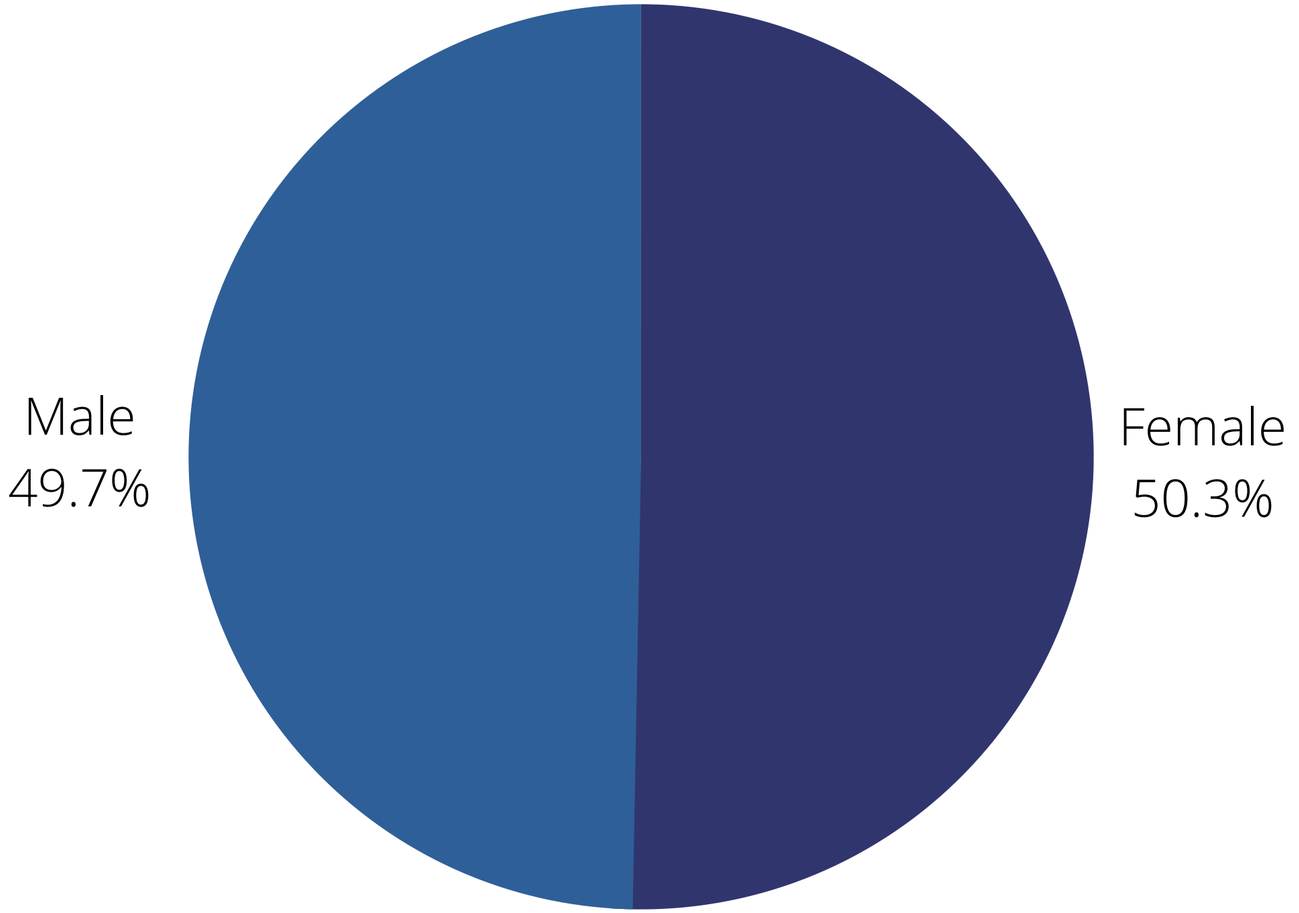
Redefine data

Example: for Regular cases: add 121 unknown to cured since cured is 475 and deceased 0

Data Visualisation

Gender distribution

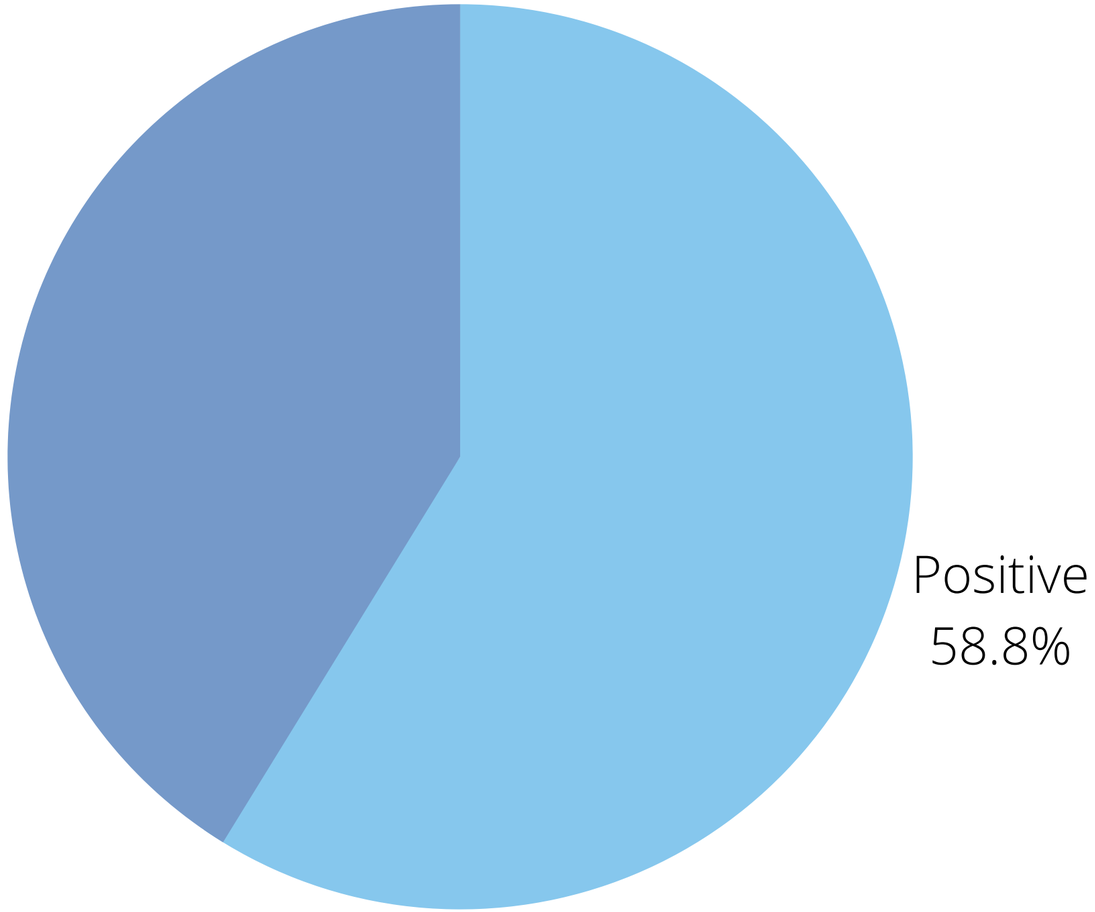
Female: 765
Male: 756



SARS-CoV-2 nucleic acids tests distribution

Positive: 894

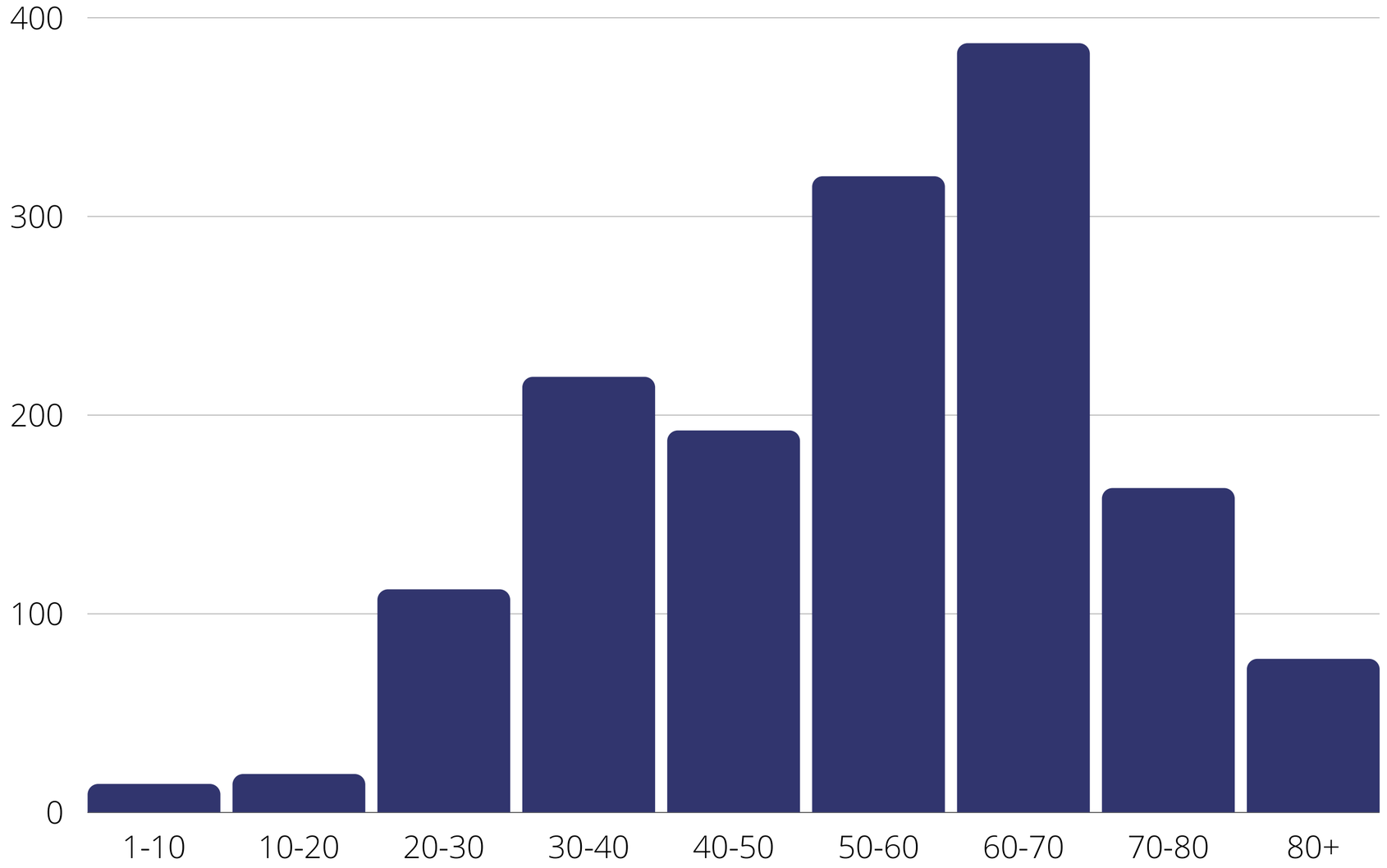
Negative: 627



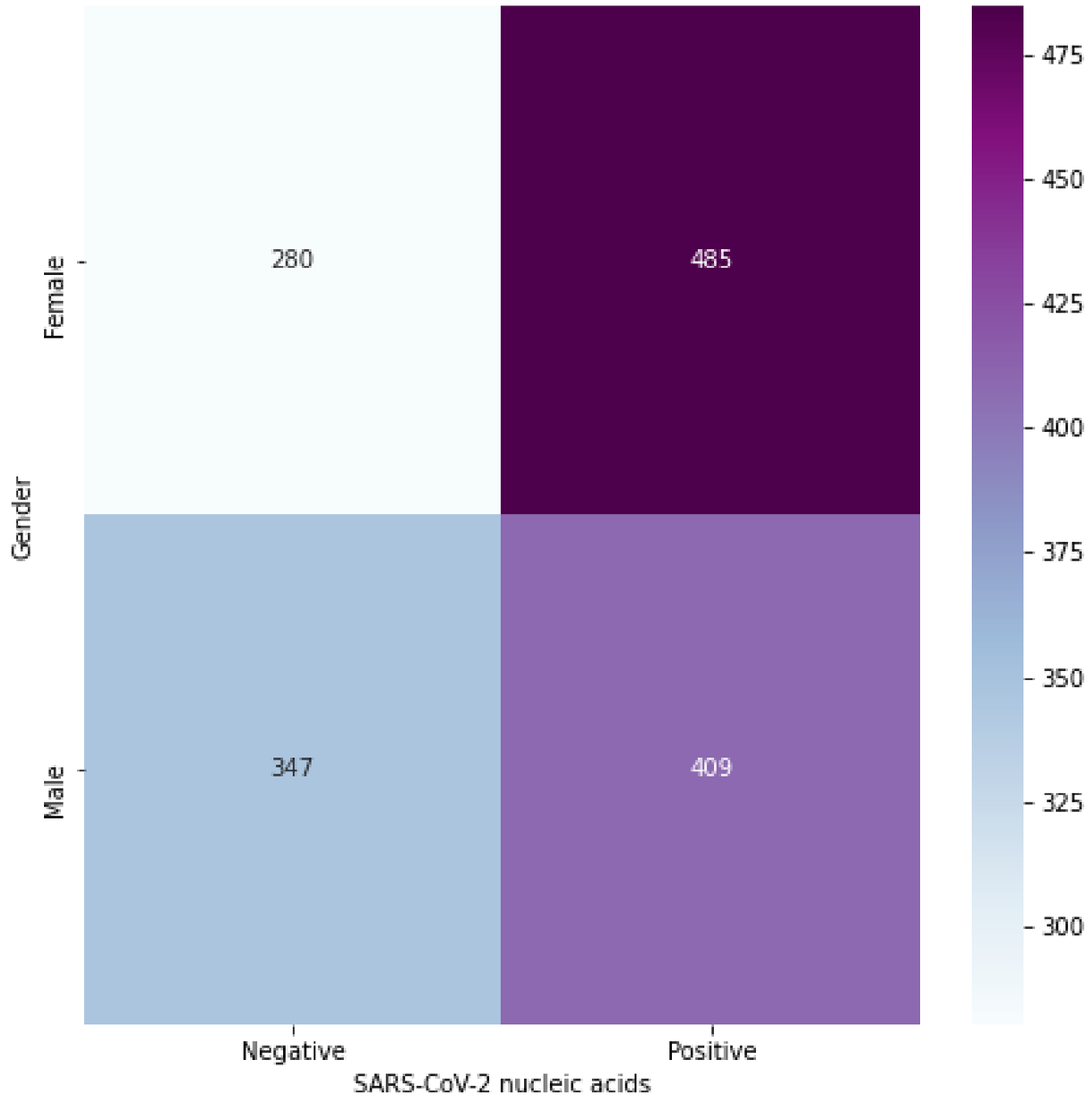
Age distribution

The average age of the dataset is 55

The most common age is 65

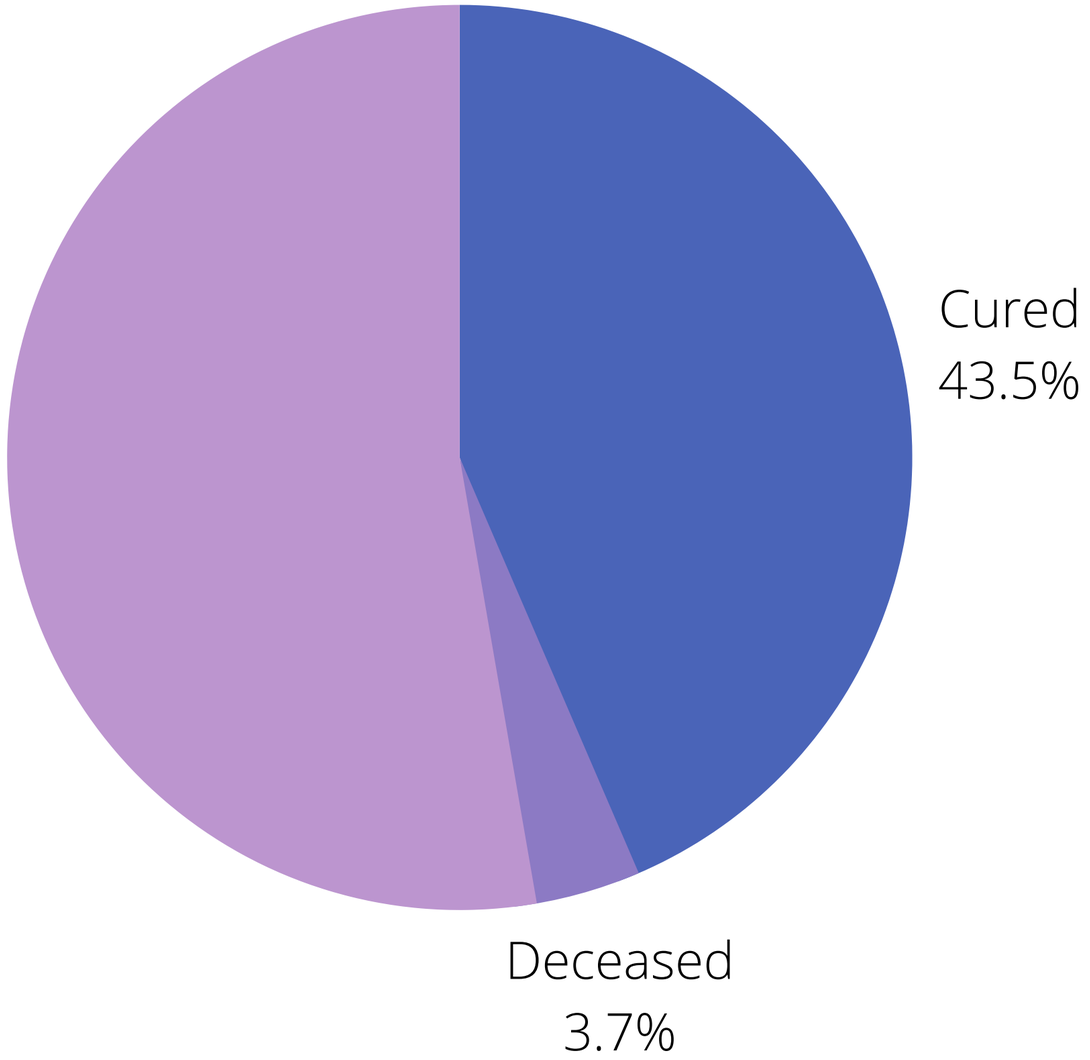


SARS-CoV-2 nucleic acids tests per gender



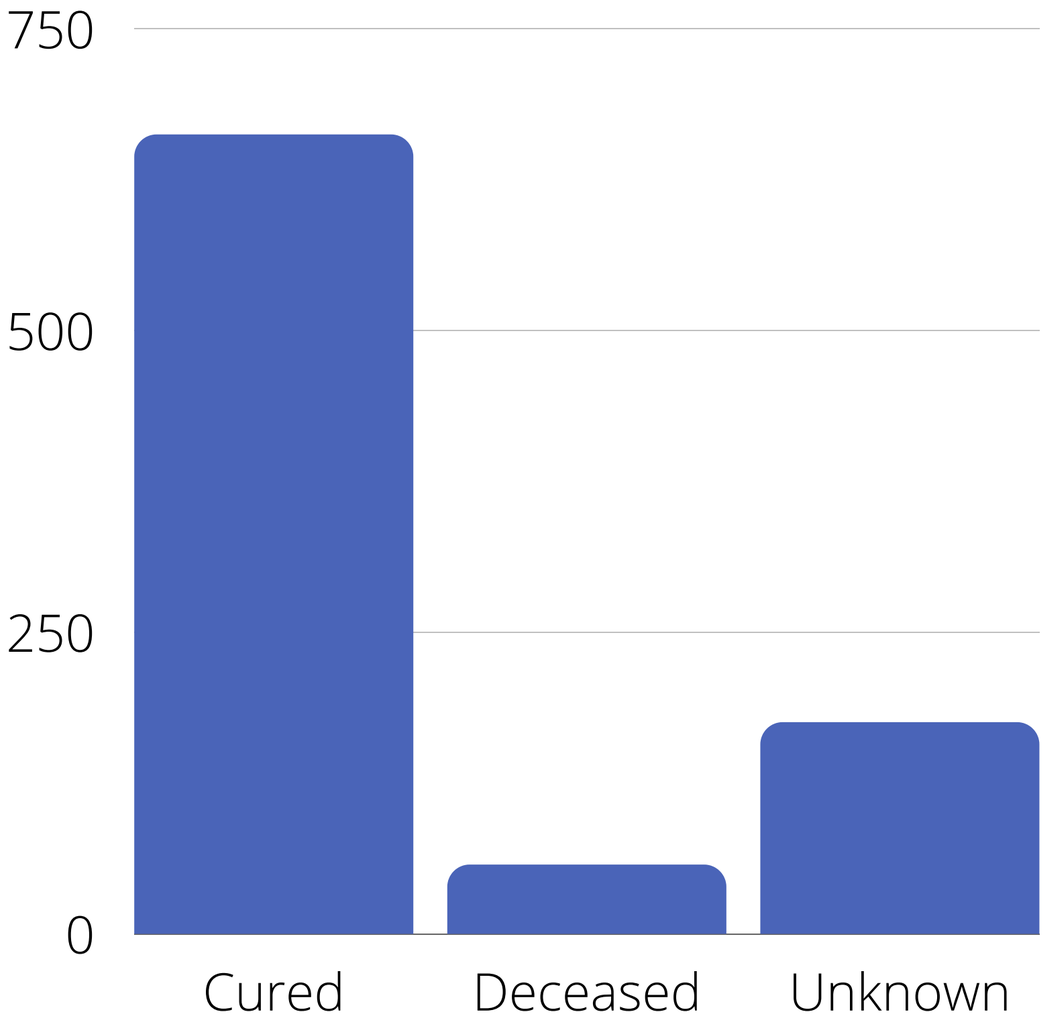
Mortality distribution

Cured: 662
Deceased: 57
Unknown: 802

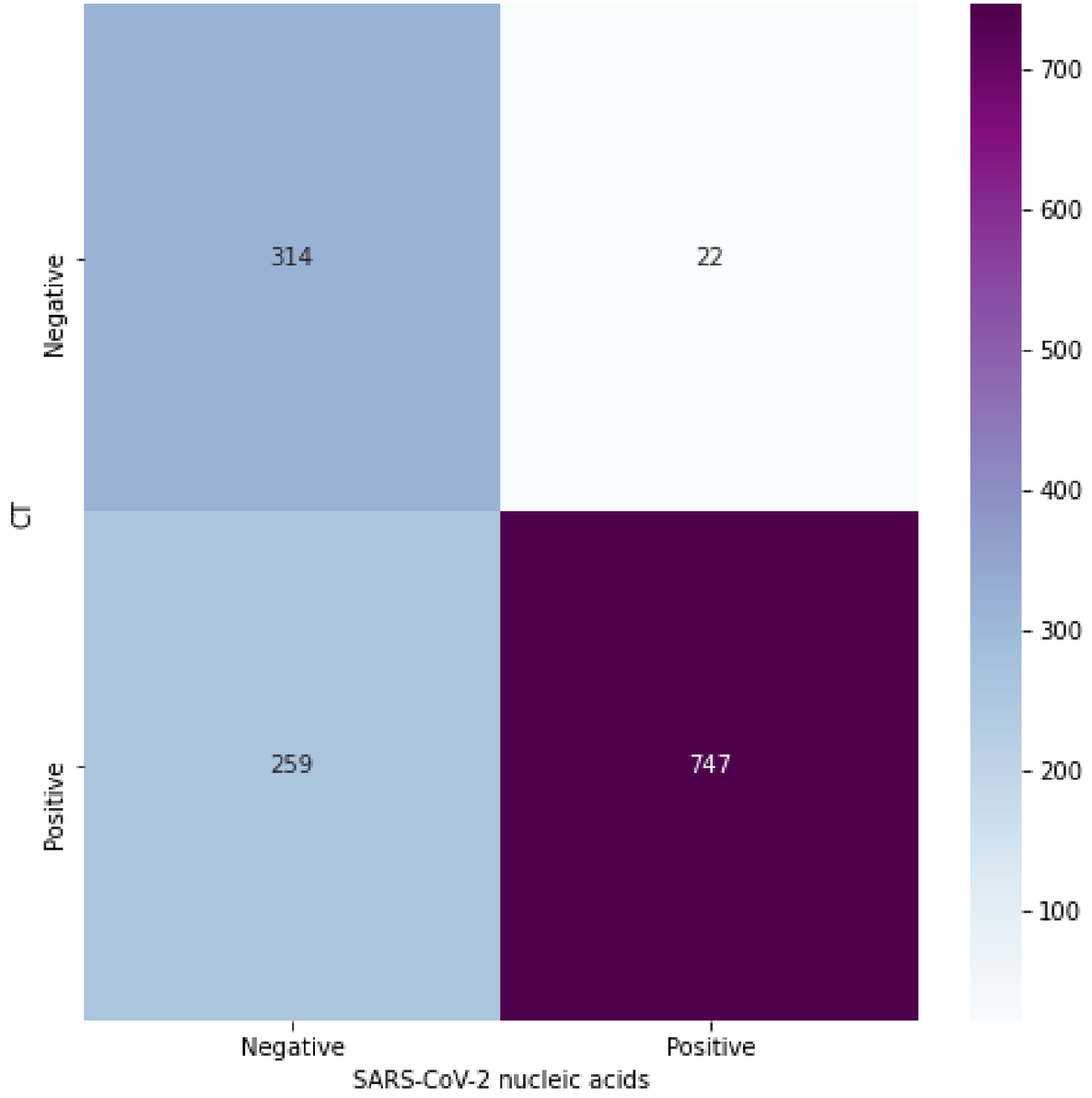


Mortality for positive tests

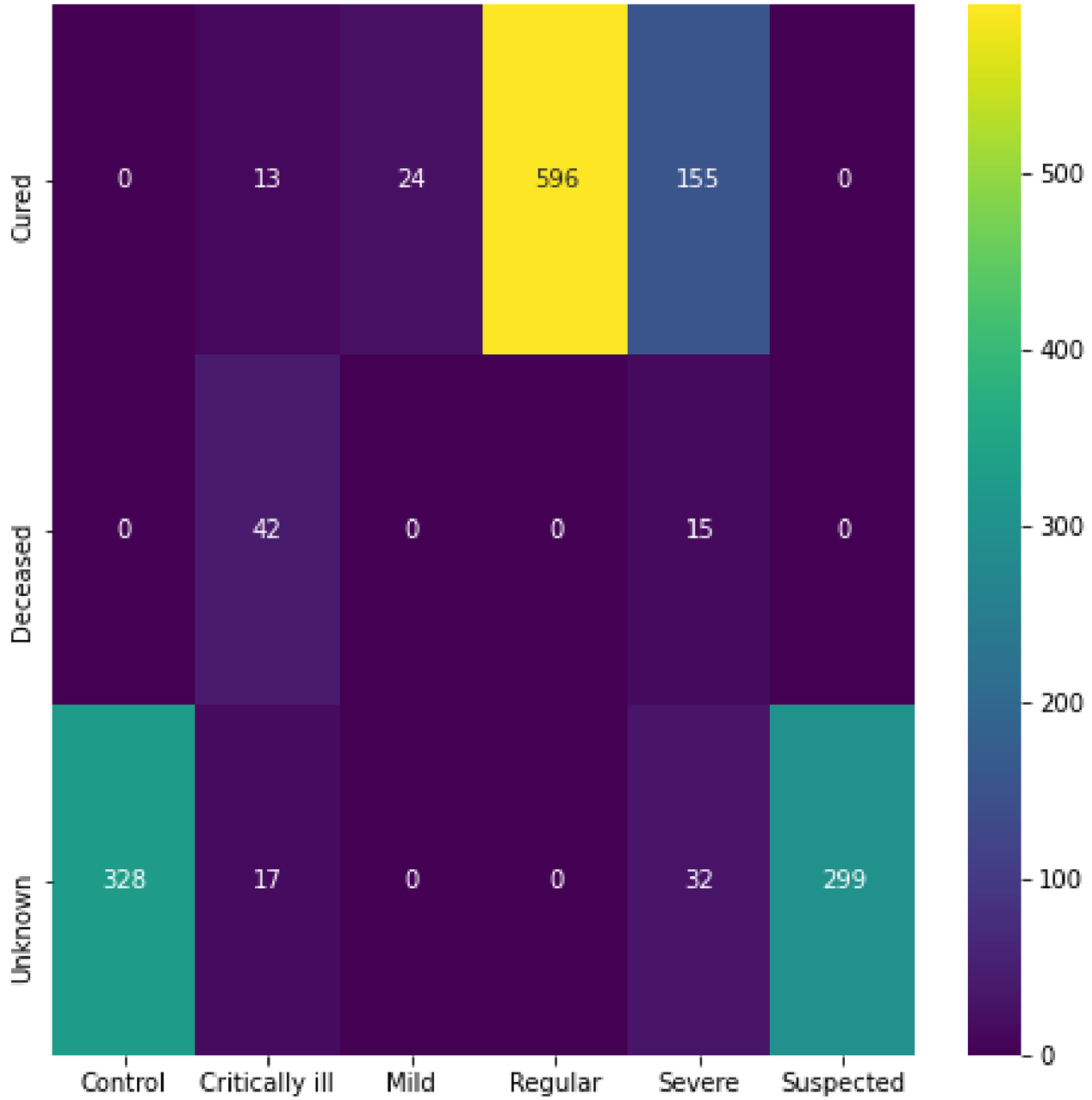
Cured: 662
Deceased: 57
Unknown: 175



SARS-CoV-2 nucleic acids tests VS CT Scans



Mortality VS Morbidity

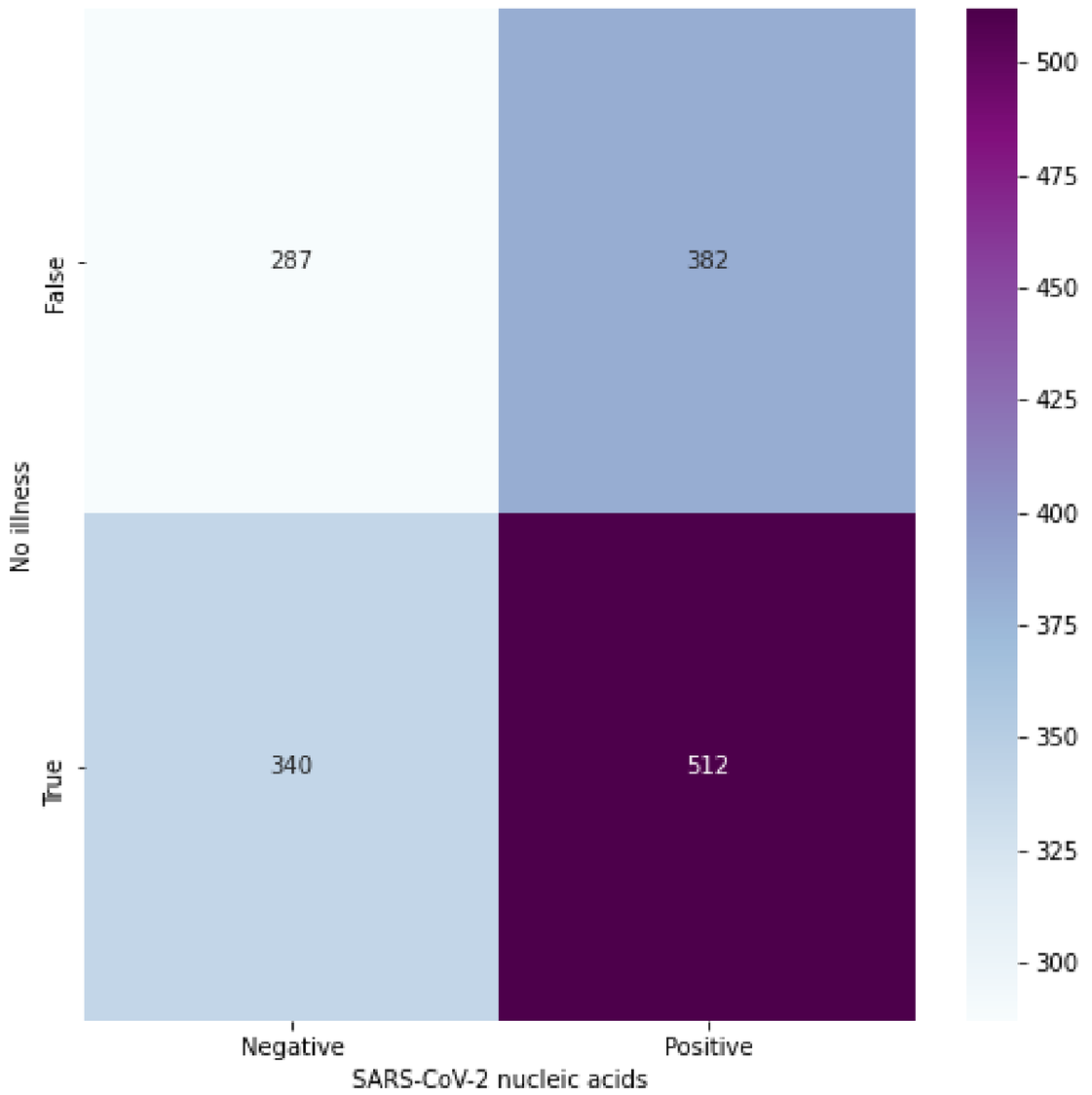


Underlying Diseases

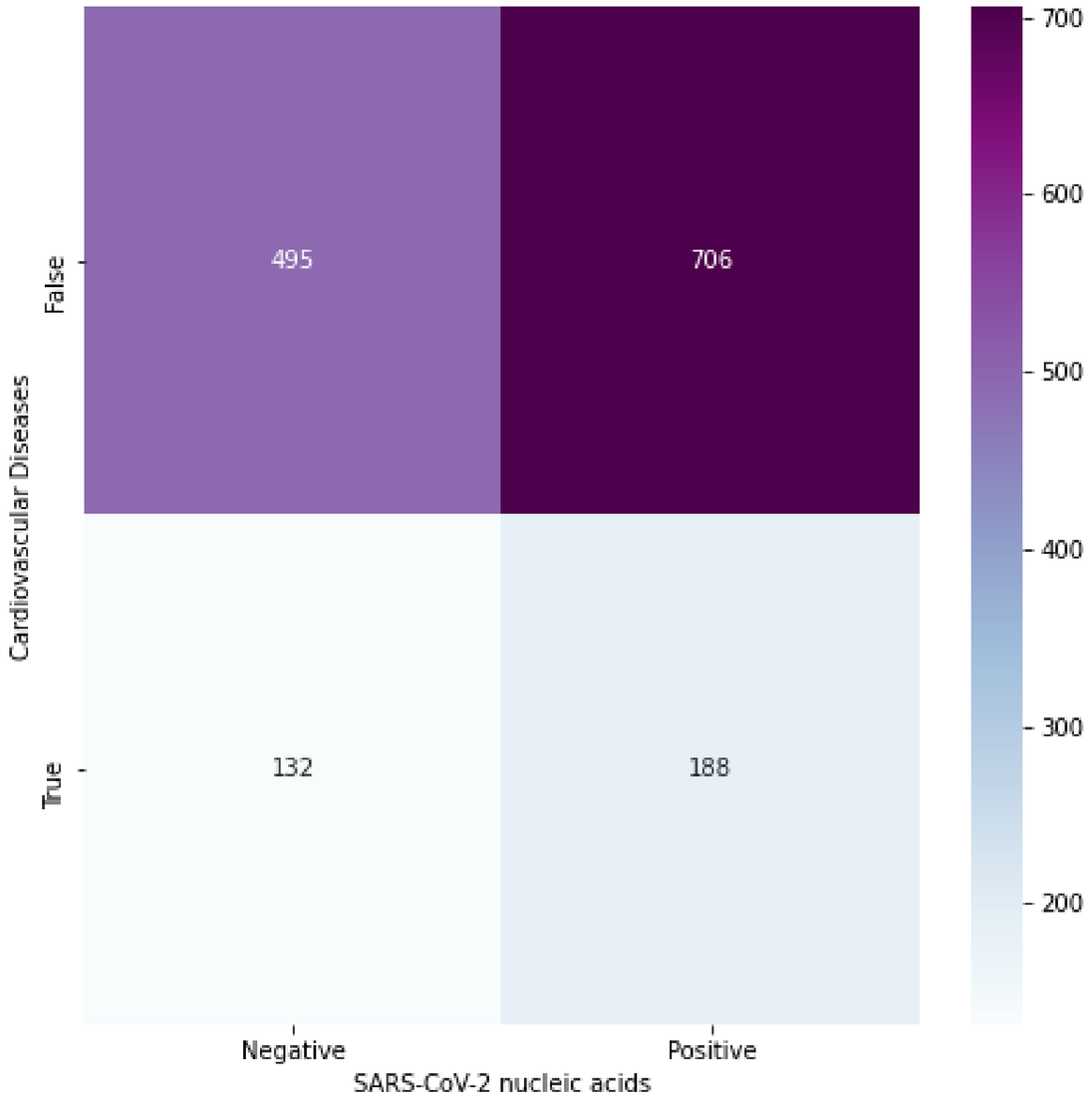
Divide:

- No illness
- Cardiovascular
- Thoracic
- Neurological
- Oncological
- Nephrological
- Endocrinological

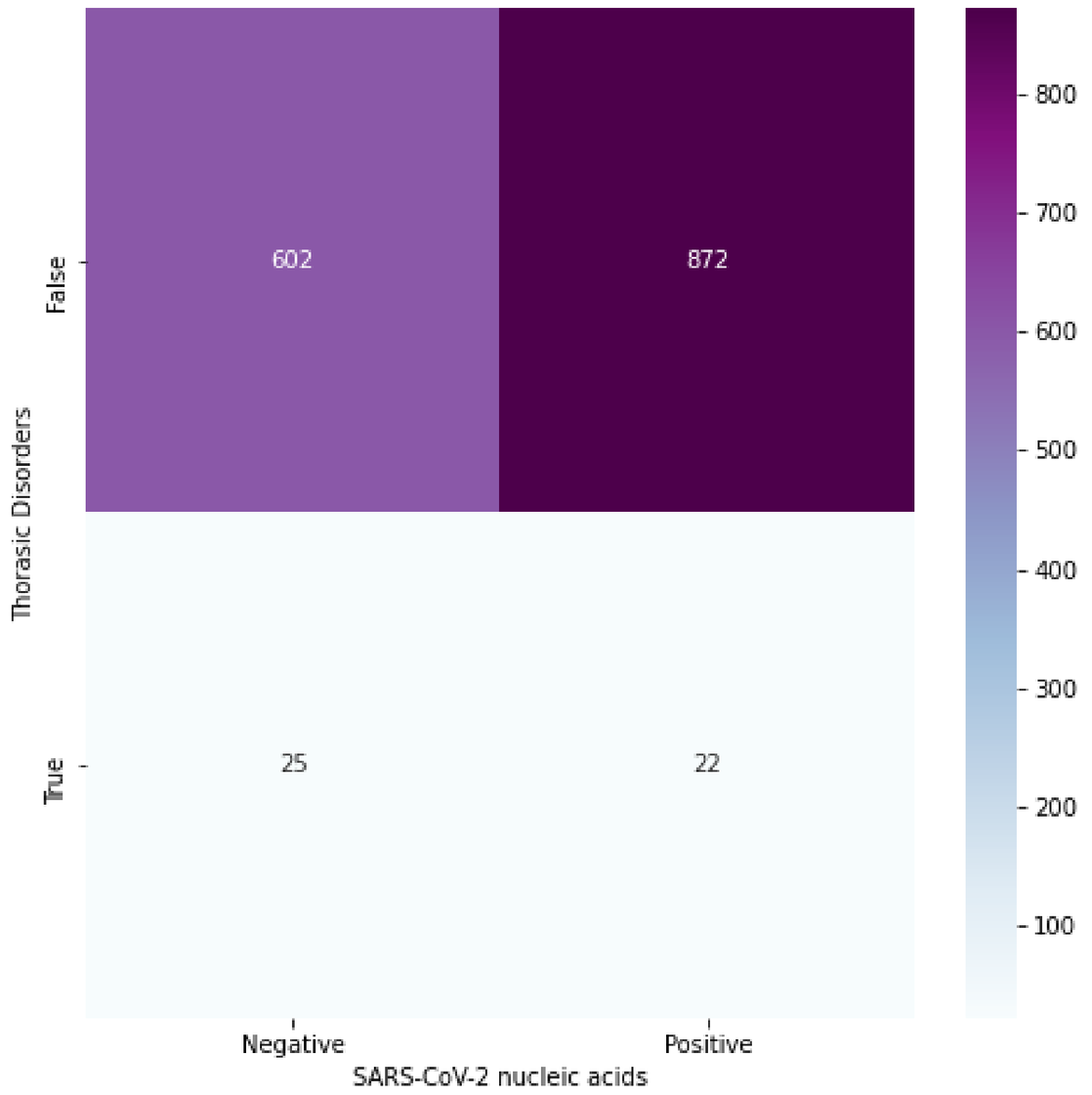
SARS-CoV-2 nucleic acids tests VS no underlying diseases



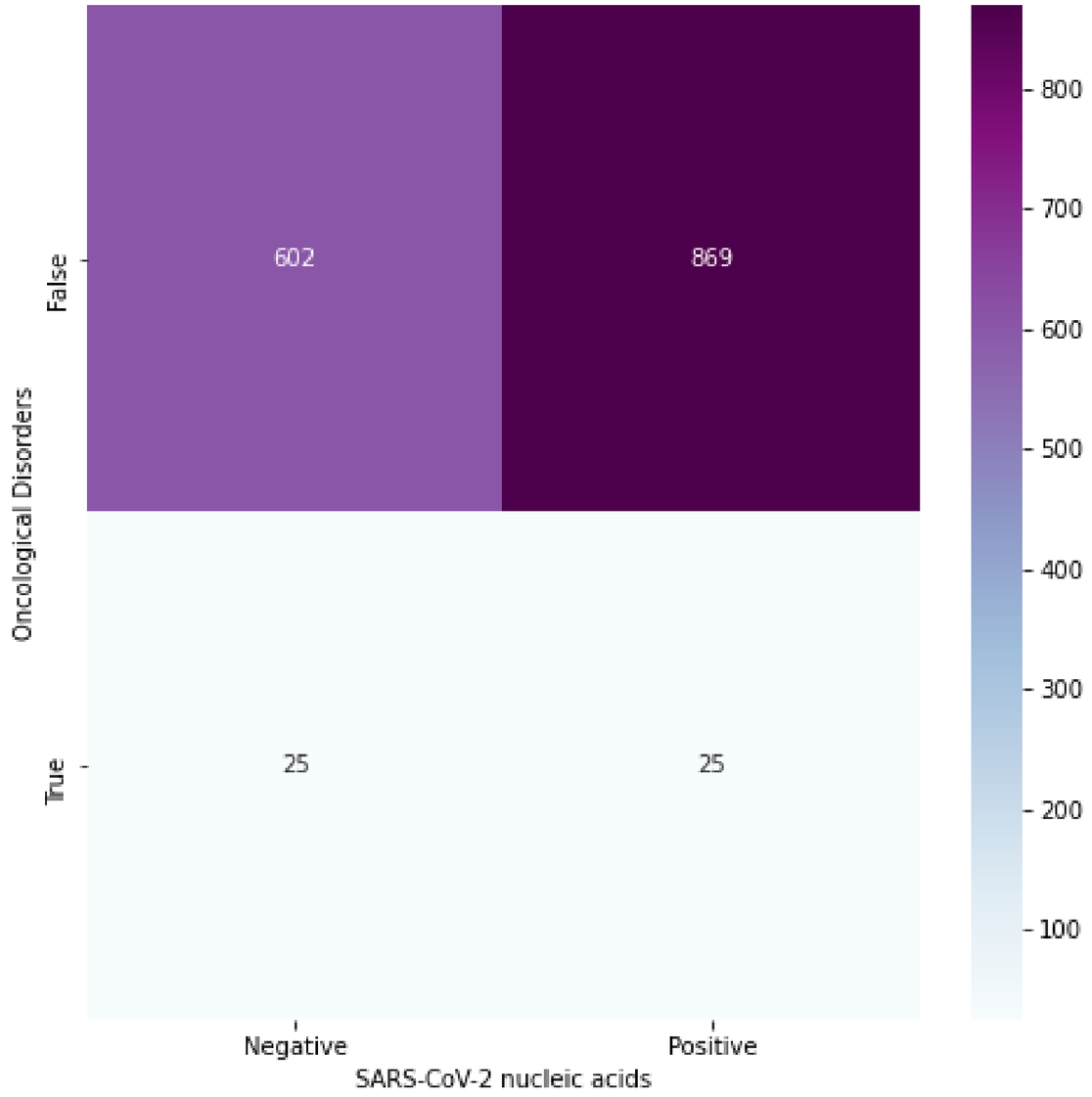
SARS-CoV-2 nucleic acids tests VS cardiovascular diseases



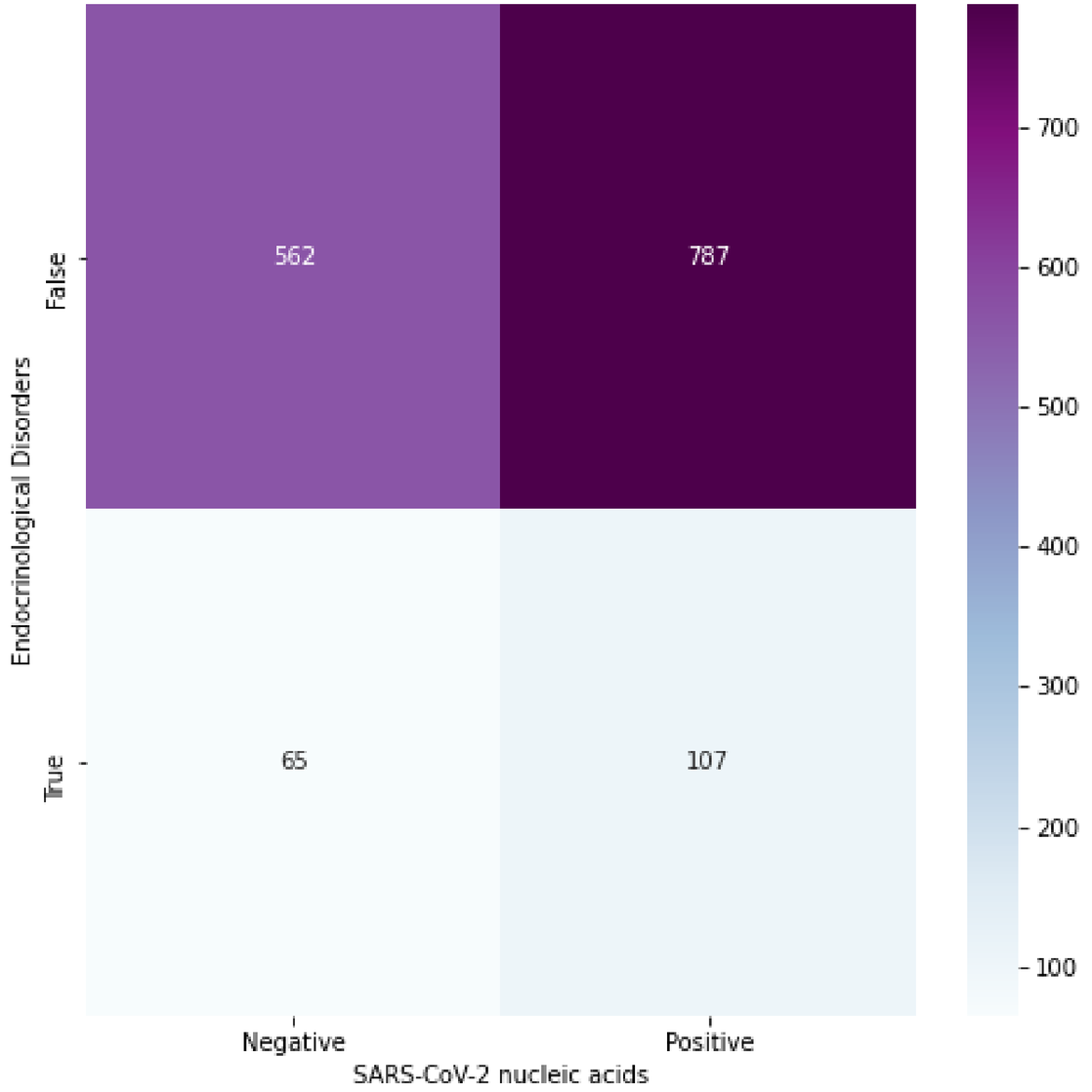
SARS-CoV-2 nucleic acids tests VS thoracic diseases



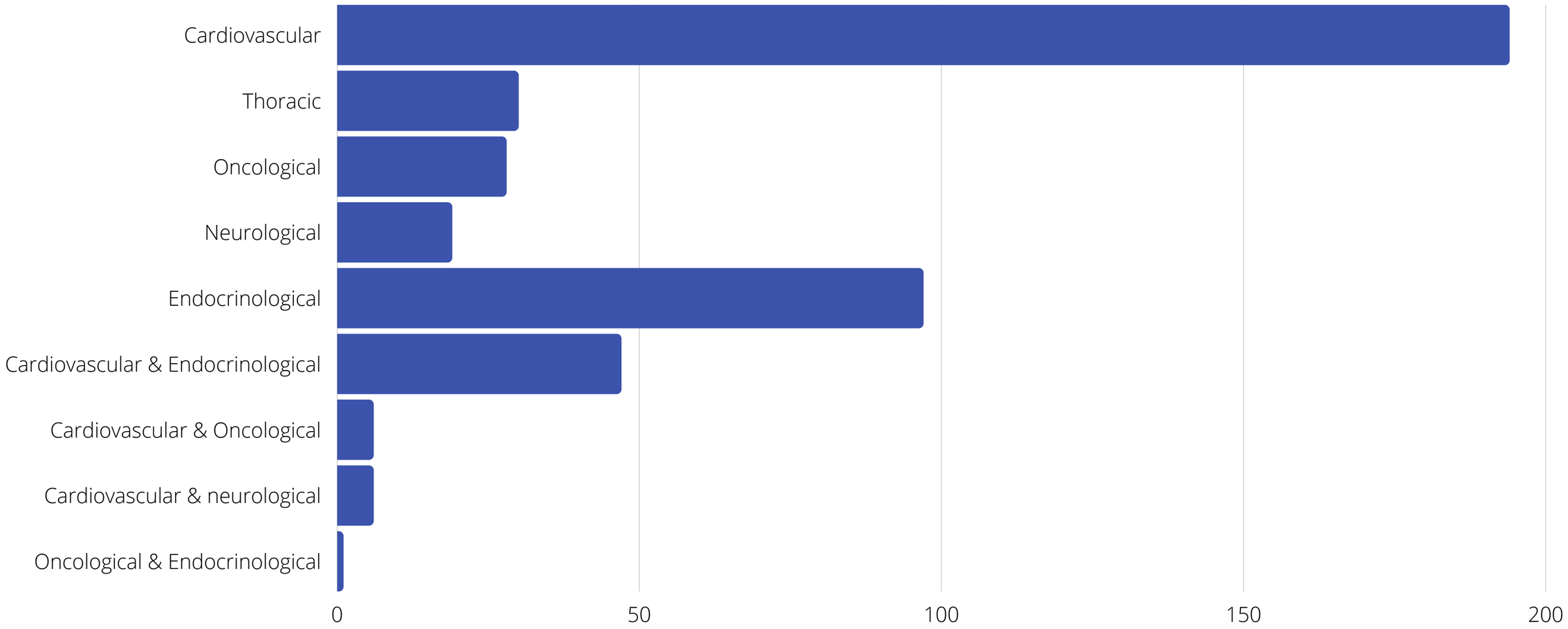
SARS-CoV-2 nucleic acids tests VS oncological diseases



SARS-CoV-2 nucleic acids tests VS endocrinological diseases



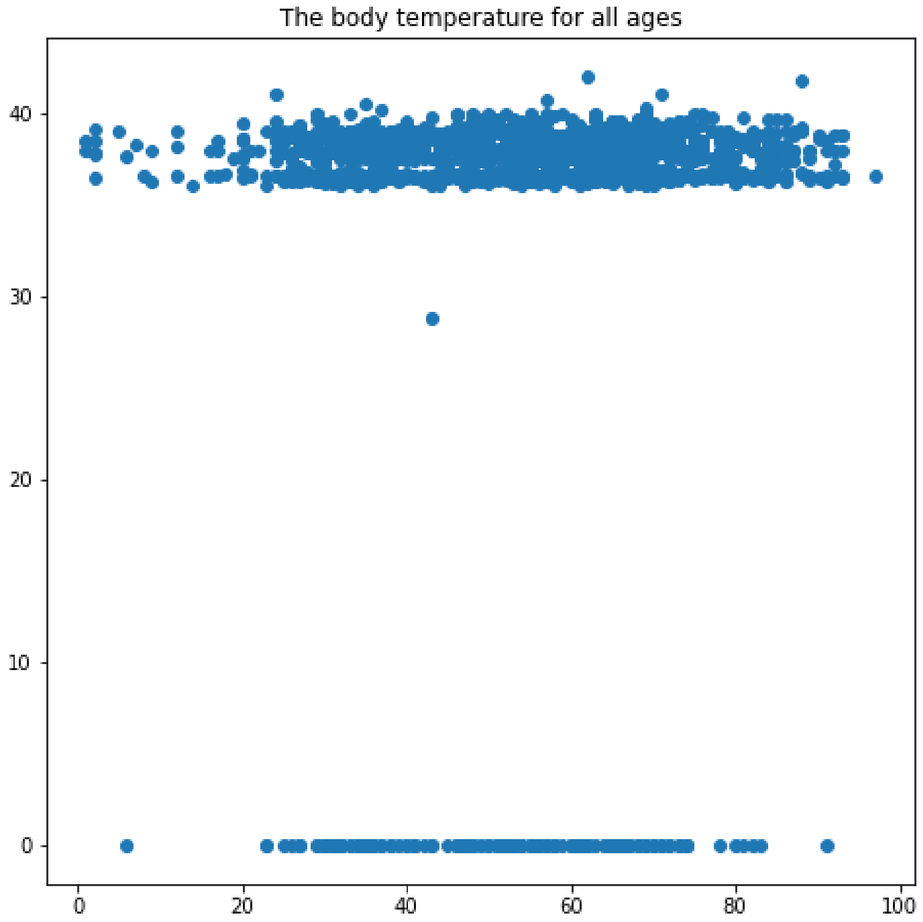
Positive SARS-CoV-2 nucleic acids tests VS Diseases



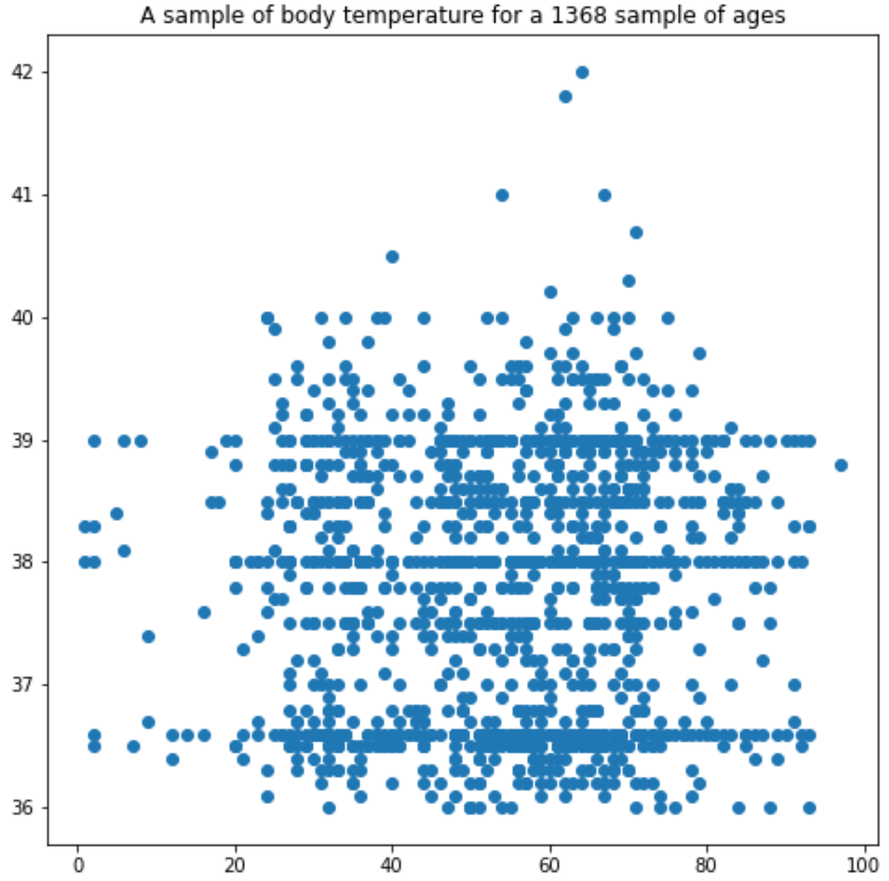
Body Temperature

Average temperature 38 °C

Population

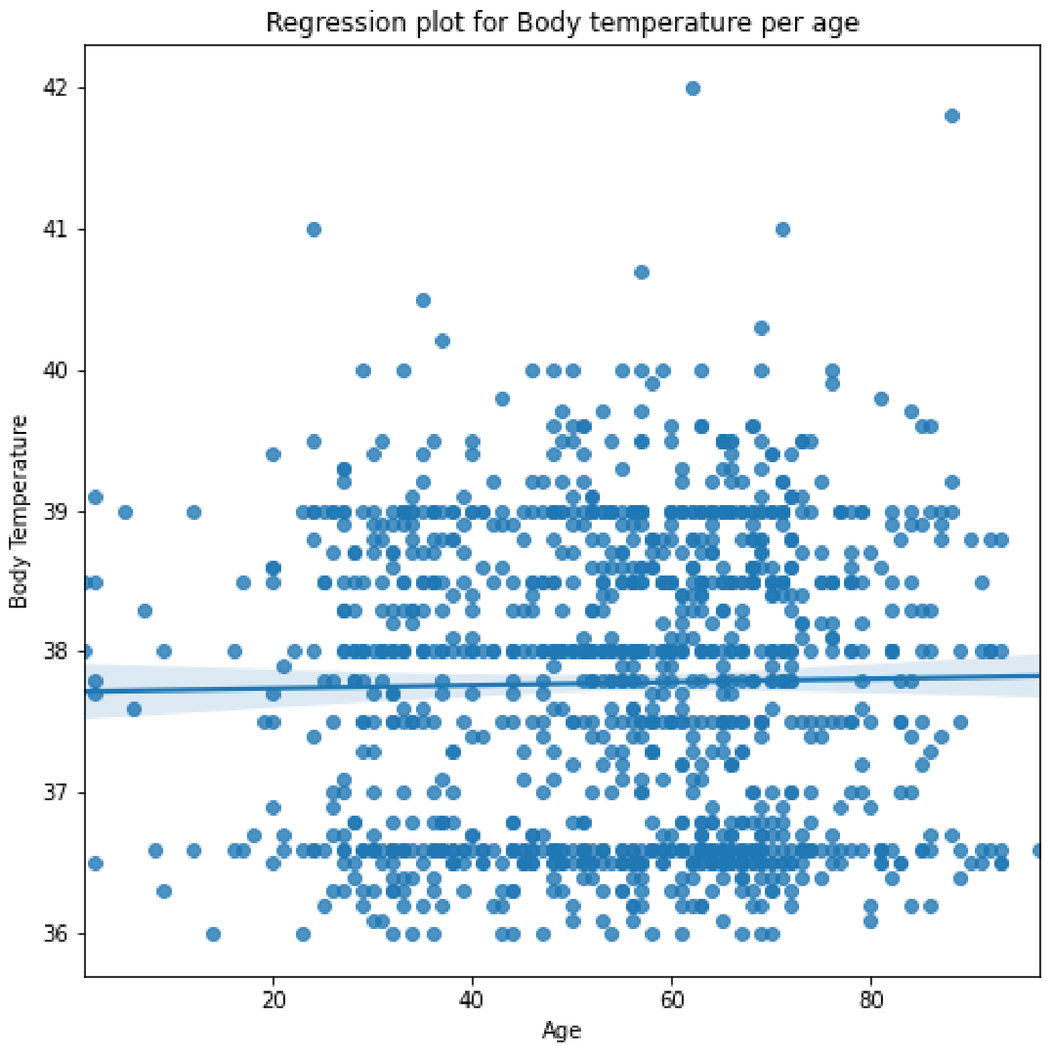


Sample

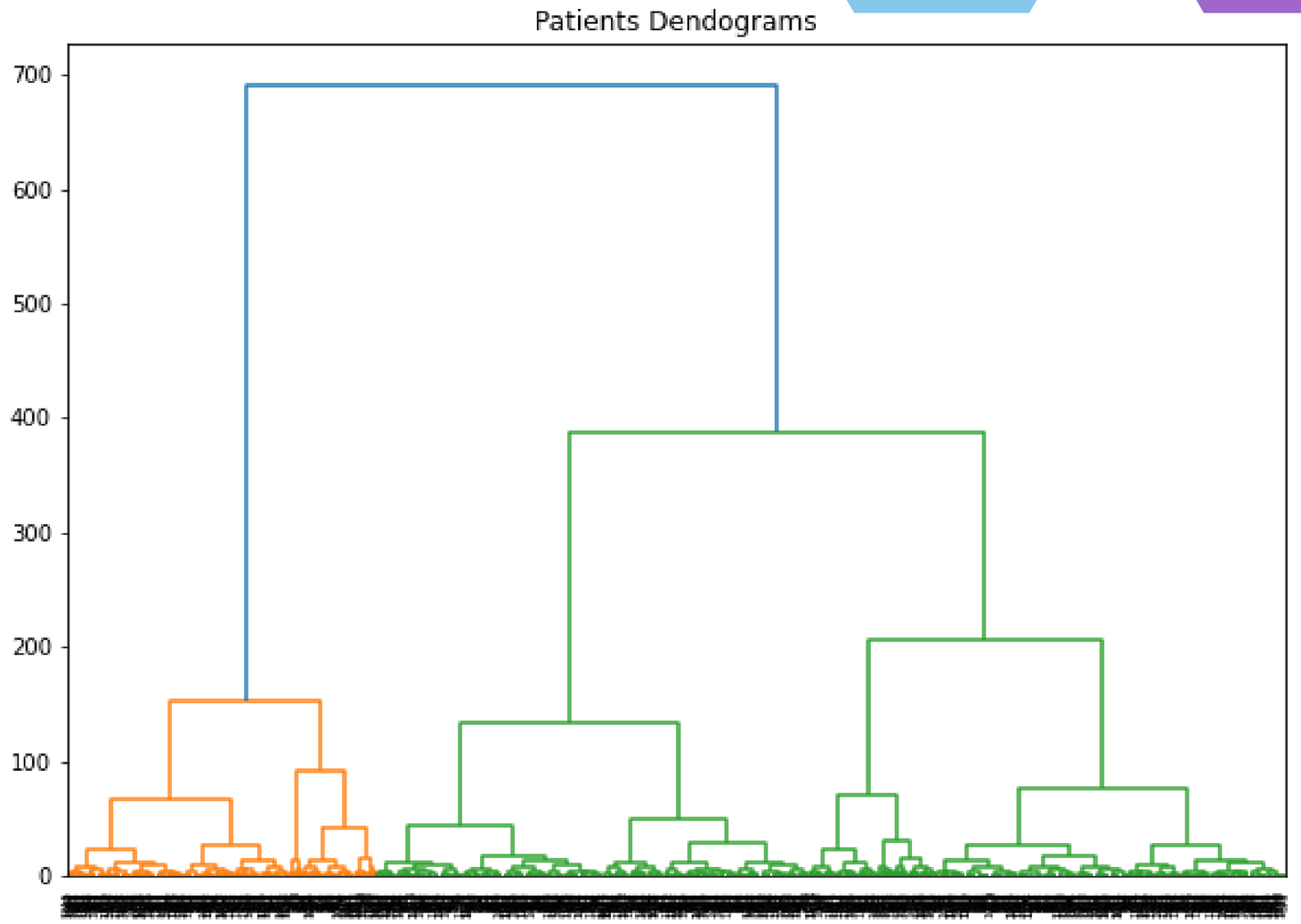


Regression plot:

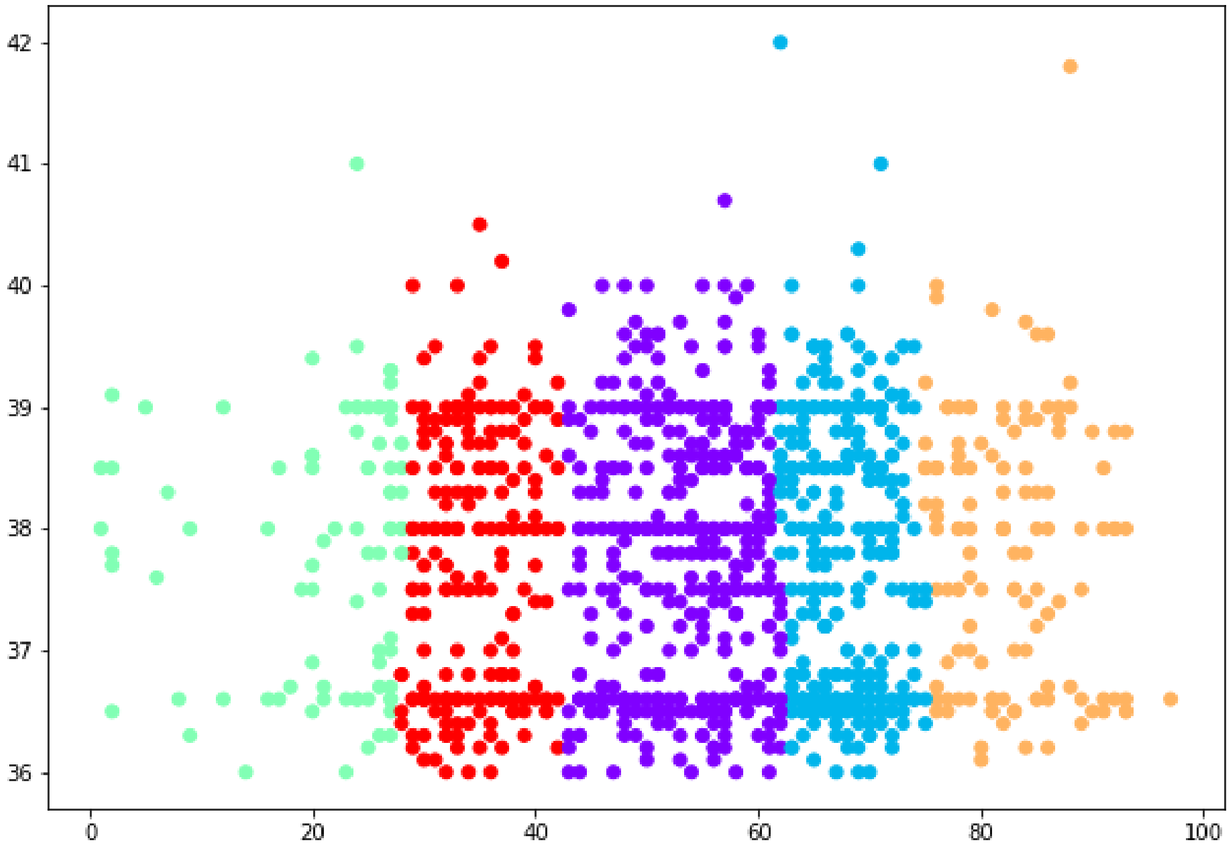
Body Temperature



Patients dendogram



Cluster plot: Body Temperature



Modeling: Prediction of future patients

Logistic Regression

- **Split the dataset in features and a target variable**

  Our target variable was SARS-CoV-2 nucleic acids tests

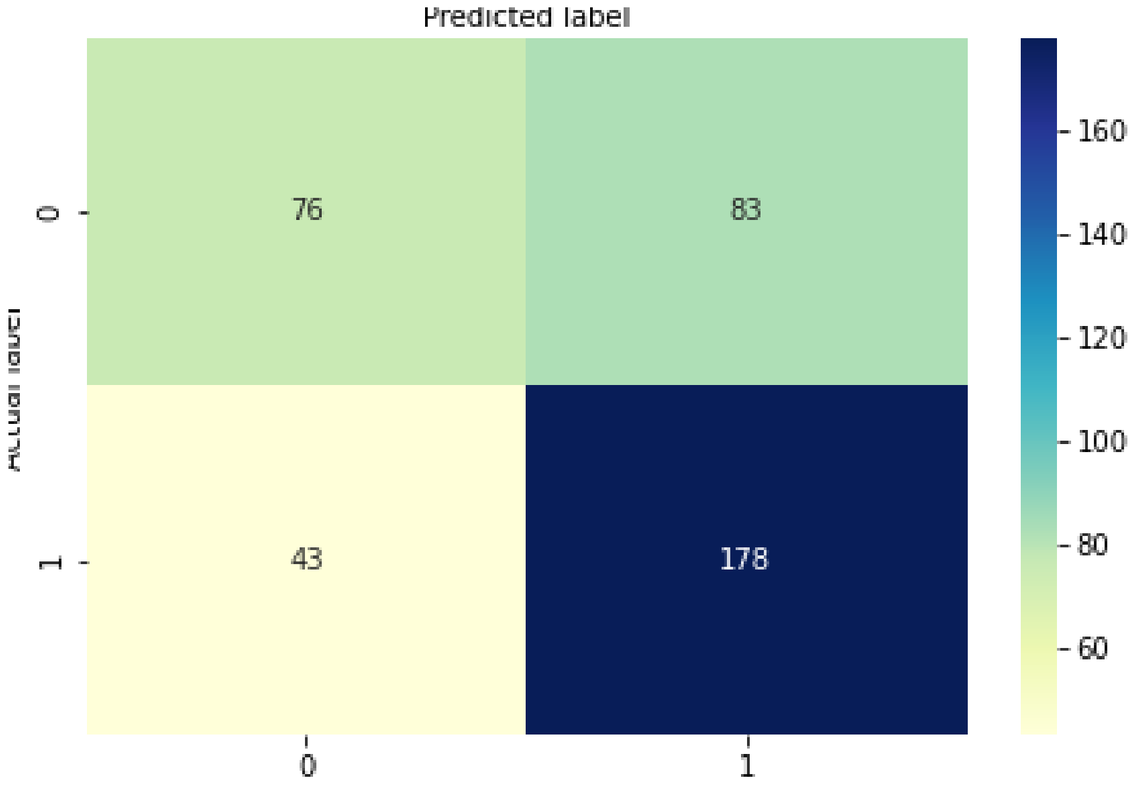
- **Split the dataset into train and test data**

  using train_test_split()

- **Instantiate and Fit the model**

  using solver='lbfgs' and .fit()

Model Evaluation using Confusion Matrix



Testing

```
PatientX= [[39.0, 0, 0, 1, 0, 0]]
ynew = logreg.predict(PatientX)[0]
if ynew:
    print("PatientX has a positive SARS-CoV-2 nucleic acids test ")
else:
    print("PatientX has a negative SARS-CoV-2 nucleic acids test  ")
```

[552] ✓ 0.2s Python

```
PatientX has a positive SARS-CoV-2 nucleic acids test
```

+ Code + Markdown

```
PatientX= [[37.0, 0, 0, 1, 1, 0]]
ynew = logreg.predict(PatientX)[0]
if ynew:
    print("PatientX has a positive SARS-CoV-2 nucleic acids test ")
else:
    print("PatientX has a negative SARS-CoV-2 nucleic acids test  ")
```

[558] ✓ 0.2s Python

```
PatientX has a negative SARS-CoV-2 nucleic acids test
```

+ Code + Markdown

Challenges

1

The web scrapping was challenging: the website had multiple pages and filters and couldn't be scrapped using BeautifulSoup like usual websites

2

The division of the underlying diseases was not easy because we needed to specify each disease and the list was long.

3

The clustering of data was difficult due to the lack of medical expertise and the variety of data

Future steps

Add more variables to the dataset

Clinical Features: Routine Blood Test, Inflammation, Blood Coagulation Test, Immune Cell Typing, etc...

Improve the accuracy of the model

training will be depending on more variables

Look for more correlations

Depending on clinical features

Thank you