Predicting Covid Cases through Machine Learning Algorithms

By: Loh, Nathan; Schultze, Nikolaus

Introduction and Problem Description

Hospital staffing is very minimal, and doctors are very busy. With the addition of Covid, this already limited hospital staffing is spread very thin, and they have to work extensive hours, causing them to get tired and worn out. The staff is already responsible for the care of their patients, but with the addition of Covid, there are more patients than ever, resulting in the staff caring for more patients and forcing them to spend less time per patient. As a result of the influx of Covid patients, the hospital staff needs more time to visit every patient and determine whether or not the person has Covid. Not only is this influx of Covid patients requiring these new Covid patients to be attended by the hospital staff, but it also requires extra resources. This influx also takes the hospital staff away from its other patients and uses resources that could be designated for the other patients. While there are other alternatives for testing for Covid, these alternatives, such as Rapid Testing, have flaws. While these rapid Covid tests give the Covid test results back quickly, they have an issue with inaccurate and false-positive results. There are more accurate Covid tests than rapid tests, but they take over 24 hours for the patient to get their result back. These long wait times result in patients having to quarantine while they await the results; however, only some patients will quarantine. The long wait time can result in a person with Covid continuing to adventure out into public and spread Covid to others. For these lab tests, if there is a higher number of patients sending in Covid tests, then the wait time for the results can be increased.

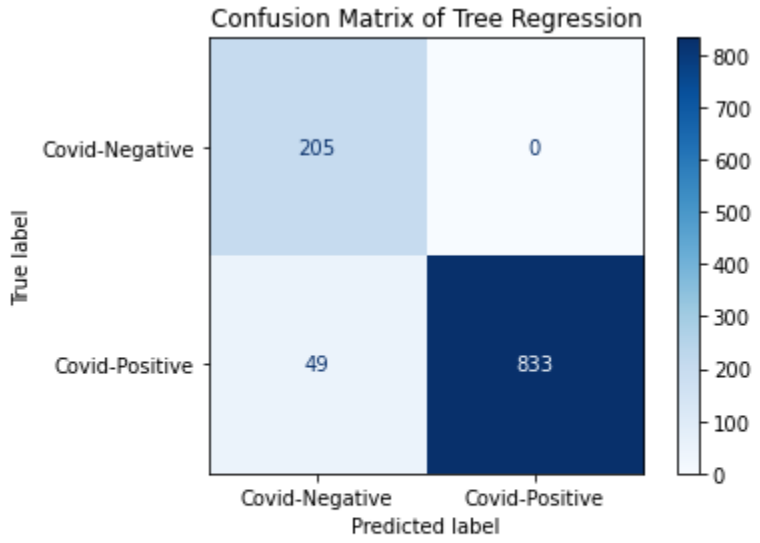
Description of the Data

The data contains a list of symptoms, including Breathing Problems, Fever, Dry Cough, Sore Throat, Running Nose, Asthma, Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hypertension, Fatigue, and Gastrointestinal. It also includes possible events in which a person could have received Covid, including Abroad travel, Contact with COVID Patients, Attended Large Gathering, Visited Public Exposed Places, and Family working in Public Exposed Places. The data also includes a COVID-19 column, indicating whether the patient has Covid. Each column in the original dataset represents “Yes” and “No,” which we converted into binary values, with Yes being converted to 1 and No being converted to 0. For the symptoms column, 1 represents the person having the symptom, and 0 represents the person not having the symptom. For the events columns, 1 represents the person attending an event of that type, and 0 representing not attending the event. For the Covid-19 column, 1 represents the person having Covid, and 0 represents the person not having Covid.

What have we done

Tree Regression Algorithm

Results



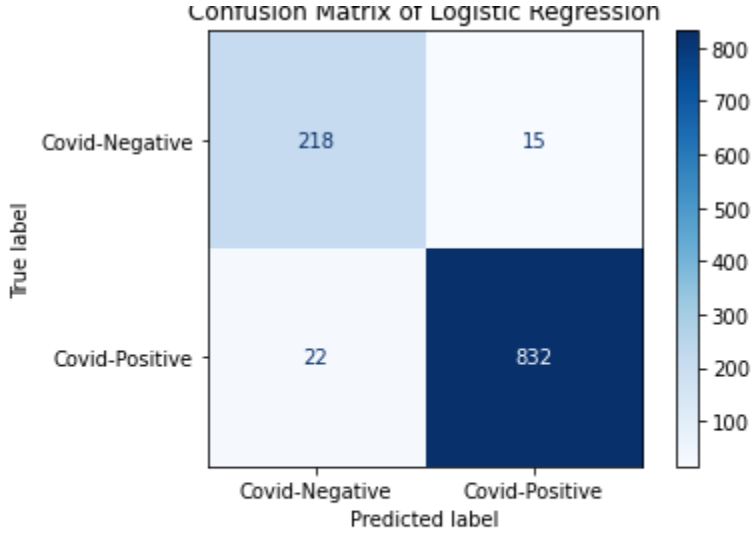

```
Accuracy = 95.49%
Precision = 100.0%
sensitivity = 94.44%
specificity = 100.0%
```

Test Number	Accuracy
Test 1	95.49%
Test 2	95.77%
Test 3	95.22%
Test 4	94.57%
Test 5	96.78%
Test 6	95.31%
Test 7	94.85%
Test 8	95.77%
Test 9	95.86%
Test 10	95.4%

Average Accuracy: 95.5%

Logistic Regression Algorithm

Results



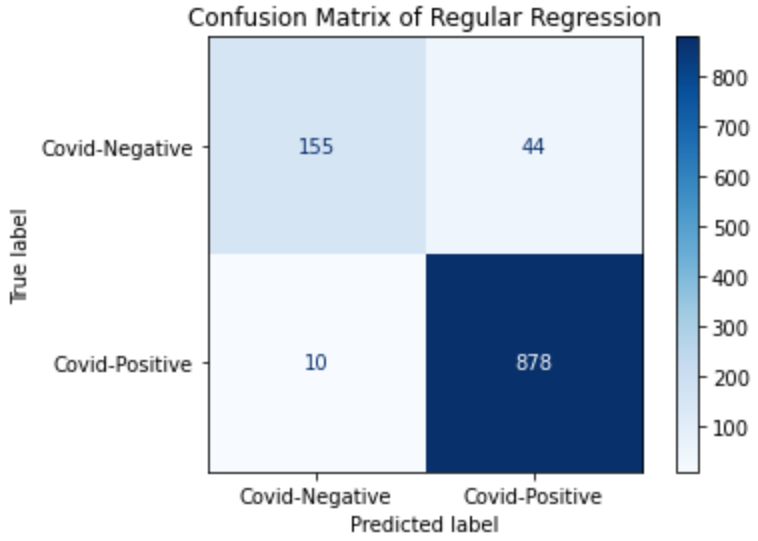

```
Accuracy = 96.6%
Precision = 98.23%
sensitivity = 97.42%
specificity = 93.56%
```

Test Number	Accuracy
Test 1	96.6%
Test 2	96.69%
Test 3	95.95%
Test 4	96.14%
Test 5	97.42%
Test 6	97.15%
Test 7	96.96%
Test 8	95.58%
Test 9	96.32%
Test 10	96.32%

Average Accuracy: 96.51%

<u>Regular Regression Algorithm</u>

Results




```
Accuracy = 95.03%
Precision = 95.23%
sensitivity = 98.87%
specificity = 77.89%
```

Test Number	Accuracy
Test 1	95.03%
Test 2	96.41%
Test 3	96.04%
Test 4	95.77%
Test 5	95.86%
Test 6	95.22%
Test 7	96.04%
Test 8	96.14%
Test 9	95.58%
Test 10	95.58%

Average Accuracy: 95.77%

What remains to be done

We finished coding the three algorithms: Tree regression, Logistic Regression, and Regular Regression. The code for each algorithm was created and developed based on the dataset we

were using. One potential work to be done would be to test the algorithms on alternate datasets. However, one issue with this is that because each algorithm was created for the specific dataset, we would have to alter each algorithm as the datasets potentially share a different list of symptoms which would impact the accuracy of each algorithm. Another potential work would be altering the algorithm to include K Fold Cross Validation. This addition could improve each algorithm's prediction and reduce the possible variation. However, this is optional as it would increase the computational costs of each algorithm and the current algorithms, which are already reasonably accurate. The Tree Regression Algorithm has an accuracy of about 95.5%, the Logistic Regression Algorithm has an accuracy of about 96.51%, and the Regular Regression Algorithm has an accuracy of about 95.77%. One remaining work that remains is the comparison of the different algorithms. So far, we have tested each algorithm ten times, placing the accuracy of each test in a table. We then took the average accuracy for the ten tests for each algorithm and compared them. We found that the Logistic Regression Algorithm performed the best, followed by the Regular Regression Algorithm, followed closely by the Tree Regression Algorithm. We still would like to perform more comparisons on these algorithms to determine best which algorithm is the most accurate algorithm for predicting Covid-19. This can include comparing the false-negative rates of each algorithm, where it would predict that the patient did not have Covid when they actually have it.