

AI MODEL TO PREDICT COVID TEST RESULTS.

by Mathew C Mammen

under the guidance of Ms.Xiaomeng Kong, Data Science Lecturer at Hofstra University
and Asst. Professor, Nassau Community College, New York

on September 20, 2023

Table of Contents

Abstract.....	3
Introduction	3
Background.....	4
Dataset	5
Methodology/Models	6
Results.....	8
Conclusion	11
Acknowledgments.....	11
References	12

Abstract

COVID-19 was declared a global pandemic by the WHO¹ on 11th March 2020. During this global pandemic, we faced a long wait for an appointment for Covid testing. The test results were even further delayed. Had there been a better system, we could have been prepared more and acted timely! The spread of Covid was also dependent on various factors like the living conditions, ethnicity, age, and health pre-conditions. This project was built using AI models for an early prediction of the test results. I used the CDC² public information available to build the logical regression to predict the Covid test results of individuals. Using these predictions, the potential red zones could be marked out to allow authorities to bring in extra care to the area to control the pandemic outbreak. This model can be extended to serve any such pandemic situations, if such a need arises.

Introduction

This research paper addresses the question of how early predictions of Covid results can be made for individual patients thereby helping them with the foresight to exercise caution. Previous test results published by CDC were used for this purpose. Using this model, we use the health and living factors of previous test cases and predict the results based on such factors. The purpose of this research is to extend this AI model to analyze and predict the results during any such pandemic. This could empower authorities to be proactive, taking into account vulnerable demographics susceptible to such a pandemic.

As the Covid pandemic unfolded, we quickly realized the profound impact it had on us. However, we had inadequate testing infrastructure to control the spread of the virus. Waiting days for an appointment and an additional two weeks for test results meant that by the time we knew we were affected, the infection had already permeated through our interaction. This delay created a cascade of consequences with the schools and offices shift to remote operations, and face-to-face interactions became a distant memory. With delayed results leaving us in the dark, the only recourse was widespread lockdowns to curb personal interactions. The imperative for more rapid testing solutions became evident, crucial for effectively managing and mitigating the impact of any future pandemics.

¹ WHO – World Health Organization.

² CDC – Center of Disease Control.

During the initial stages of Covid, The New York Times³ reported on a series of missed chances by the federal government to ensure more widespread testing, when containment would have been easier. The article sheds insight on how ethical alternatives are important in handling such pandemics.

Analyzing the CDC dataset, I attempted to establish connections among various factors that appeared susceptible to Covid infection. Ethnicity, age, and health pre-conditions like the body mass index factor, glucose reading, blood Pressure were analyzed to build a model to predict Covid infections. Designating red zones for authorities, determined by pre-screening results, is crucial for towns that are elderly segregated, who are more susceptible to potential outbreaks.

Many of our towns have segregated demographics. Some of them have elderly segregations, while some others have segregations by ethnicity. A new patient trying to get to a Covid test could be undergoing the pre-test screening using my model. Based on the factors described, my model was able to predict the test result, which is not assured to be accurate. However, such an early prediction based on the previous data on similar cases would be helpful to give sufficient warnings. This model could be a life saver for many such cases. How do we help the authorities to bring sufficient care in time? I tried to mark the red zones based on the pre-screen results. Based on these readings, the model can be extended for authorities to enable them to be proactive for hospitalizations and disseminate awareness in vulnerable towns.

Background

The COVID-19 virus has been responsible for over 2.5 million deaths across the world. This virus has severely affected the economy and the public health status of the people in various countries. Historically, similar viruses have sparked numerous outbreaks, exemplified by the Extreme Acute Respiratory Syndrome Coronavirus (SARS-CoV) in China and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in the Middle Eastern countries. The COVID-19 virus was initially identified in China in December 2019, after which it swiftly transcended borders, traversing the globe as individuals came into contact with the infected. It is noted that the coronavirus significantly affects the health of people and even causes death, either directly or through exacerbating pre-existing health problems. As a large proportion of people have been affected by the COVID-19 pandemic

³-Fin, Sheri and Baker, Mike "It's Just Everywhere Already": How Delays in Testing Set Back the U.S. Coronavirus Response." The New York Times, 10 March 2020.

throughout the world, and there is no cure available for the disease, it becomes important that we develop early predictions to prevent the spread of infection.

Many researchers, including data scientists, have been working intensely to design prediction models that highlight the probable activities of this virus. Most of them focus on predicting the Covid cases by States and territories based on daily incidence reported. The main contribution of this paper is that it proposes a prediction model which can present individuals' awareness by early prediction of potential positive COVID-19 cases. The model drills down to segregated demographics based on these potential cases, thereby warning authorities of potential spikes in such towns.

Dataset

For our project, the dataset that we used to perform the analysis was downloaded from the CDC website. Various data cubes on Covid testing were published in the CDC website <https://data.cdc.gov>. We used them to prepare the dataset for machine learning and testing. This data set was hosted in a MySQL relational database and was served to the Python program used to build the logical regression model. A summary of the data used is as follows.

SUMMARY OF THE DATA:

PersonID - ID Number of the patient/Individual.

Age - Age (years)

Ethnicity - Social/cultural/national group

Ethnicity_group - Group sorted alphabetically and numbered

<i># Ethnicity</i>	<i>Ethnicity_Group</i>
<i># African American</i>	<i>1</i>
<i># Asian</i>	<i>2</i>
<i># Hispanic</i>	<i>3</i>
<i># Latino</i>	<i>4</i>
<i># Native American/Alaskan Native</i>	<i>5</i>
<i># Other</i>	<i>6</i>
<i># Pacific Islander</i>	<i>7</i>
<i># white</i>	<i>8</i>

Fever - Temperature recorded in degrees Fahrenheit

Asthma - Breathing issue recorded - Positive or negative

Asthma_group - Breathing issue recorded - Positive = 1 or negative = 0

BloodPressure - Diastolic blood pressure (mm Hg)

HeartRate - Beats per minute

```

### Glucose - Plasma glucose concentration
### BMI - Body mass index ( $\frac{\text{weight}}{\text{height}^2}$  height2weight in kg/m)

### Outcome(COVID Test Result - Class variable (0 - healthy or 1 - diabetic)

```

To better visualize the data set, given below is a snapshot of the data analysis.

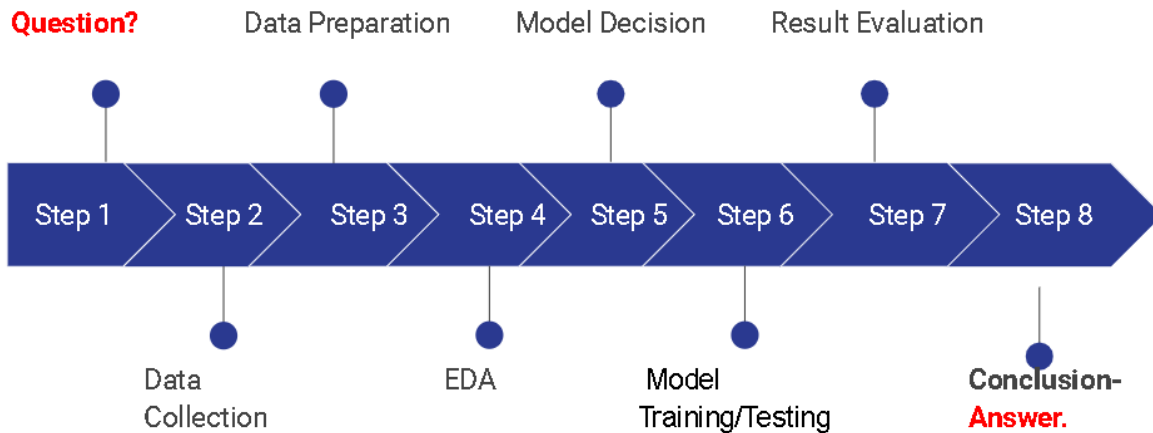
AI Model to Predict Covid test results using Logistic Regression/Binary Classification.													
Data Analysis.													
covid_analysis													
	PersonID	Age-years	Fever-temperature	Ethnicity	Ethnicity_Grp	Asthma	Asthma_Grp	BloodPressure-mm_hg	HeartRate-beatspermin	Glucose-SugarLevel	BMI-BodyMassIndex	COVID-TestResult	
0	1340386	21	96.0	African American	1	negative	0	48	97	77	47.150986	0	
1	1807582	21	96.0	African American	1	negative	0	81	97	77	21.246453	0	
2	1680684	21	96.0	African American	1	negative	0	49	97	77	41.289096	0	
3	1560929	21	96.0	African American	1	negative	0	87	97	77	44.686495	0	
4	1138937	21	96.0	African American	1	negative	0	82	97	77	21.999207	0	
...
13995	1187803	59	100.2	Asian	2	positive	1	53	98	109	33.186520	1	
13996	1843979	59	100.4	Asian	2	positive	1	61	98	111	33.854893	1	
13997	1201886	59	100.2	Asian	2	positive	1	95	98	112	27.430496	1	
13998	1228510	59	100.4	Asian	2	positive	1	50	98	115	34.692154	1	
13999	1416388	59	100.4	Asian	2	positive	1	62	98	118	34.504598	1	

I split the dataset as the training and testing datasets. Seventy percent of the data set was used to build the training set and the remaining thirty percent was used as a testing dataset.

Methodology/Models

To solve the research problem, we had to analyze the dataset to extract what is required for our model. There was a steady inflow of data on the CDC website and hence we got sufficient data to build our training and testing datasets. The below diagram explains the steps of the methodology used for solving the problem.

AI Model to Predict Covid test results using Logistic Regression/Binary Classification.



Analysis of the data was performed in various angles by grouping them by various parameters. For example, I tried to plot the age – ethnicity combination to see the Covid results.

```
covid_analysis.groupby(['Ethnicity', 'Age-years']).count()
```

Glucose level by ethnicity was a key check to analyze the data.

```
covid_analysis.hist(column='Glucose-SugarLevel', by='Ethnicity')
```

After performing the initial analysis, I moved on to preparing the data sets for training and testing. scikit-learn package was used to prepare the train_test_split function. Using this package we can get a statistically random split of TRAINING and TEST data.

At this stage, I split the data into 70% for TRAINING and held back 30% for TESTING.

The next step was to decide the model to be used for training. I used a logistic regression model on the training set `x_train`, `y_train`. I used solver `liblinear`, a library used for large Linear Classification. This Uses a coordinate descent algorithm. Coordinate descent is based on minimizing a multivariate function by solving univariate optimization problems in a loop. In other words, it moves toward the minimum in one direction at a time.

```
model = LogisticRegression(C=1/reg, solver="liblinear").fit(x_train, y_train)
```

I also checked the accuracy of the predictions - what PERCENTAGE of the labels did the model predict correctly? Mathematically it represents the ratio of the sum of true positives and true negatives out of all the predictions.

```
Accuracy: 0.9921428571428571
```

The accuracy reported is good enough to proceed. But What are the other ways that we can check if this model is good enough to use? I used the Classification report that is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False? Precision is the ability of a classifier not to label an instance positive that is negative. Recall is the ability of a classifier to find all positive instances. The F1 score is a weighted mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

```
Overall Precision: 0.9936373276776246
```

The Model was checked and was found good to proceed for further analysis and results.

Results

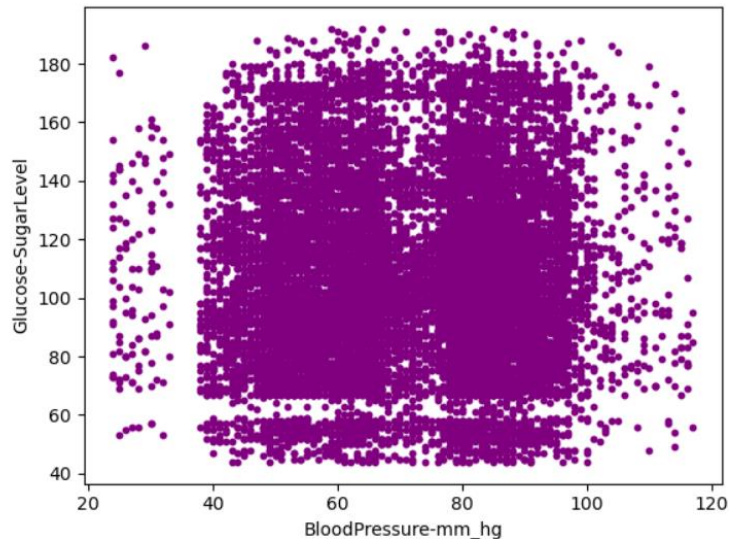
When creating results to analyze to address my original research question of—
“How would we use the earlier results to predict the result on Covid testing? Can we plot the red zones based on results of different demographics”?

I created charts and visualizations to analyze the dataset. Different combinations of factors and test scores were plotted against each other. The first comparison that was conducted examined the potential correlation between the ethnicity against the positive cases reported.

In the next step I included Glucose level reading as a parameter to the above analysis. I also created a scatter plot by choosing glucose level (x) against blood pressure reading (y).

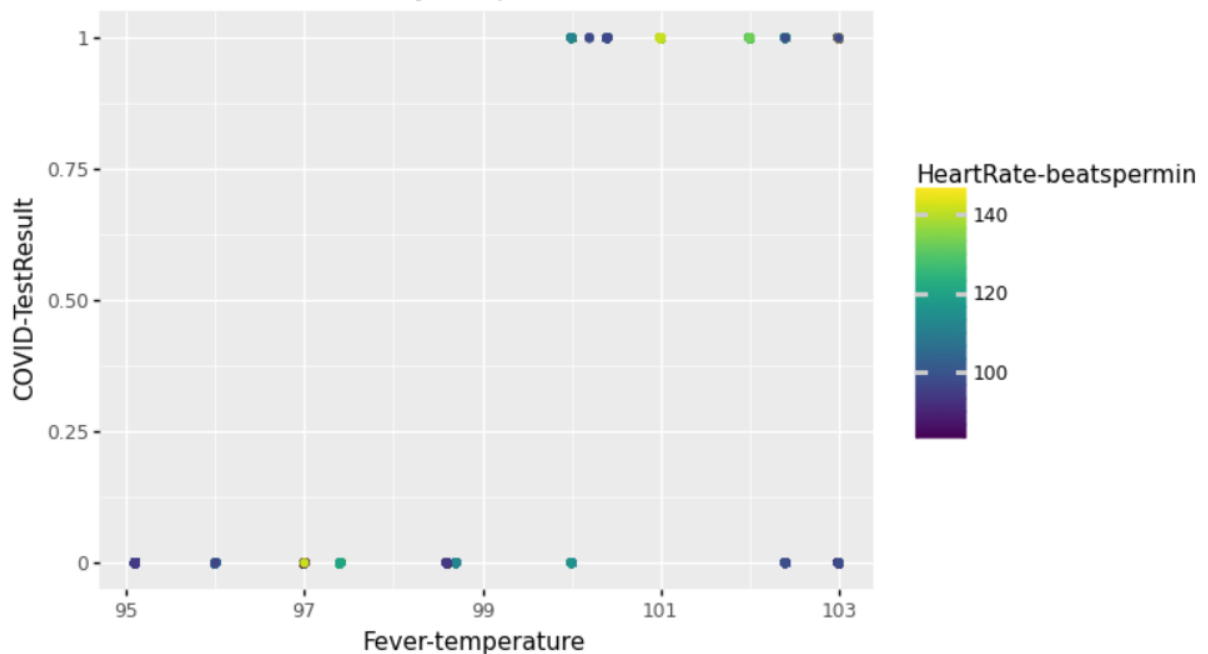

```
In [36]: # Create scatter plot for min_height and min_weight
covid_analysis.plot.scatter(x='BloodPressure-mm_hg', y='Glucose-SugarLevel', s = 10, c='purple')

Out[36]: <Axes: xlabel='BloodPressure-mm_hg', ylabel='Glucose-SugarLevel'>
```

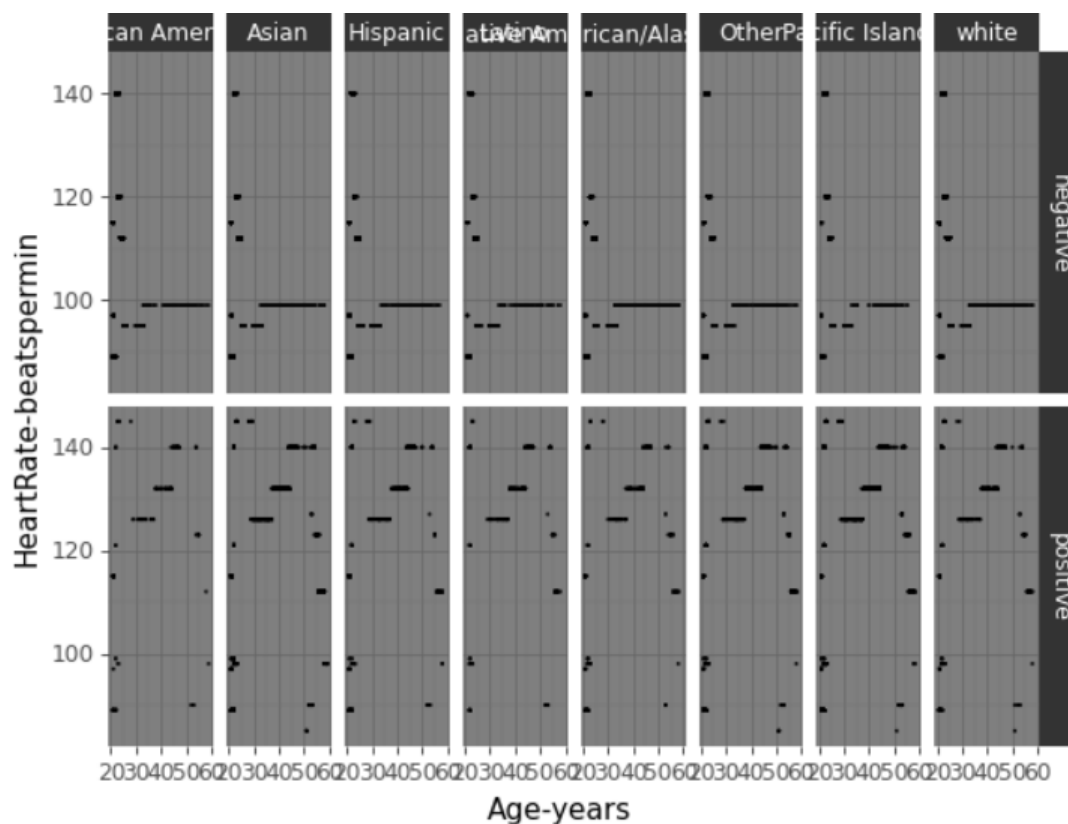


I also combined certain key factors to perform further data Investigation - Individuals with Fever and covid results including the heart rate reading to the plot.

COVID-TestResult Versus Body-Temperature - (Heart Rate Included)



Similar Plots were prepared bringing in ethnicity, heartbeat readings, age in years to compare the Covid results.



Patient Sample Data Testing

Enter the New sample data that you want to check:
 [30.0, 96.0, 1.0, 1.0, 60.0, 97.0, 79.0, 45.33189957]

Predicted class is 0

Predicted - Covid Testing Result = 0 = Negative

Enter the New sample data that you want to check:
 [40.0, 101.0, 3.0, 1.0, 69.0, 132.0, 136.0, 40.81699943]

Predicted class is 1

Predicted - Covid Testing Result = 1 = Positive

Conclusion

This research named AI Model to Predict COVID Test Results was conducted to check if an early prediction of test results during a pandemic outbreak is possible. I used the published data on the CDC website on Covid test results which do not use any personal information. However, various factors like ethnicity, age, body temperature, height, weight, heartbeat reading, oxygen level, glucose reading, blood pressure were given along with the Covid test result. I used these test results with some alterations as the data set for my model. Binary classification on Logical Regression was used to build the model using Python as Programming Language. Python scripts for Exploratory data analysis EDA-marginal and GGLOT for EDA-pairs were used. Data was visualized using the Plotnine python package, which provides a grammar of graphics.

I would like to take this research to a further level. The goal is to extend this research to prepare a ready to use model for the local authorities to support pre-screening while an individual is doing a test during a pandemic. Along with the prediction of results for the patient to be cautious until the actual test results arrive, I wish that the new model would mark spots to keep track of the potential positive cases in a town. This would help the authorities be notified of a possible outbreak and be prepared to handle such an emergency.

Acknowledgments

My first step into AI, Machine Learning and Data Science began with a summer course taught by MIT Alumni. I'm happy that I took this course right after my freshman year, which gave me inspiration to think and explore deep into this area of computer science.

I would like to thank all the instructors who gave me a big picture of what I can expect with data analysis. I thank CDC (data.cdc.gov) for providing the dataset (open data repositories) that was used for my research. This quality data gave accurate predictions.

I thank my Data Science research project instructor (Ms.Xiaomeng Kong -Data Science Lecturer at Hofstra University / Asst. Professor at Nassau Community College). My sincere thanks to Ms. Kong's guidance and support throughout this research.

I also thank my parents for supporting me for academic success. Love you Mom and Appa!

References

Textbook:

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.
2. Thinkstats: Probability and Statistics for Programmers
3. Python Data Science Handbook

Open Data Repositories: Dataset

Data Access - Public-Use Data Files, Documentation (cdc.gov)
<https://data.cdc.gov/browse?tags=covid-19>

Introduction to Statistical Learning:

An Introduction to Statistical Learning (statlearning.com).
<https://www.statlearning.com>.

Scikit Learn – Machine Learning in Python. <https://scikit-learn.org/>

Wikipedia Covid - Link: <https://en.wikipedia.org/wiki/COVID-19>.

The New York Times Article -

<https://www.nytimes.com/2020/03/10/us/coronavirus-testing-delays.html>