Covid-19 Classification and Regression Tree

By Nikolaus Henderson,
Cristian Ocampo,
Gerardo Valenzuela

Background to Covid-19’s Differing outcomes

COVID-19 is an RNA virus that causes a potentially lethal cytokine storm which can damage vital organs. It was first detected in 2019, and according to the CDC, as of May 17, 2021, it has been involved in 571,488 deaths in the United States alone. This number rises to 3.38 million deaths worldwide.

However, not all COVID-19 infections yield such drastic consequences. Indeed, some COVID-19 patients are asymptomatic while others quickly deteriorate to critical condition. These critical cases have driven up the need for ICU beds. Within these two extremes, however, there exists an assortment of outcomes. Our system focuses on asking important questions regarding the facts of patients like health, age, sex, and pre-existing conditions and how these facts relate to fatality.

Our Research

In our research we sought to implement a machine learning classification algorithm to determine, based off previous data, whether an infected individual would experience a fatal case of COVID-19. The algorithm of choice is the Classification And Regression Tree (CART) algorithm.

A proof of concept has already been completed using “Pseudo-Data”. Depending on the quality and quantity of data, our methods should yield a reliable classification of features that represent ‘at-risk’ patients.

Additionally, it could be used as to better determine treatment order replacing a first-come-first-serve system.

Finally, it is possible to implement other algorithms that may be better suited to the data collected.

CART Background and Overview

- Classification and Regression Trees are useful in situations where the answer you are trying to find is categorical.

- A true or false question about a feature is asked on each object of a dataset to partition it into two groups. The question that gets asked is always the one that results in the most information gain.

- The process is repeated, and new question is asked on each partition, until asking questions no longer results in information gain, or there are no more questions to ask.

- The resulting decision tree is used to classify an object based on its features.

C

What Question to Ask?

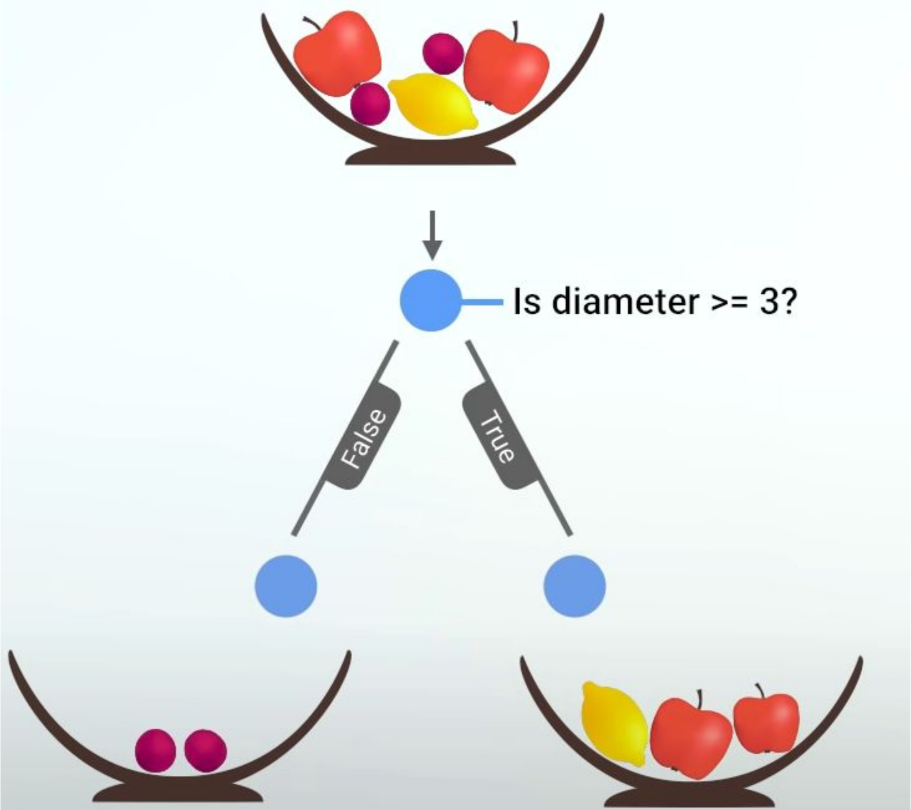
- Asking just *any* question may not lead to a significant conclusion

- Gini Impurity. $$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- We must ask the question that results in the most information gain.

  Impurity Before - Impurity After

Example CART



Data

G

- Our data is pseudo-data that is idealized such that it works as a "proof-of-concept" for the CART algorithm. We have taken extra care to include data that has features such as fatality, pre-existing conditions, and demographic data.

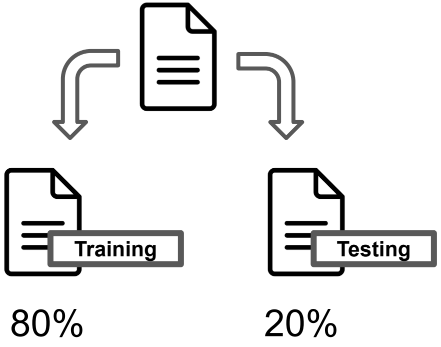
- Additionally, there are other methods of obtaining data which can yield important results. Through collecting contact tracing data, researchers are immediately familiar with first symptoms of patients who are exposed to infected people, shedding more light on early symptoms and risks. COVID-19 test site data also allows for raw data collection which has not been tampered with. This type of raw data has its benefits in that this data is often the most rapidly available for models, however it does not include data pertaining to fatality. Hospital data is further developed and filtered, making it ideal for many training scenarios, especially for a model dealing with fatality such as ours.

C

Avoiding Statistical Pitfalls

How data was\will-be partitioned to hold statistical significance: When creating a training set and a testing set it’s possible to accidentally create one(s) that are not homogenous or similar to each other.

To mitigate this a larger dataset would be preferable and building multiple trees with differing partition methods and comparing them would help prevent faulty-partitioning even more.



Sex	Age	Underlying conditions	Label
Female	24	Yes	Lived
Female	28	No	Lived
Male	25	No	Died
Female	44	Yes	Died
Female	32	No	Lived
Female	36	No	Lived
Male	37	Yes	Lived
Female	26	No	Lived
Male	45	Yes	Died
Female	33	Yes	Died

Training

Sex	Age	Underlying conditions	Label
Female	24	Yes	Lived
Female	28	No	Lived
Male	25	No	Died
Female	32	No	Lived
Female	36	No	Lived
Male	37	Yes	Lived
Female	26	No	Lived
Male	45	Yes	Died

Testing

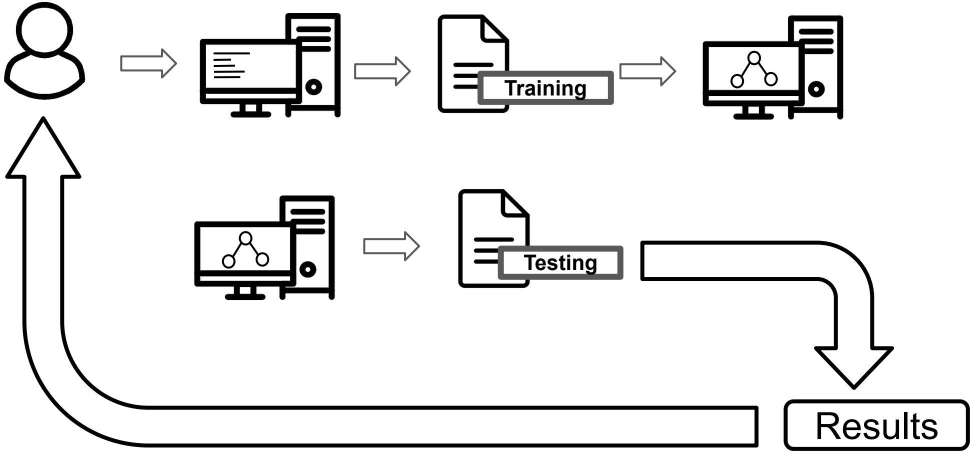
Sex	Age	Underlying conditions	Label
Female	44	Yes	Died
Female	33	Yes	Died

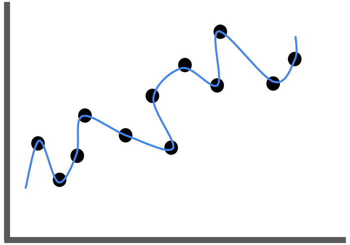
Avoiding Statistical Pitfalls

Overfitting via P-hacking/Data-debriding:

It is possible for the person to unintentionally become the element that allows the ML model to view the testing data.

To avoid this we will set aside a partition of our data as a “final test” to determine if any adjustment(s) of our model resulted in overfitting.



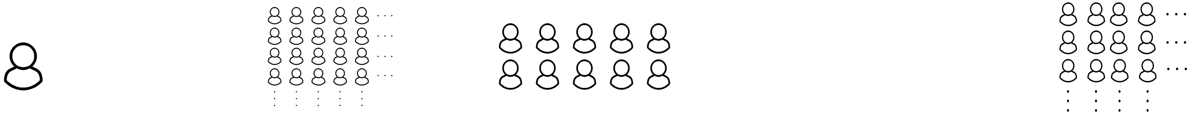


Avoiding Statistical Pitfalls

Accuracy paradox: The basic metric of accuracy (Correct/Total) can lead to some very correct-faulty models as it does not necessarily include False-positives and False-negatives.

To avoid this we will be using Cohen’s Kappa as our metric of accuracy.

$$k = \frac{p_o - p_e}{1 - p_e}$$



1 in 30,000,000 300,000,000 10 10,000

$$\frac{\text{Correct}}{\text{Total}} = \frac{300{,}000{,}000 - 10 - 10{,}000}{300{,}000{,}000} = \frac{299{,}989{,}990}{300{,}000,\ 000} = 0.999966 \approx 99.997\%$$

Applications

- Assist hospitals and medical facilities in determining if a patient requires immediate medical attention or can recover at home, which patients are at the most risk, and the order in which to direct patient care

- Adapted to assist researchers in determining what groups of people the vaccine is appropriate for

- Adapted to assist people in determining their chance of catching the virus, if extra precautions are necessary, and the likelihood of recovering at home as opposed to in a hospital

Conclusion

Thus far we have shown what methods we utilized for our research and how this classification system can predict whether an infected individual would potentially face grave outcomes. We have also covered how we have circumvented statistical pitfalls regarding our data and algorithm choice. Our model features benefits such as utilizing measures which can be replicated quickly for future pandemics, and most importantly, it can potentially prevent needless death and hospitalizations. However, there are different machine learning algorithms which may be used for this problem such as Neural Networks, Naive Bayes, among others. We will use these algorithms and create models to compare against CART to determine the best.

Thank you for joining us today

&

Thanks to the MREC^3, a project funded by the NSF

Q&A