Parker Barnes

STAT 426 Final Project Proposal

Question: How accurately can we predict if someone will die of Covid-19 based on basic demographic data (age range, sex, race/ethnicity, etc)? How accurately can we predict if they will be hospitalized? What are the major risk factors associated with Covid-19?

Data Source: I am scraping my data using the Center for Disease Control API (<https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf>).

Target: **death\_yn** – did the patient die?

Features:

1. **cdc\_report\_dt** – date case was reported to CDC
2. **pos\_spec\_dt** – date of first pos specimen collection
3. **onset\_dt** – symptom onset date, if symptomatic
4. **current\_status** – Laboratory-confirmed case or probable case?
5. **sex**
6. **age\_group**
7. **race\_and\_ethnicity**
8. **hosp\_yn** – admitted to hospital or not
9. **icu\_yn** – admitted to icu or not
10. **medcond\_yn** – did they have an underlying comorbidity or disease?

I am planning on creating a few more features. There are over 5 million cases in the dataset. I plan on taking a random sample of a few hundred thousand to limit computation time.

I will answer the research question by fitting a model to these variables and seeing how accurately I can predict if the patient died or not. I will look at which features are most significant to assess the risk factors of Covid-19.