COVID 19 Hospitalization Predictions

Stacie Allen   
Kennesaw State University   
Kennesaw, Georgia, United States  
[sallen16@students.kennesaw.edu](mailto:sallen16@students.kennesaw.edu)

Stephen Patton  
Kennesaw State University   
Kennesaw, Georgia, United States  
[spatto14@students.kennesaw.edu](http://spatto14@students.kennesaw.edu)

Anna Kurenkov  
Kennesaw State University  
Kennesaw, Georgia, United States  
[akurenko@students.kennesaw.edu](http://akurenko@students.kennesaw.edu)

Winston Kone

Kennesaw State University   
Kennesaw, Georgia, United States  
[akone1@students.kennesaw.edu](http://akone1@students.kennesaw.edu)

Abstract—In this class, we learned about Machine Learning and how it can be applied to analyze data. Data analysis helps us transform and manipulate information from data and give it meaning. To do so in this research project, we utilize an open-source development environment called “Jupyter Notebook” to apply machine learning techniques to automatically learn our data and draw out predictions for our Covid research study.

# **Introduction**

In December 2019, Coronavirus, also known as Covid-19, was first discovered in Wuhan, China [1]. It spread rapidly around the world and impacted the lives of everyone in a multitude of ways. During times of health crisis, the hospitals provide a critical element. A health crisis of this magnitude affects the hospitals and their ability to provide care for the communities [2].

This study will analyze data collected for positive COVID patients and the factors that lead to extensive medical treatments.  In an attempt to predict the number of cases, massive amounts of data have been collected that will lead to medical treatments, including hospital stays and ICU care.

Our team applied machine learning algorithms to research a very large Covid-19 dataset of confirmed positive cases. We analyzed the “Covid-19 Case Surveillance Public Use Dataset” that compiles data from US cases reported to the CDC [3]. We explored the predictions from specific machine learning techniques to explain the correlation among the hospitalized demographics and its death rate.

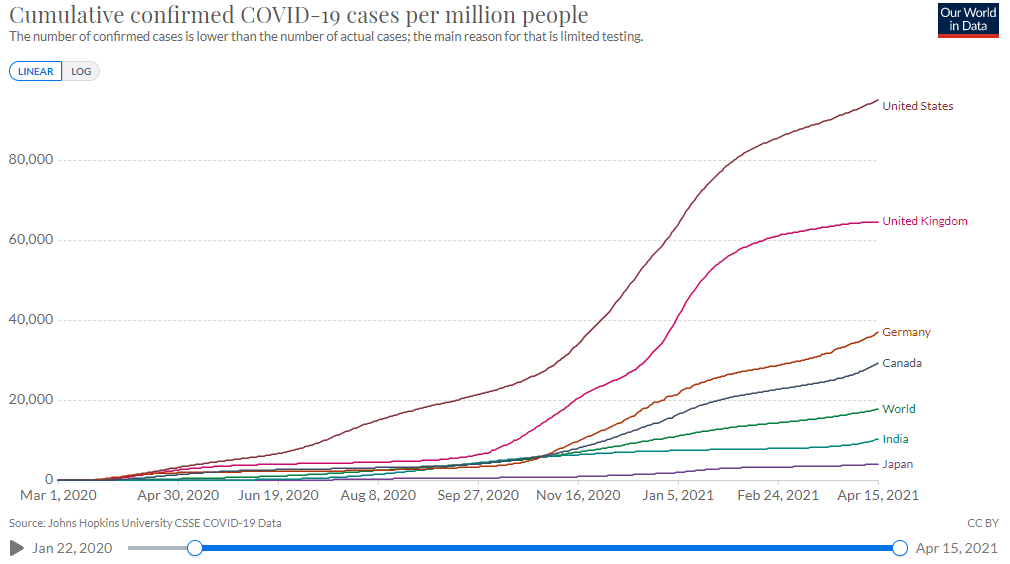
This research paper will be organized in parts. Part II will look at research that has already been conducted in this area. Part III will present our approach and the dataset used. In Part IV, the results will be presented and analyzed.

*Keywords—COVID, Medical Condition, Naïve Bayes, KNN, SVM, Logistic regression*

# **related research**

## **General research and Data Collection**

Ever since the Covid world state of emergency was announced in January 2021 by the WHO [4], Covid has been extensively studied to isolate statistics, demographics, and risk factors. At the time of writing, worldwide Covid deaths have reached 379.6 million [5], with deaths in the US surpassing every other developed country as well as the world average daily number of Covid deaths early on [6] as seen in Figure 1. Therefore, our research focusing on US Covid data looks at the population most impacted by Covid worldwide.

[Figure one - visualization of cumulative Covid cases per million people, with USA statistics exceeding the World’s average]

Many research papers have been written about it. An early 2020 paper by Wang, Zhang et al identified having preexisting conditions, old age, and male sex to be risk factors for having a severe or deadly Covid diagnosis by looking at hospital patients in Wuhan. [7] More recently in 2021, a similar study was conducted by Mikami, T., Miyashita H. et al, which studied Covid-positive New York hospital patients. The findings of the latter study were consistent with the previous one - patients who are older than 50, more likely to be male, and with a history of preexisting conditions or smoking, adding a lifestyle component. [8] From either of those studies, we cannot isolate demographic factors beyond sex, though we are also considering the factors such as race in our research. Mahajan, et al conducted an overview of race demographics in relation to Covid in the US. The study found that there is a positive correlation between the population of African Americans and Asians in a county and positive Covid cases or deaths, the inverse being true for White populations.[9]

Given all of those previous results, we form the hypothesis that

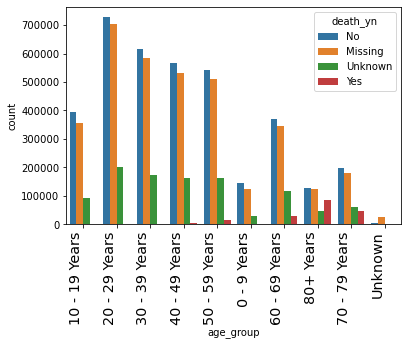
a. hospitalization and medical conditions are correlated with Covid deaths

b. old age is correlated with Covid hospitalization and death

c. certain race groups such as African American, as well as (to a lesser extent) Male sex, correlate (to a lesser extent) with Covid positive diagnosis, hospitalization, and death.

## Machine Learning in COVID Predictions

The Kaggle public-use Covid data set used in this research presented many machine-learning case studies. User Harsh Bhatnagar proved a code which presented some interesting predictions. [10] His findings, shown in Figure 1, confirmed that age over 50 years old correlates with a number of Covid deaths.



[Figure 1 - Bhatnagar’s findings about age groups in relation to Covid-19 deaths]

It is useful to take note of the classification methods used in project files as Bhatnagar’s as well as the relative accuracy scores, which we used a benchmark to build and compare our research against. He has achieved a decision tree weighted score of .78 and a Logistic Regression score of 0.73. This is consistent with Athar’s results when using the same classification techniques [11].

# **Approach**

In this research, our team selected a large dataset that required cleaning. The first task was to clean the data. Analyzing the features, the platform and machine learning methods were selected. After the methods were applied, the machine learning methods were compared through standard statistical evaluation tools.

Analyzing the features, the variables and machine learning methods were selected. Per our hypothesis, the independent variables selected for the X values were “medcond\_yn” and “age\_group”. The dependent Y variables analysed were first “hosp\_yn” and later “death\_yn”. Training/Testing split was 70/30.

After the methods were applied, the machine learning methods were compared through standard statistical evaluation tools. .

## **Dataset with Features**

The dataset is a public dataset, “Covid-19 Case Surveillance Public Use Dataset” provided by www.kaggle.com of compiled data from US cases reported to the CDC. Features are listed below in Table 1[3].

**Table1: Dataset Features**

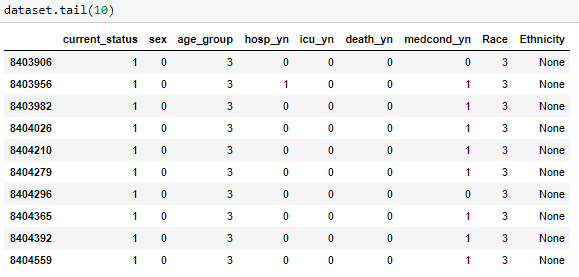
|  |  |  |
| --- | --- | --- |
| Features | Type | Values |
| cdc\_report\_dt | date | multiple dates |
| pos\_spec-dt | date | multiple dates |
| onset\_dt | date | multiple dates |
| current-status | string | ['Laboratory-confirmed case' 'Probable Case'] |
| sex | string | Male, Female |
| age\_group | string | '10 - 19 Years','20 - 29 Years','30 - 39 Years','40 - 49 Years', '50 - 59 Years', '0 - 9 Years','60 - 69 Years', '80+ Years' ,'70 - 79 Years'  'Unknown' |
| race and ethnicity | string | ['Black',White' 'Missing','Hispanic/Latino' 'Unknown','Asian', 'Multiple/Other','American Indian/Alaska Native', 'Native Hawaiian/Other Pacific Islander'] |
| hosp\_yn | string | ['No','Missing','Unknown','Yes'] |
| icu\_yn | string | 'No','Missing','Unknown','Yes'] |
| death\_yn | string | 'No','Missing','Unknown','Yes'] |
| medcond\_yn | string | 'No','Missing','Unknown','Yes'] |

## **Platform**

Our analysis was performed using Jupyter Notebook, an open-source python interpreter that helped transform the dataset by cleaning and implemented machine learning techniques with statistical analysis.

## **Cleaning of the dataset for use**

Out of the 11 variable columns, we dropped the first 3 pertaining to dates as well as the “laboratory confirmed” as irrelevant to our research. For simplicity we also split the “Race” and “Ethnicity” columns and dropped the original combined column. The dataset contained a lot of “missing” “unknown” values which we hypothesized would produce poorer results, so we dropped all entries with those values present in any of the columns. Given that the dataset is large we saw no need to generalize values in place of the missing data. After removing all “missing” or “unknown” values, our dataset retained over 400,000 of the original 8 million entries - a sizable dataset. Finally, for the purposes of machine learning algorithm analysis, we replaced all text values with numerical values. Figure 3 portrays the look of our resulting data set.

Figure 3 - Dataset after cleaning

## **Machine Learning Methods Applied**

1. Logistic Regression: Logistic regression is based on the Sigmoid function and is a great classifier for binary Logistic regression is not good on datasets that have many features and small training samples. The COVID19 dataset studied here has limited features and is very large [14].
2. Naïve Bayes: A commonly used classifier based on the Bayes Theorem is called Naive Bayes. This a simple and effective classifier that calculates the probability of the results. In the Bayes Theorem, all features are considered independent of one another. The Gaussian Naive Bayes model uses the average value and the standard deviation of each feature for the class. Naive Bayes are fast and robust. In addition, they perform well on large datasets such as this COVID-19 dataset analyzed in this paper [13, p.70-72].
3. K-Nearest Neighbors: Considered the most simple machine learning language, K-nearest neighbors predicts the data points based on the closest data points in the training set. N is the value of how many neighbors will be used by the algorithm. Being able to change the N value allows for fine tuning. When N is 1, it is the simplest nearest neighbor algorithm [13, p. 36-41].
4. Decision Tree: Widely used for regression and classification, decision tree algorithms learn from a series of if/else questions in a hierarchical pattern [13, p.72]. They are powerful, fast and robust without the need to scale the data [13, p.130]

## **Evaluations of Data**

In order, to evaluate the machine learning results measurements for analysis were chosen.

Confusion Matrix: A confusion matrix is a visual display of the positive or negative classes represented on the rows and the predicted negative or positive classes represented on the columns. TN represents true negatives. FN represents false negatives. FP represents false positives. TP represents true positives. The confusion matrix is a two-by-two matrix in the following format:

|  |  |  |
| --- | --- | --- |
| Actual  Negative | TN | FP |
| Actual  Positive | FN | TP |
|  | Predicted  Negative | Predicted  Positive |

Figure 4 - Confusion Matrix

1. Accuracy: Accuracy is the correct predictions true positive or true negative, added together and divided by the sum of the number of all entries [13, p.289].
2. Precision: Precision is used to measure the true positive to all the positive results both true and false. It is a metric that is used to limit the number of false positives [13, p.289].

1. Recall: Recall is used to identify all positive samples. It is calculated by dividing the true positives by the true positive and false negative results. The goal of a recall metric is to avoid false negatives [13, p.289].

1. F1 score: The F1 score uses both the precision and recall. It gives a measure of the accuracy of the two together [13, p.290]

# **results**

In our analysis, cross-validation of the accuracy was used. Cross Validation takes random splits of the training data and validation data. This is a defined number in the algorithm, and the number used in this analysis was 10. Therefore, there are 10 different performances tested. The model is trained and then measured to the validation data. An average of the cross validations on accuracy were calculated [15].

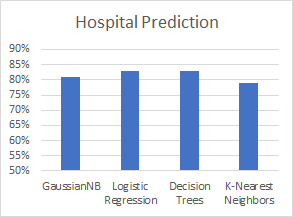
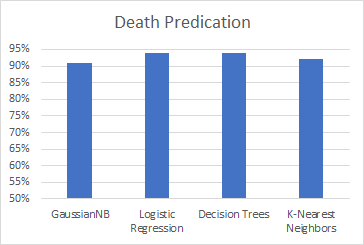
For overall accuracy, logistic regression and decision tree algorithms gave the best performance. For predicting hospitalization, logistic regression and decision tree algorithms were 83% accuracy. Whereas the same algorithms gave a 94% when predicting deaths. 

Figure 5-Accuracy for Hospitalization



**Figure 6-Accuracy for Deaths**

Precision gives the number of correctly predicted positive results [13, p.289]. In our analysis Gausian Naive Bayes gave the best prediction with 82% precision when predicting hospitalization.

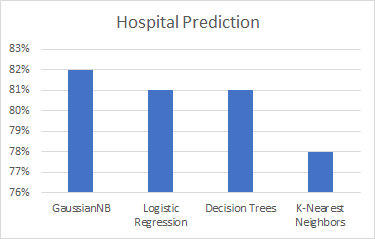
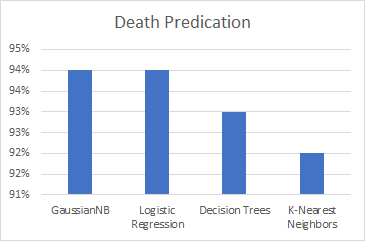


Figure 7-Precision for Hospitalization

For predicting deaths, Gausian Naive Bayes and logistic regression showed a precision of 94%.

Figure 8-Precision for Deaths

Recall is the number of true positives compared to all positive results [13, p.289]. Again, logistic regression and decision tree algorithms performed the highest recall at 83% for predicting hospitalization and 94% for predicting deaths.

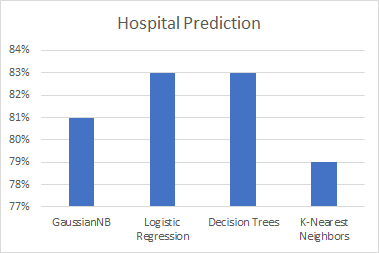


Figure 9-Recall for Hospitalization

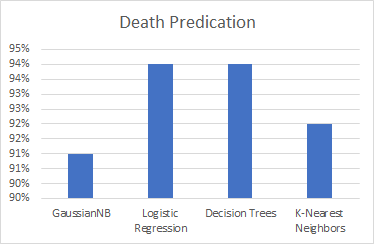


Figure 10-Recall for Deaths

F-measure shows the combination of precision and recall performance [13, p.290]. for predicting hospitalization, Gausian Naive Bayes, logistic regression and decision trees algorithms all performed a F-1 score of 81%. While for predicting deaths, the logistic regression, decision trees, and K-nearest neighbors performed the best F-1 score of 93%.

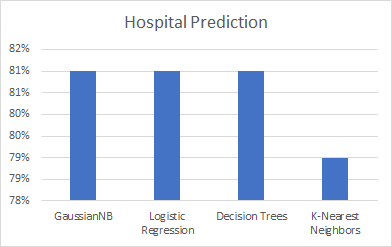


Figure 11-F1 for Hospitalization

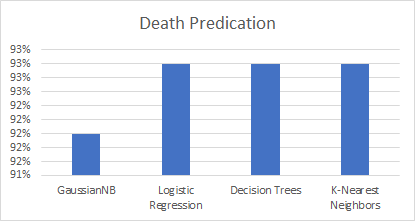


Figure 12-F1 for Deaths

# **Future research**

Possible future research proposals would look more closely at the types of preexisting conditions that are most likely to result in Covid deaths, provided such data exists. In addition, co-occurrence of factors such as old age and preexisting conditions with different race or sex demographics could be examined to determine if some are affected more than others.

# **Conclusion**

Our machine learning code positively supports our hypothesis of a direct relationship between the independent variables of preexisting medical conditions and age and the dependent variables of Covid hospitalization or death. The cross-validation score mean of .80 affirmed this hypothesis. For predicting hospitalization, averages for Logistic Regression, and GaussianNB algorithms were all at 83%.. Also, for predicting deaths, Logistic Regression, and GaussianNB algorithms, the weighted average yielded 93%. This suggests the Logistic Regression and GaussianNB algorithms were successful predicting deaths and hospitalizations the best out of all algorithms used.

Using virtually all algorithms and by all metrics, death predictions were more accurate than hospitalization predictions. This could be attributed to the fact that death is simply easier to predict using age and preexisting condition variables, whereas hospitalizations were more widespread among the population overall, making them harder to predict. We could not find significant differences between different metrics results, suggesting they are all equally valid for our research purposes.

##### **Acknowledgment**

Thank you to our patient and considerate professor Doctor Seyadamin Pouriyeh and our very wonderful and very responsive advisor Doctor Li Lei

##### **References**

1. World Health Organization. (‎2020)‎. Origin of SARS-CoV-2, 26 March 2020. World Health Organization. https://apps.who.int/iris/handle/10665/332197. License: CC BY-NC-SA 3.0 IGO
2. Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1–30, 2020. MMWR Morb Mortal Wkly Rep 2020;69:458–464. DOI: http://dx.doi.org/10.15585/mmwr.mm6915e3external icon
3. Möbius. “Covid-19 Case Surveillance Public Use Dataset.” Kaggle, 21 Dec. 2020, www.kaggle.com/arashnic/covid19-case-surveillance-public-use-dataset.
4. A timeline of COVID-19 developments in 2020. (2021, January 1). Retrieved April 14, 2021, from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020
5. Coronavirus (covid-19) deaths - statistics and research. (2021, April 13). Retrieved April 14, 2021, from <https://ourworldindata.org/covid-deaths>
6. Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020, March 05). Coronavirus pandemic (covid-19) - statistics and research. Retrieved April 16, 2021, from https://ourworldindata.org/coronavirus
7. J. Zhang, X. Wang, et al. Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China, Clinical Microbiology and Infection, Volume 26, Issue 6,

2020, Pages 767-772, ISSN 1198-743X, https://doi.org/10.1016/j.cmi.2020.04.012.

1. Mikami, T., Miyashita, H., Yamada, T. et al. Risk Factors for Mortality in Patients with COVID-19 in New York City. J GEN INTERN MED 36, 17–26 (2021). https://doi.org/10.1007/s11606-020-05983-z
2. Uma V Mahajan, Margaret Larkins-Pettigrew, Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties, Journal of Public Health, Volume 42, Issue 3, September 2020, Pages 445–447, https://doi.org/10.1093/pubmed/fdaa070
3. Harsh Bhatnagar. COVID-19 Case Surveillance. *Kaggle*, 21 Dec. 2020, https://www.kaggle.com/harshbhatnagar/covid-19-case-surveillance
4. Athar, A. (2021, January 06). Covid-19: Case surveillance + modelling. Retrieved April 17, 2021, from https://www.kaggle.com/afafathar3007/covid-19-case-surveillance-modelling
5. Ray, S. (2020, December 23). Commonly used machine learning algorithms: Data science. Retrieved April 17, 2021, from https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
6. Guido, Sarah. *Introduction to Machine Learning with Python*. O'Reilly Media, 2016.
7. Jun Liu, Jianhui Chen, and Jieping Ye. 2009. Large-scale sparse logistic regression. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). Association for Computing Machinery, New York, NY, USA, 547–556. DOI:https://doi.org/10.1145/1557019.1557082
8. George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explor. Newsl. 12, 1 (June 2010), 49–57. DOI:https://doi.org/10.1145/1882471.1882479