**Abstract:**

We analyze two datasets regarding Covid-19 confirmed cases for race, ethnicity, death, and ages for the state of California. The two datasets contain information regarding deaths caused by Covid-19 and which race/ethnicity and age group was the most and least affected. For the predictive model using Microsoft Azure Machine Learning we focused on two columns from the datasets. The two columns contained data regarding the confirmed total cases and the death total. We also wanted to include the recovery data, but that data was not available. We tested for the accuracy of the data we trained using linear regression and Bayesian linear regression. Furthermore, we checked for the permutation level of important and cross validated the two data sets to determine the level or errors and accuracy. Using Kibana we were able to display the data. The visuals we created helped us as a team to elaborate which data points seemed to be the most important. Whatever we were able to derive from the graphs was represented in the Azure ML tests.

**Introduction:**

We decided to do an analysis on the Covid-19 confirmed cases total and death total in the United States, specifically focusing on the state of California. We decided to work with two datasets from the Los Angeles Times’ tally of coronavirus cases in California obtained from GitHub. The two datasets contain information about the confirmed cases and deaths and which race/ ethnicity and age group was the most and least affected by Covid-19. We believe that it is important to know the impact Covid-19 has had on our state. We will be analyzing the total confirmed cases and death totals. We can determine which age group and race/ethnicity was the most and least affected. With the visualizations of the data we were able to display what the main focus of using this data was. We wanted to show in a clear and vivid matter how this data affected the California community.

**Related work:**

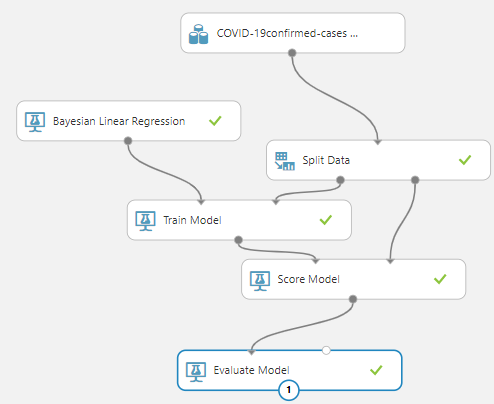
We were not able to find any journal articles/papers with similar work as us. There were only news articles that had similar information to our project. In a Los Angeles Times article titled, tracking coronavirus vaccinations in California, it provides statistics of doses administered, partial doses administered, and demographic information regarding vaccination. As well as confirmed cases and deaths. The work that they did is like ours only that their work is being constantly updated. Their data looks different from our due to their data being constantly updated. Another article that contains similar work is from ABC 7 news titled, California Covid-19 vaccine tracker:  How the state is doing, when you can get a coronavirus vaccine. This article also contains information regarding the confirmed cases and deaths in regard to demographics. The vaccination information is more generalized and does not include specific information.

**Background/existing work:**

The data that we used the current Covid-19 vaccination data as of May 18, 2021 for the state of California. It contains many datasets regarding how Covid-19 impacted various organization, demographics ages etc. We only focused on the dataset for “confirmed cases and death for race and ethnicity” and the “Ages” dataset. These files come from a continual Times survey of California's 58 county health agencies and three city agencies. The figures are typically ahead of the totals compiled by California's Department of Public Health. By polling local agencies, The Times database also gathers some information not provided by the state. The tallies here are mostly limited to residents of California, the standard method used to count patients by the state’s health authorities. Those totals do not include people from other states who are quarantined in California, such as the passengers and crew of the Grand Princess cruise ship that docked in Oakland.

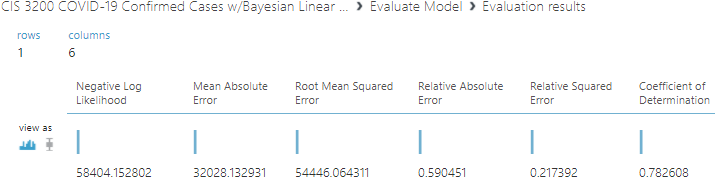
**Your work:**

After importing the two datasets into azure machine learning. We needed to test the level of errors and accuracy for the data column that we wanted to train. We created a simple model and trained two types of data. We trained the confirmed total cases of people found testing positive for COVID-19 and death total. We used the Bayesian linear module and linear regression module to test the accuracy and level of error of our two data sets as shown below. (Replace the Bayesian Linear Regression module with the Linear Regression module to test both types of modules)

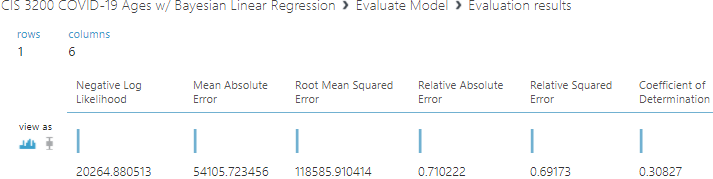


After running the model, we visualized both data sets for Bayesian and linear regression and found that the level of errors and accuracy were the same so we will only include the Bayesian linear regression visualization as shown below.

Confirmed Cases Total Dataset:



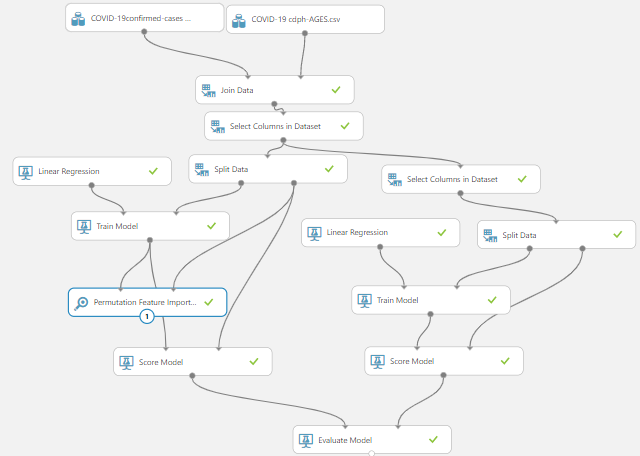
Ages Data set:



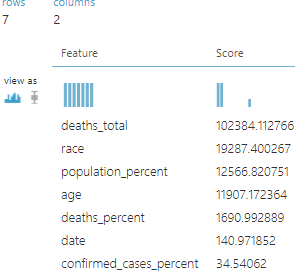
The coefficient of determination for the confirmed cases dataset is a 0.78 and 0.30 for the age’s dataset. The closer it is to one the more accurate the data is. Meaning that the confirmed cases dataset level of accuracy is not bad while the level of accuracy for the age’s dataset is bad and filled with errors.

We decided to create a new experiment and join both datasets in the hopes of building a better model. We used linear regression (left side) to test the level or accuracy and error level. Furthermore, we used the permutation level of importance module to help us find out what data columns are the least important. We found out the confirmed cases percent column for both datasets is the column of least importance. Hence, we decided to remove the confirmed cases percent column. We created another model next to the existing one to attempt to cross validate the data to improve accuracy and decrease the level of errors. In the new model we excluded the confirmed cases percent column with the project columns module. We placed a score module on both models and evaluate the models and ran the experiment.

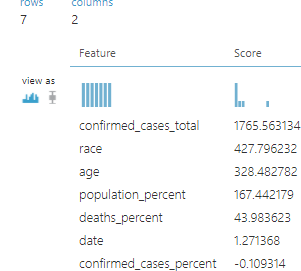
Final Experiment Model:



Permutation level of importance For Confirmed Cases:



Permutation level of importance For Death Total:

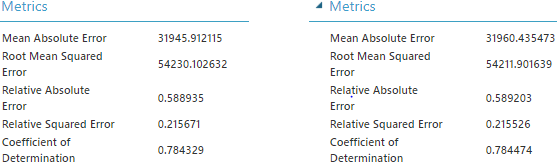


Results:

After many attempts we concluded that the only column we were able to remove to get a more accurate model is the confirmed cases percent column. When we also excluded the date column for both confirmed total cases and death total when training the models, the level of accuracy deceased. The accuracy slightly increased when we only removed the confirmed cases percent column. The comparison of both models for confirmed cases total and death totals is shown below.

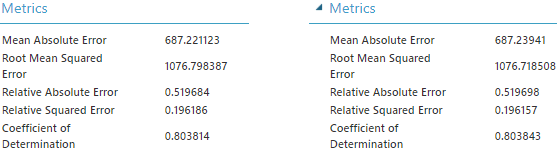
Confirmed Cases Total Data:

With Every Column | Confirmed Death Percent Column Excluded



Death Total Data:

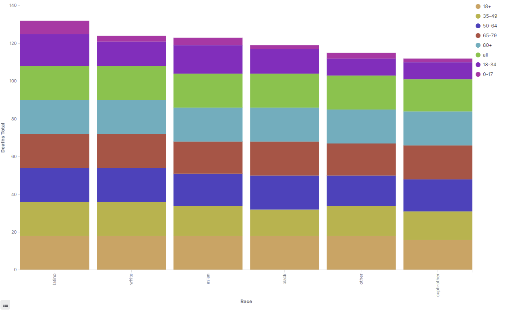
With Every Column | Confirmed Death Percent Column Excluded



The new models on the right made the results more accurate, but the difference is very insignificant.

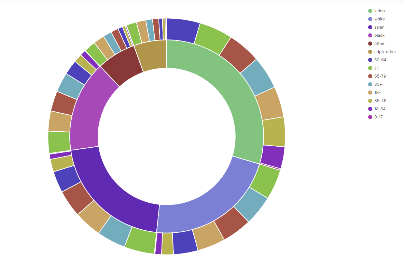
Kibana:

Through the use of elastic search and Kibana I was able to present the data in a clear format to understand and visualize it. The graphs were easy to produce as it was simple to use the Kibana platform. The only trouble I encountered came when using both covid vaccination data sets as the demographic value column contained both the age and race of people who received the vaccination. Unlike the other dataset which we included to see the death totals in California by the demographics. The death totals were of note because of how similar the amount of death totals were when put in a graph as well as scary. But overall, the data proved to be informative and interesting.

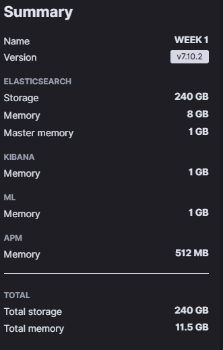




Here I have a vertical graph displaying the total deaths related to covid by age. I also have the age groups of the races shown in each bar separated by color. From what me and my partner could tell was how the deaths were spread out. What is noticed first is which demographic groups were affected the most by covid in California this year. The Latino demographic suffered the most deaths as they have the most total numbers in death. Then the least total number of deaths were from the black demographic. To get the percentage of the demographics death’s, I utilized a Pie Chart.



The percentage of the total deaths were fairly even but we can get distinct percentages. What we already know from the vertical graph is who has the most deaths and who has the least. With the pie graph me and my partner can attain exact percentages of the total deaths shared between demographics and age groups. As I said before the Latino demographic resulted in the most total deaths due to covid from what we could tell from the graph but with the help of the pie graph we determined that 29.62% of the covid related deaths came from the Latino population in California. The percentages tell us which communities have been adversely affected by the Covid virus the most. Not only does the pie chart split the total deaths by percentages, it can also represent the ages of the people who died from covid in each race demographic. We can see clearly in the pie chart which age group was affected the most in each race group. In the Latino demographic the age group affected the most by COVID is people aged 50-64. What is most interesting is how this is the only demographic which has the people of this age to have suffered the most deaths. Most of the other demographics have this age group to have suffered the least amount of losses. It is apparent how people who are older are more susceptible to the COVID disease because most of the demographics of their older generations have higher total deaths. What this confirms is the fact that older people (which we consider to be 50 and up) have a higher mortality rate when fighting covid. Having this data allows us to recognize the need to be aware of how deadly covid is towards older people. That is why families in California could not visit their older relatives for a time.



**Conclusion:**

In conclusion, when using azure machine learning we were able to model, train, test and evaluate the data we selected. Furthermore, we were able to test the accuracy of the model and determine which data were needed and not needed by finding the permutation level of importance and cross validating. The data was not entirely accurate and had many errors. We were able to exclude the Confirmed cases percent column to slightly increase our model’s accuracy when training the confirmed total cases and death total. Based on our analysis we determined that the number of cases and deaths have been decreasing significantly. This may be due people being vaccinated or also due to people building up immunity. Afterall, there is no dataset that includes information on their recovery. Many people did not even know they contracted the virus and recovered without being tested so their case was never known. When using elastic, we have come to several conclusions about the data and where the majority of the deaths were occurring. We were able to make profound conclusions in our topic.

**References:** Papers, articles, URLs (your github, data source,, Azure Studio link, …) that you referred to

LINK FOR AZURE ML:

<https://studio.azureml.net/Home/ViewWorkspaceCached/e839ba1d4cd4455e9af1e8610742b5e4?#Workspace/Experiments/ListExperiments>

Data Source:

<https://data.chhs.ca.gov/dataset/vaccine-progress-dashboard>

Articles of Similar work:

<https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/covid-19-vaccines-distribution/>

<https://abc7news.com/covid-vaccine-california-tracker-ca-newsom-update/9567632/>

Elastic Link:

<https://4f3b2b9038c04b85a6a137abfde9b509.us-central1.gcp.cloud.es.io:9243/app/dashboards#/view/7f31cf40-b2a9-11eb-9680-e7ae60df816f?_g=(filters:!(),refreshInterval:(pause:!f,value:30000),time:(from:'2021-05-01T07:00:00.000Z',to:'2021-05-09T06:30:00.000Z'))&_a=(description:'',filters:!(),fullScreenMode:!f,options:(hidePanelTitles:!f,useMargins:!t),query:(language:kuery,query:''),timeRestore:!f,title:'Vaccinations%20in%20California',viewMode:view)>

GitHub Link:

<https://github.com/XALFONZO90X/CIS-3200-PROJECT>