Data 200 Graduate Project 2023

Checkpoint 3 Write - Up

Topic 1 - COVID 19

Lakshya Aggarwal
Apratim Banerjee
Anusri Sreenath

November 27, 2023

1 Abstract

This paper presents a analysis of COVID-19 data, focusing particularly on the United States. The primary datasets utilized include the 'time series covid19 confirmed US' and 'time series covid19 deaths US' from the CSSE at Johns Hopkins University and CDC, alongside 'Provisional COVID-19 Death by Sex and Age' from the National Center for Health Statistics (NCHS). Our research aims to understand the dynamics of COVID-19 spread and impact through exploratory data analysis (EDA), feature engineering, and the development of predictive models using advanced statistical and machine learning techniques.

The methodology involves data preparation, feature engineering to extract temporal and demographic dimensions, and the application of various machine learning models, including linear regression and decision trees, for both predictive and classification tasks. Specifically, the paper explores time series analysis to forecast COVID-19 cases, employing models that consider the cumulative daily cases and demographic characteristics. The study pays particular attention to the states of New York and California, due to their significant data volume and distinctive COVID-19 impact patterns.

The findings reveal intriguing geographical nuances in the pandemic's impact, with distinct trends in confirmed cases and deaths in the studied states. The paper evaluates the models' performance using metrics like Root Mean Squared Error (RMSE), providing insights into their accuracy. The study also identifies challenges such as the simplification inherent in linear models and suggests the potential of more sophisticated machine learning models for future research.

The research questions guide our entire analysis. Firstly, we investigate how to accurately forecast the future number of COVID-19 cases using time series analysis of historical data. This forecasting aims to equip healthcare systems and policymakers with predictive insights, facilitating proactive planning and resource allocation to prevent the potential overburdening of medical infrastructure. Secondly, we aim to determine the age groups most susceptible to COVID-19's impact by analyzing cases and fatalities. This information is crucial for tailoring health communications and interventions, specifically targeting vulnerable populations

2 Introduction

In the realm of public health, the exploration and analysis of COVID-19 data have emerged as crucial components for understanding the dynamics of the pandemic. Conducting exploratory data analysis (EDA) on COVID-19 datasets provides invaluable insights into the spread, impact, and patterns of the virus across regions and populations. EDA helps identify trends, hotspots, and potential influencing factors, enabling

policymakers and healthcare professionals to make informed decisions. Furthermore, the development of prediction models becomes essential for forecasting the future trajectory of the pandemic, anticipating potential surges, and optimizing resource allocation. By leveraging advanced statistical and machine learning techniques, these models contribute to proactive planning, allowing for the implementation of targeted interventions and mitigation strategies

Our research is mainly focused on the cases and data of The United States. 3 Data set files under "csse_covid_19_time_series," were considered. We collected data from 'time_series_covid19_confirmed_US' and 'time_series_covid19_deaths_US', which mainly give us information regarding the number of deaths and the number of cases recorded every day in various states. This data set contains US reports on COVID-19 testing and cases from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and CDC (Centers for Disease Control and Prevention). The second dataset, which we will be using for this analysis is "Provisional COVID-19 Death by Sex and Age," provided by the National Center for Health Statistics (NCHS). Some potential reasons that may cause this data to be inaccurate can be Misclassification of Cause of Death: Determining the cause of death can be complex, and misclassification can occur. Some COVID-19 deaths may be misclassified as other causes, and vice versa. In some cases, data may be suppressed or not reported for privacy or security reasons. This can lead to incomplete datasets.

After thorough literature review, it was found that among all the proposed methods of machine learning to solve problems of binary classification, the algorithm of decision-making trees best coped [1] and we wanted to use this algorithm for our use case. We also performed logistic regression, which is is a well-known statistical method for determining several factors' influence on a logical pair of results. The name "logistic regression" reflects the fact that the data curve is compressed by applying a logistic transformation to reduce the effect of extreme values. [1] We want to employ this method, as well as run tests on multiple other algorithms to define the best fitting model.

To predict the number of Covid 19 cases in the upcoming months, time series analysis would best suit our use case.Time series analysis keeps time interval in specific period into consideration while deal with data.

Stipulated Time interval generates behavior and pattern in data series which help to design a forecasting model [2]. Hence the following paper intends to find reason for trends, patterns, and compare different models for the target data set.

By employing an approach that integrates state-level testing and case data with demographic information on COVID-19-related deaths, our research aims to provide a holistic understanding of the pandemic's impact on different age groups, contributing novel perspectives to inform public health strategies.

3 Description of data

In our study of the impact of the COVID-19 pandemic, we utilized two crucial data sets:'time_series_covid19 _confirmed_US' and 'time_series_covid19_deaths_US' from the "csse_covid _19_time_series" collection, capturing details on COVID-19 testing, cases, and deaths in the United States. Complementing this primary data source is the "Provisional COVID-19 Deaths by Sex and Age" data set from the National Center for Health Statistics (NCHS), offering valuable insights into COVID-19-related deaths categorized by sex and age.Columns such as "COVID-19 Deaths," "Total Deaths," "Pneumonia Deaths," "Pneumonia and COVID-19 Deaths," "Influenza Deaths," "Pneumonia," and "Influenza, or COVID-19 Deaths" contain numerical data, likely integers, representing the counts of deaths attributed to specific categories.

The "State" column is a text column indicating the name of the state or territory.

The "Sex" column is a text column indicating gender, with possible values like "Male" and "Female."

The "Age Group" column is a text column denoting age categories, with age ranges such as "0-17 years," "1-4 years," "5-14 years," "15-24 years," "18-29 years," etc.

Data preparation commenced with a cleaning process to address the granularity inherent in the CSV files,

which reported data at the provincial level within each state. The scope of the data extended from January 22nd, 2020, to March 28th, 2022. We modulated the dataset into 60 days per point, and grouped the data by Province State. We cleaned the data set by removing columns that had stray data and was affecting the trend. We also aggregated the data to find out the total confirmed cases and total confirmed deaths for the top 10 states in the USA.

We processed the "Provincial COVID-19 deaths by Sex and Age" dataset by filtering the rows to only include monthly cases and encoding categorical data, such as age group and state columns. Our objective was to predict the category label Y, which represents the age group for each entry/row in the dataset. This was a categorical prediction task, where we used the input features X (Year, Month, Covid-19 Deaths, Pneumonia Deaths, Influenza Deaths, State - encoded, and Total Deaths) to train a model and predict the corresponding age group for each row.

In summary, our approach involved data cleaning and reduction, enabling analysis and exploration. Feature engineering efforts enriched the data set, incorporating temporal and demographic dimensions for predictive modeling. This preparation helped build a foundation for our study's goal of providing insights into the impact of COVID-19 in the United States by analyzing whom it impacted and predicting the virus trend.

4 Methodology

Feature engineering played a pivotal role in enhancing the dataset for predictive modeling. Time series components were isolated, emphasizing trends, seasonality, and residuals from cumulative daily cases, with a specific focus on California for prediction purposes. Demographic features were also engineered, categorizing and encoding age to enrich predictions related to age groups affected by COVID-19, pneumonia, or influenza.

The first study aims to forecast the number of COVID-19 cases for the upcoming month in the United States by employing time series analysis of historical data. The objective is to provide healthcare systems and policymakers with predictive insights to prepare and mitigate the potential overburdening of medical infrastructure.

The dataset comprised daily confirmed COVID-19 cases and deaths for all U.S. states over the last two years. The dataset was preprocessed by grouping the data by state. Preliminary analysis identified New York and California as states with the highest number of deaths and confirmed cases, respectively. These states were selected for in-depth analysis due to their significant data volume, which is likely to yield more reliable insights.

The primary method for forecasting COVID-19 cases is time series analysis. Time series analysis is particularly suitable for this research due to its effectiveness in analyzing data points collected or recorded at regular time intervals. This approach helps in identifying patterns, trends, and seasonal variations in the data, which are crucial for accurate forecasting.

Model Selection and Development
Linear regression models were developed for New York and California. The choice of linear regression is grounded in its simplicity and effectiveness in capturing trends in time series data. The models predict the number of COVID-19 cases based on historical data trends.

We employed various models for the Categorical Prediction task, including linear regression to predict the age group and decision trees via Random Forest. The latter showed the best results among the models tested, specifically for the 17-class classification task of all age group labels within the dataset!

Feature Engineering
Feature engineering involved extracting meaningful features from the time series data. This included creating time-based features like month, week, and day to capture seasonal and cyclical trends. Additional features such

as moving averages were also computed to smooth out short-term fluctuations and highlight longer-term trends.

We used label encoding and filtered the data for the classification task of 17 classes for the "Provincial COVID-19 deaths by Sex and Age" dataset.

Regularization
Given the complexity of the dataset, we applied lasso and ridge regularization techniques to prevent overfitting for the time series task and reduce the potential for highly variable COVID-19 case numbers.

We experimented with various hyperparameters for the classification task using different models. After comparing the results, we found that the decision tree classifier performed the best. We then focused on adjusting the n_estimators parameter, which determines the number of decision trees in the random forest ensemble. Additionally, we set the random_state to 42, which sets the seed for the random number generator and ensures that the results are reproducible when the algorithm involves random processes.

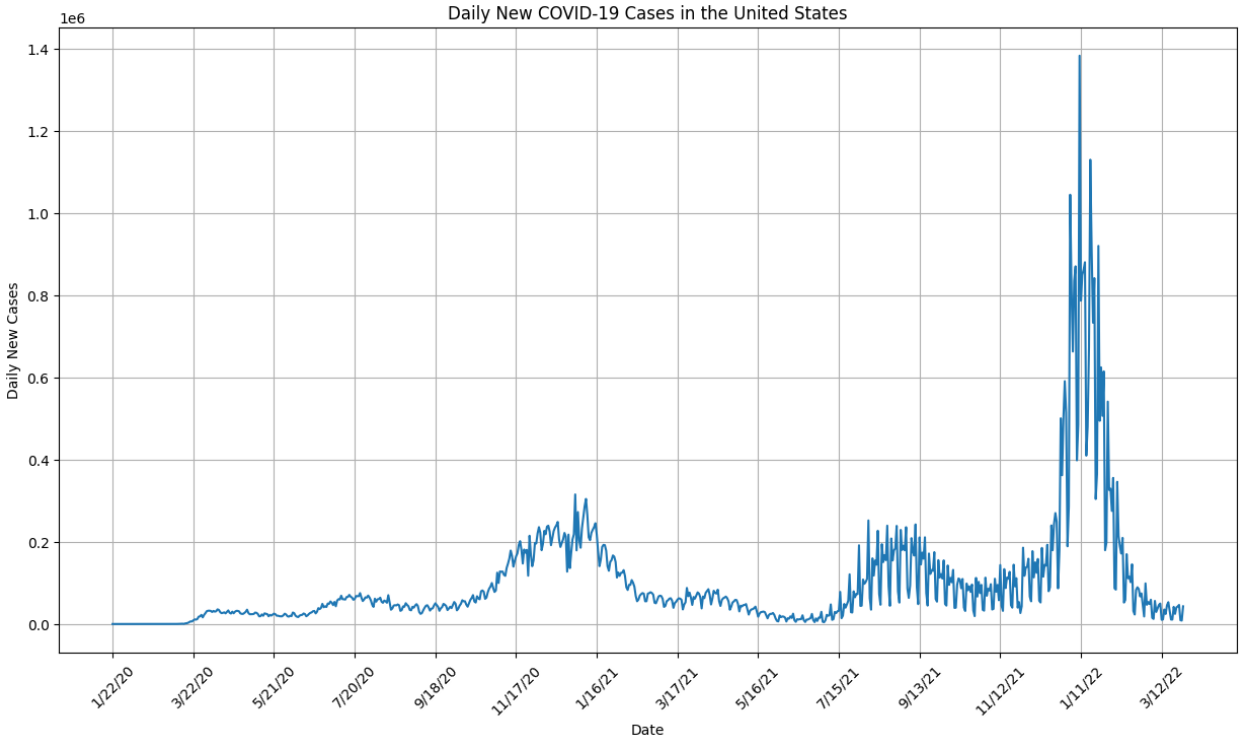
Inference and Prediction Methods
The models were used to make inferences about the trends in COVID-19 cases and to forecast future cases. The forecasting was conducted for a one-month horizon, aligning with the research objective to provide short-term predictive insights.

Model Evaluation and Validation
To ensure the model's robustness, cross-validation was performed. This involved dividing the datasets into training and testing sets to validate the model's performance on unseen data. The model was trained on a subset of the data and tested on another subset to evaluate its predictive accuracy.

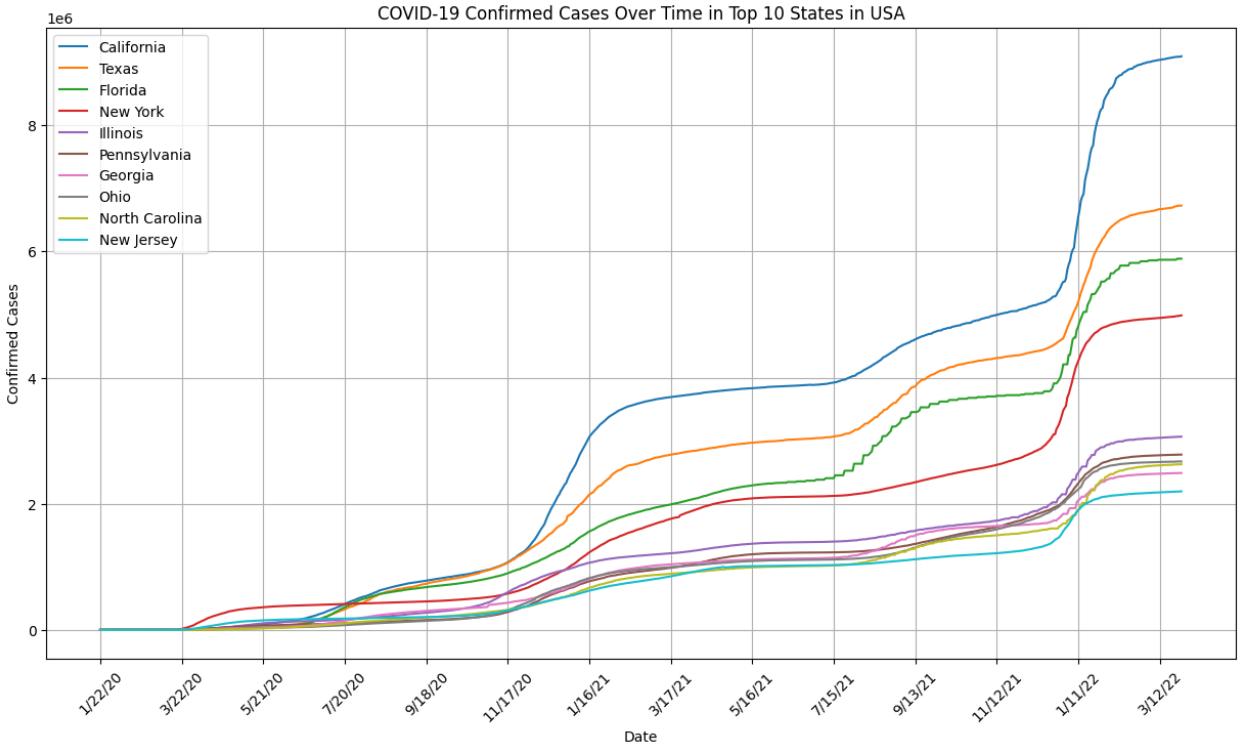
Model Selection and Evaluation Metrics
The model's performance was evaluated using Root Mean Squared Error (RMSE). These metrics provide insights into the model's accuracy and its ability to generalize to new data.

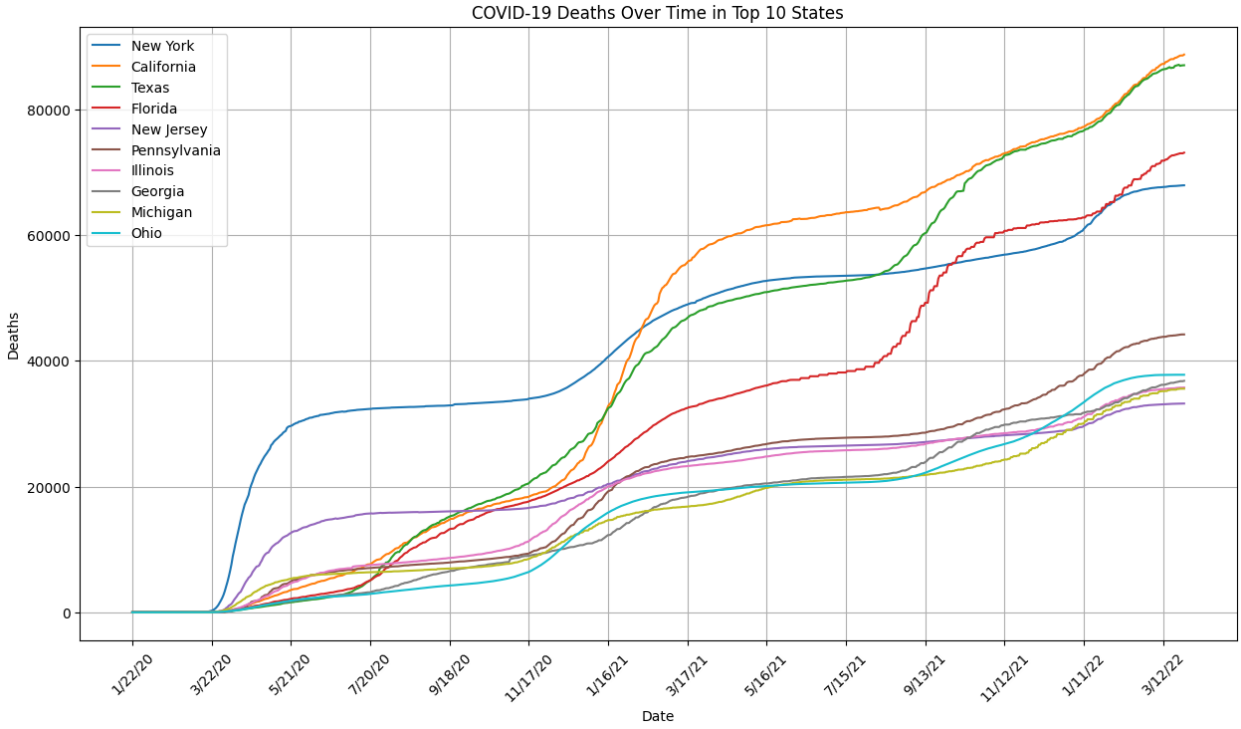
5 Summary of results

We first plotted the 'Daily New COVID-19 Cases in the United States' graph which accurately shows the number of different waves in which the pandemic hit the United States of America within the two year period.



Then we plotted the top-10-states with Confirmed COVID-19 Cases and Confirmed COVID-19 Deaths.

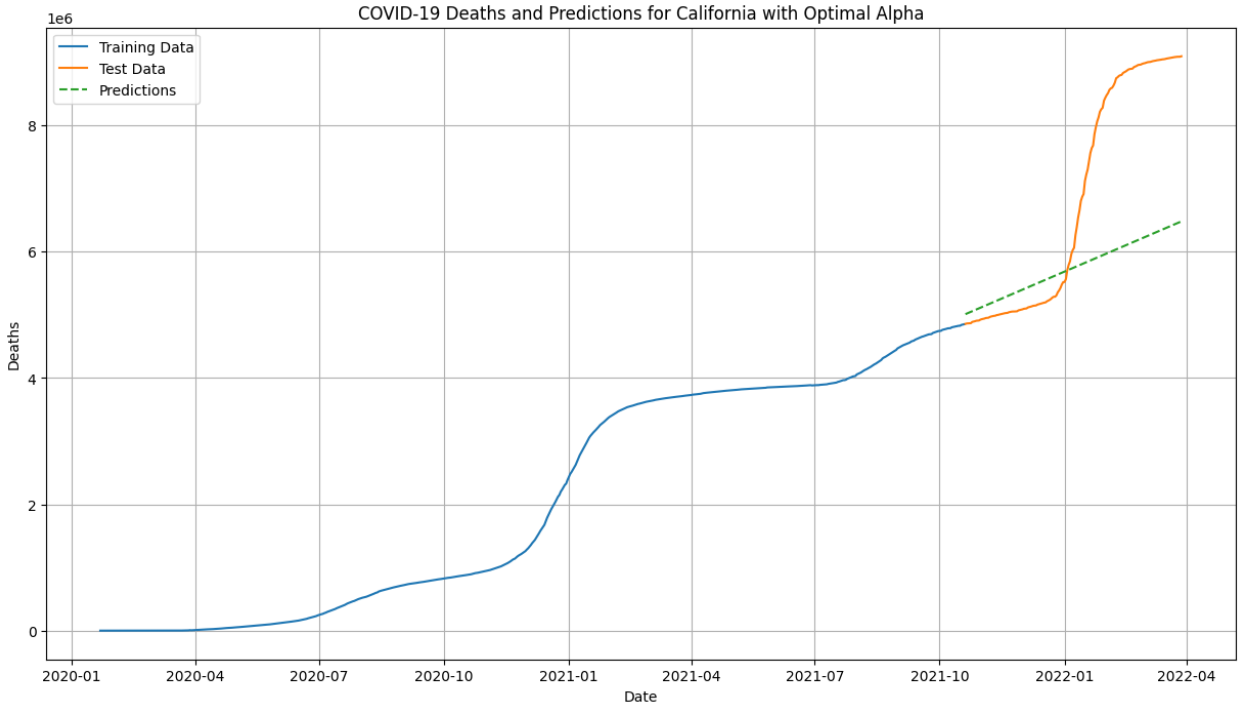


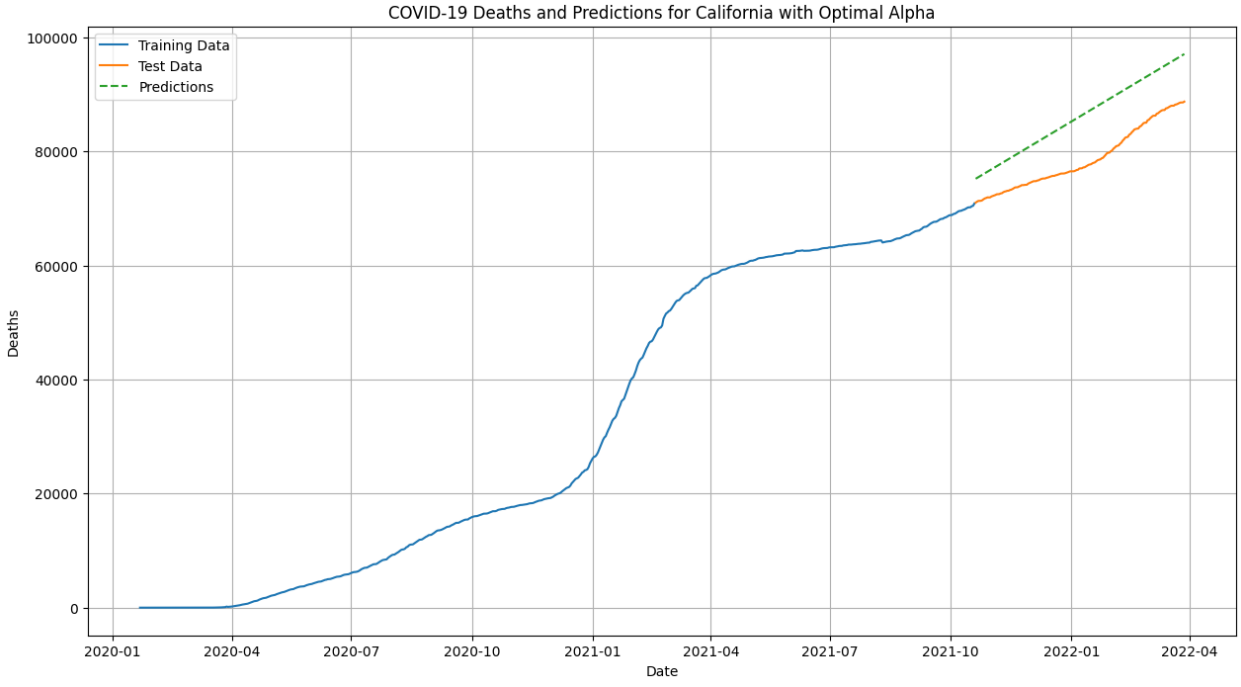


Notably, There were several observations for both the graphs:
Significant Variability: The states exhibit significant variability and the rate at which these cases and deaths occurred over time.
Major Peaks: The plot highlights major peaks in confirmed cases and deaths, which likely correspond to different waves of the pandemic.
New York and California: These states, in particular, show a higher burden of deaths compared to other states in the top 10.

California emerged with the highest number of COVID-19 cases, prompting further investigation. Intriguingly, while California had the most cases, New York recorded the maximum number of deaths, with California a close second. Upon further research we realized that New York's maximum deaths were recorded during the first wave of the pandemic, since it reached first in the United States and the states were least prepared. New York also had a slightly higher proportion of senior citizens and had colder weather, which kept people indoors and transmission risks became higher.

Next, Let us perform some time-series modelling and analysis for California, using ridge regression. The plot is as follows :





From the two prediction graphs:
COVID-19 Confirmed Cases and Predictions for California with Optimal Alpha (**RMSE: 1736812.14**)
COVID-19 Deaths and Predictions for California with Optimal Alpha (**RMSE: 7649.45**)

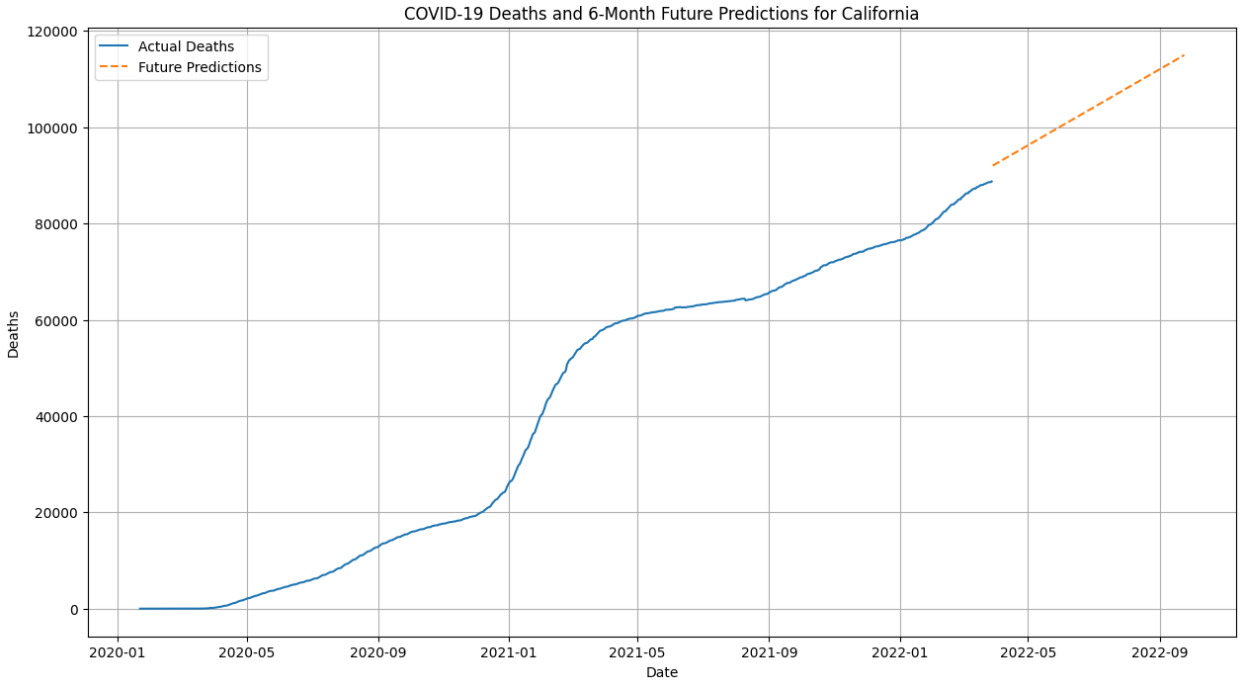
Linear regression models for California demonstrated a moderate degree of accuracy in predicting COVID-19 case numbers.

We can conclude that the time series prediction worked better for the data-set of confirmed deaths. Further research shows that during the last six months of both the data-sets, there was a sharp increase in the number of confirmed cases in California versus a very steady rise of confirmed deaths.

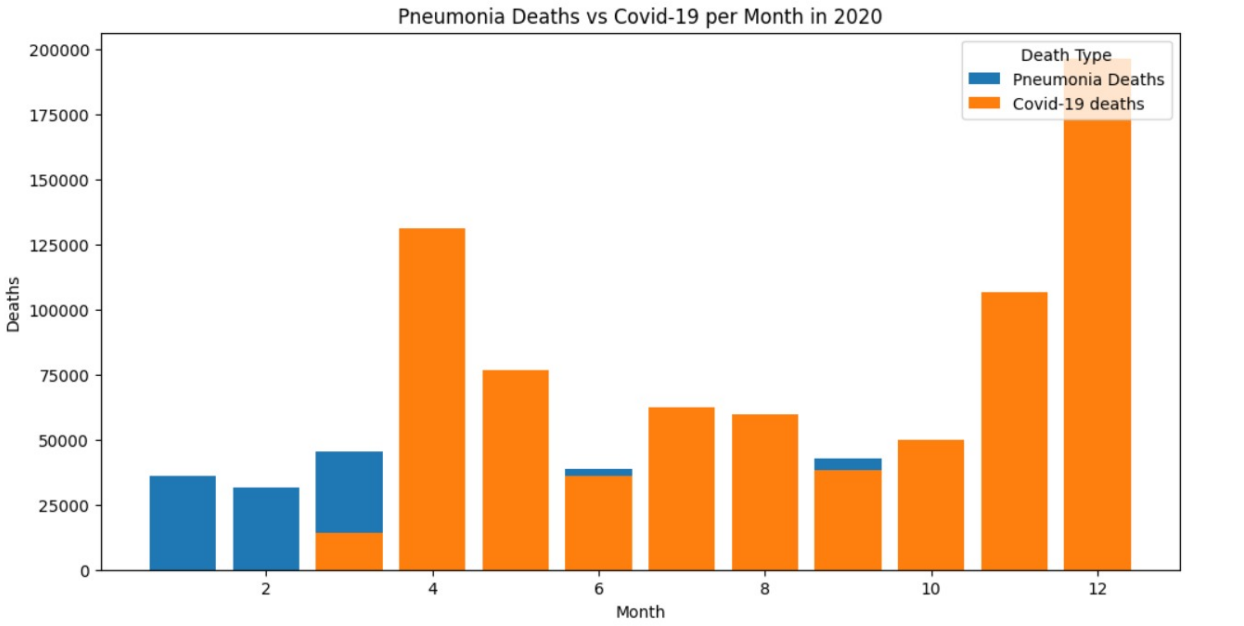
Since our regression model works best for linear data sets, the RMSE for the second graph is much lower than the first graph.

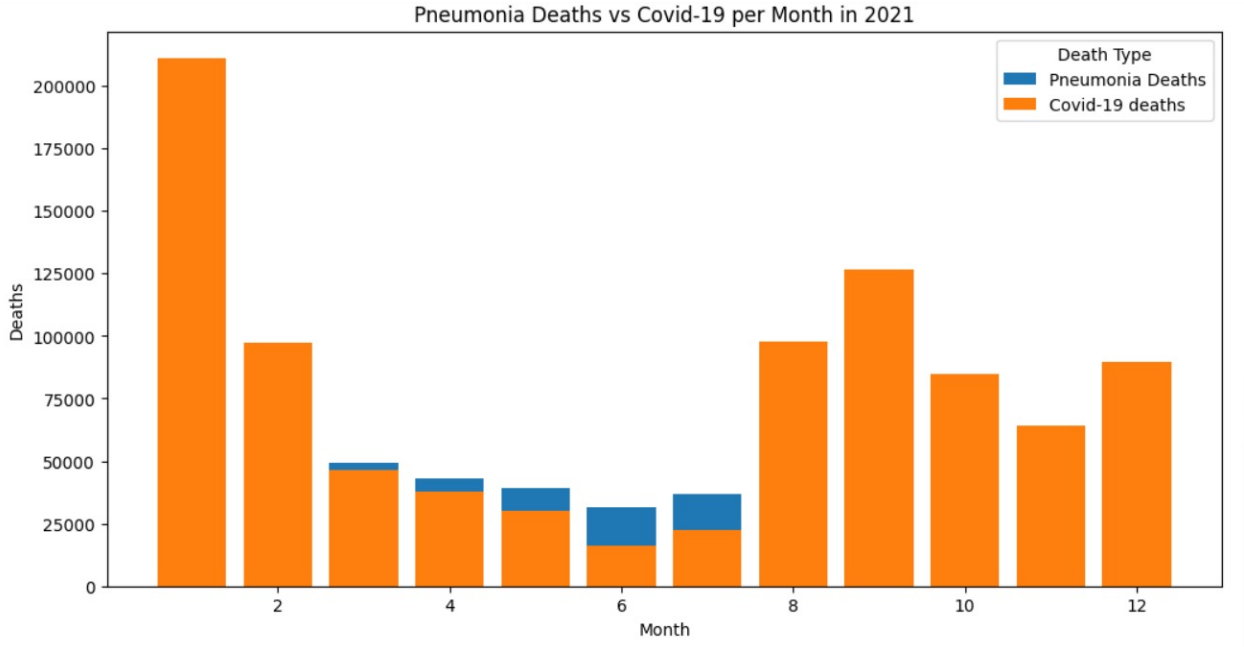
This is synchronous for what happened in real life. In the last COVID-19 wave, the state of California was much better equipped to handle the situation, leading to less number of deaths despite a sharp rise in number of cases.

Finally, we are trying to predict the next six months of COVID-19 deaths for California.



Insights and relations between respiratory diseases to test if any correlation and causation is present because one of the severe complications of COVID-19 is viral pneumonia, yielded the following result. COVID-19 patients can also develop bacterial pneumonia as a secondary infection. We first plot graphs to inspect the number of deaths due to COVID-19 in different years.





We then looked into a year-wise monthly comparison of COVID-19 and pneumonia deaths and noticed that the pneumonia deaths decreased considerably during covid phase. We noticed that it might be a misclassification of death as COVID-19 during this phase where COVID-19 deaths were high in number and overwhelming the US healthcare system.

To determine the age groups most likely to be impacted by COVID-19 by analysing the number of cases and fatalities, the data sets were tested with multiple models. Following are the best results for the different models in the 17-class classification task with their train and test (Year 2022 data).

```
Training Accuracy% :  5.509387503847337
Test Accuracy% (Year 2022 data) :  7.1714476317545035
```

The linear regression model is not suitable for a classification task. The following are rest of the results from various models:

```
Training Accuracy - Ridge Regression (alpha=0.5)% :  5.509387503847337
Test Accuracy - Ridge Regression (alpha=0.5)% (Year 2022 data) :  7.1714476317545035
```

```
Training Accuracy - SVM% :  16.08187134502924
Test Accuracy - SVM% (Year 2022 data) :  12.241494329553035
```

```
Training Accuracy - Logistic Regression% :  16.389658356417357
Test Accuracy - Logistic Regression% (Year 2022 data) :  13.075383589059372
```

```
Training Accuracy - Random Forest (n=500)% :  53.81271160357033
Test Accuracy - Random Forest (n=500)% (Year 2022 data) :  13.942628418945963
```

It was observed that the Random Forest Classifier was best suited for the targeted data set.Random Forest is capable of capturing complex relationships and interactions between features. This is particularly important in tasks with a larger number of classes, where the decision boundaries can be intricate and nonlinear.The ensemble nature of Random Forest helps reduce over fitting, which is crucial when dealing with a substantial number of classes. Individual decision trees might over fit the training data, but the aggregation of multiple trees tends to generalize well to unseen data.

6 Discussion and Conclusion

The primary limitation of the time-series analysis of the COVID-19 Data was the assumption of a linear relationship between the time variable and the number of cases, which oversimplified the complex dynamics of the COVID-19 spread. This model struggled to adapt to sudden changes in the pandemic's trajectory, such as new outbreaks or the impact of public health interventions. Most importantly, COVID-19 came in waves leading to spikes in the graph. This caused issues when adapting to a linear model.

Performing a multi-class classification task can be challenging, especially when there are 17 classes to segregate with limited data from the dataset. After analyzing the situation, we have concluded that reducing the number of classes to 5-7 major classes and modifying the overlapping labels to correspond to the new major classes Y' can provide more training samples for the models. This modification will help the models to better discriminate between the Y' classes.

Sophisticated machine learning models, such as Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs), and Wide Deep Learning (Wide Deep), offer a more advanced approach to capturing complex relationships within the "Provincial COVID-19 deaths by Sex and Age" dataset. These models are particularly useful for handling sequential or structured data, allowing them to learn intricate dependencies and patterns within the age labels for improved predictions. However, it's important to note that these models require more computational resources and power compared to simpler models like decision trees and SVCs. Therefore, the choice of models for this report was limited by the available resources and the desired level of model sophistication, ensuring a practical and efficient result for the classification task.

7 Video Link

Link for presentation on Youtube

References

[1] Yuri Kravchenko, Nataliia Dakhno, Olga Leshchenko, and Anastasiia Tolstokorova. Machine learning algorithms for predicting the results of covid-19 coronavirus infection. In *IT&I Workshops*, pages 371–381, 2020.

[2] Raghavendra Kumar, Anjali Jain, Arun Kumar Tripathi, and Shaifali Tyagi. Covid-19 outbreak: An epidemic analysis using time series prediction model. In *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 1090–1094, 2021.