A picture containing hand, decorated, colorful, close

Description automatically generated

**DAB 304** **Healthcare Analytics**

**Project Report**

**Prediction Of COVID-19 Cases**

**Created By:**

* **Mirenkumar Patel (0782036)**
* **Priyal Patel (0795099)**
* **Yashkumar Patel (0797825)**
* **Navjot Kaur (0796641)**

# Introduction

The World Health Organization (WHO) has designated the current pneumonia outbreak, which started in December 2019 near Wuhan City, Hubei Province, China, as being caused by a novel coronavirus (CoV) known as "2019-nCoV," "2019 novel coronavirus," or "COVID-19." The pathogenic virus COVID-19 is. Bats appear to be the COVID-19 viral reservoir, according to the phylogenetic analysis done with the entire genome sequences that are now available, but the intermediate host(s) has not yet been identified.

The Coronavirus Disease Pandemic of 2019 (COVID-19) serves as a timely reminder of the characteristics and consequences of Public Health Emergencies of International Concern. Since the beginning of the pandemic, there have been over 5.5 million reported fatalities and over 314 million reported illnesses as of January 12, 2022. In terms of instances and fatalities, the COVID-19 pandemic manifests itself in many ways across the globe's various regions and nations.

According to the system of equations, it should be possible to estimate the model's parameters using time series data on each state's population size. In reality, it is uncommon for modellers to obtain a precise count of participants at each step, and the variables might vary over time. Various strategies have been used to address the issue.

This paper's goal is to present a technique that, using just recorded incidence instances, may be used to accurately estimate how an epidemic will develop over the course of the following two to three weeks. We won't use the death rate in our models since overall mortality is rather low, thus we won't. Practical planning queries raised by local and state officials in charge of allocating resources and safeguarding population health served as the inspiration for this study. When the pandemic first began. They later utilized the predicted cases to support estimates of hospitalization and related health resource consumption.

# Related Work

In this section, we will be looking at the work done by other groups and comparing it with our work. As this is to confirm that our present project results are comparable to or in line with theirs, the fields of work should be similar But there were none found for modelling this particular dataset. So here are some insights from our model.

The best model, which employed the Decision Tree algorithm and the Split Validation technique, achieved an overall accuracy, precision, and sensitivity of 96.25%, 99.91% and 92.58%, a specificity of 99.92%, and an AUC of 0.963. The findings were encouraging. The predictive model may therefore be used in clinical decision support systems in the future after more testing. But this model like all other models our model also had a limitation.

To get to this result we had to go through all the evolutionary statuses of the patients in our data. And we also had to divide the dataset into more parts to achieve this result.

Then we used Tableau for analysis purposes after cleaning the data using Python there were a few changes that needed to be made for the cleaning purpose.

# Methods

To run a certain piece of the code or all of the code sections in the Python notebook(Jupyter lab), we loaded the necessary libraries and packages that were essential to us in order to perform certain tasks (ipynb file). Then, for further processing, we loaded our dataset in the file.

We also performed some analysis using various tools. We did had many ideas on how we wanted to go with the analysis part of this project. So in order to achieve this we needed to make many changes to achieve the desired output.

We did use data preprocessing which is a technique to clean the data. We had to clean the data by removing the null values and thereafter we removed some unwanted data fields we had no use in our project. After doing so we had to save the cleaned file and then we performed the analytics part of the project using that clean file.

The preprocessed dataset revealed no abnormalities, and everything appeared to be in order. This means that when our dataset is put into the machine learning models, it is ready for additional processing.

## Machine Learning Models:

Algorithms that can be enhanced over time are the subject of machine learning research. It is a branch of artificial intelligence that is used to create models that can analyze sample data, sometimes referred to as train data, in order to train and test the model to make more precise predictions for test data in the future. Our dataset was processed using machine learning models. The methods are

* **Random Forest (RF)**
* **Logistic Regression (LR)**
* **LSTM(Long Term-short memory)**

These were the few models used by us to get the correct prediction.in order to get the best prediction and result we needed to use some custom function which help us to get the best result possible because without this we were able to achieve this much perfect accuracy and the prediction. But after doing this also we weren’t able to process a large amount of data at the same time and make a prediction.

## Evaluation Metrics:

Evaluation Metrics were used to define a model's capacity. These measures are used to assess the model's performance. It is critical to utilize these metrics since they provide a better knowledge of how each model is operating to its full potential and what may be done to raise the scores later. After doing this we started comparing the accuracy of every model.

# Results:

Below given are the charts:-

Table

Description automatically generated

*Fig:1 Top10 countries affected by COVID-19*

* This table represents the number of COVID-19 cases according to the top 10 affected countries. We can observe that the United States reported 35,689,184 cases, which was the most cases reported by any country in the world, followed by India which also had more than 30 million covid patients. However, the recovery numbers were more amazing for India compared to the United States.

Chart, pie chart

Description automatically generated

*Fig:2 Region Wise case comperision.*

Chart, histogram, waterfall chart

Description automatically generated

*Fig:3 Month Wise covid cases(Top-10 Countries most affected)*

* These charts Represent monthly coronavirus cases for the top 10 countries. It is clearly visible that Austria had reported the highest number of corona patience throughout this period among other countries.

Chart, bar chart

Description automatically generated

*Fig:4* Avg. Death based on country

* Average Deaths for each Province broken down by Country. The color shows details about Province. The data is filtered on a minimum of the Last Update. The view is filtered on Province, which keeps England, Maharashtra, New York, Rio de Janeiro and Sao Paulo.

Chart, bar chart

Description automatically generated

*Fig:5* Deaths and Recovered for each Province/State

* Deaths and Recovered for each Province/State. The color shows details about Deaths and Recovered. The view is filtered on Province/State for Average Death.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

*Fig:6* Model Comparison

* We compared the model which was created by us using different techniques. The second model is predicting the most accurate result of covid-19 cases.

# Conclusion:

IT systems are revolutionizing the healthcare sector in ways that were never imagined, from the creation of novel treatments and illness cures to the advancement of patient diagnostics and clinical status monitoring. As a result, from small clinics to massive hospitals throughout the globe, the advantages of adopting IT technologies in clinical settings—such as enhancing patients' quality of care and maximizing health institutions' resources—have come to be generally acknowledged.

We created a model which wasn’t able to predict if we feed a large amount of data then the prediction would not be even close to the actual data. So we need to limit ourselves so that we could make a proper prediction. After doing so we were able to get the best prediction we could possibly get.

But not being able to make a prediction with all of the data we were having was the biggest drawback of our project. So we will be trying to improve this in future.

In the future, data from various healthcare facilities should be gathered in order to provide a dataset that is richer, more diverse, and has a more equal distribution of classes. This would reduce the need for data sampling approaches and improve the accuracy and realism of the models. The data capture step of this study also does not take the collection window into account, which is another disadvantage.

# Contributions

|  |  |
| --- | --- |
| Name | Contribution |
| **Mirenkumar** | * We had set up a meeting for the selection of the dataset. * I was leading the group and divided work according to everyone’s specialty * also created ML models and visualization, Documentation, Weekly meetings etc. |
| **Priyal** | * She did work on virtualization on Tableau. |
| **Navjot** | * She was working on data cleaning and transformation * She also did some research about the project that helped the project |
| **Yashkumar** | * created some tableau visualization as well as helped in making ppt and made some weekly reports and final report |

In the end, it was teamwork..

# References

* **Datasets (Kaggle):**
  + [**Novel Corona Virus 2019 Dataset | Kaggle**](https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset)
* **Corona:**
  + [**https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e**](https://towardsdatascience.com/lstm-recurrent-neural-networks-how-to-teach-a-network-to-remember-the-past-55e54c2ff22e)
  + [**https://intellipaat.com/blog/what-is-lstm/**](https://intellipaat.com/blog/what-is-lstm/)
  + [**https://www.who.int/en/activities/tracking-SARS-CoV-2-variants**](https://www.who.int/en/activities/tracking-SARS-CoV-2-variants)
  + [**https://www.nature.com/articles/s41598-022-15968-z**](https://www.nature.com/articles/s41598-022-15968-z)
* **Machine Learning:**
  + [**https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/**](https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/)

# Tools:

* **Anaconda Navigator 3**
* **Python 3**
* **Tableau 2019.1**
* **MS Excel 2016**