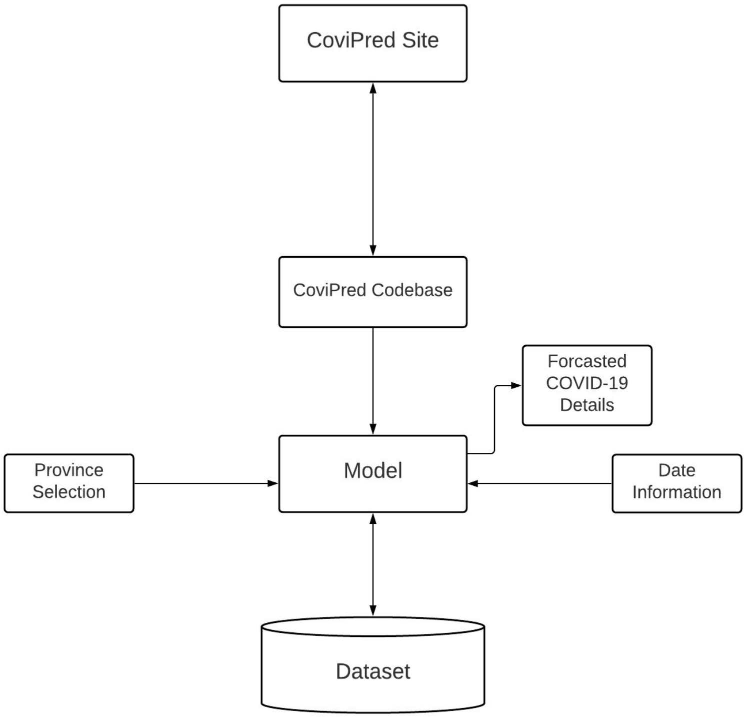
NASH-ML
Codefest Datathon 2020 Report

Group Number - DA2049

Dataset - https://bit.ly/3kNGrcD

High Level Diagram



The **COVID-19 Site** is the Website where the User can **Forecast COVID-19 Details** by entering **Date Information** and **Selecting a Province,** that is predicted using a **Model** that learns to forecast through a **Dataset** containing Columns such as Province, Prefecture, Date and Cases.

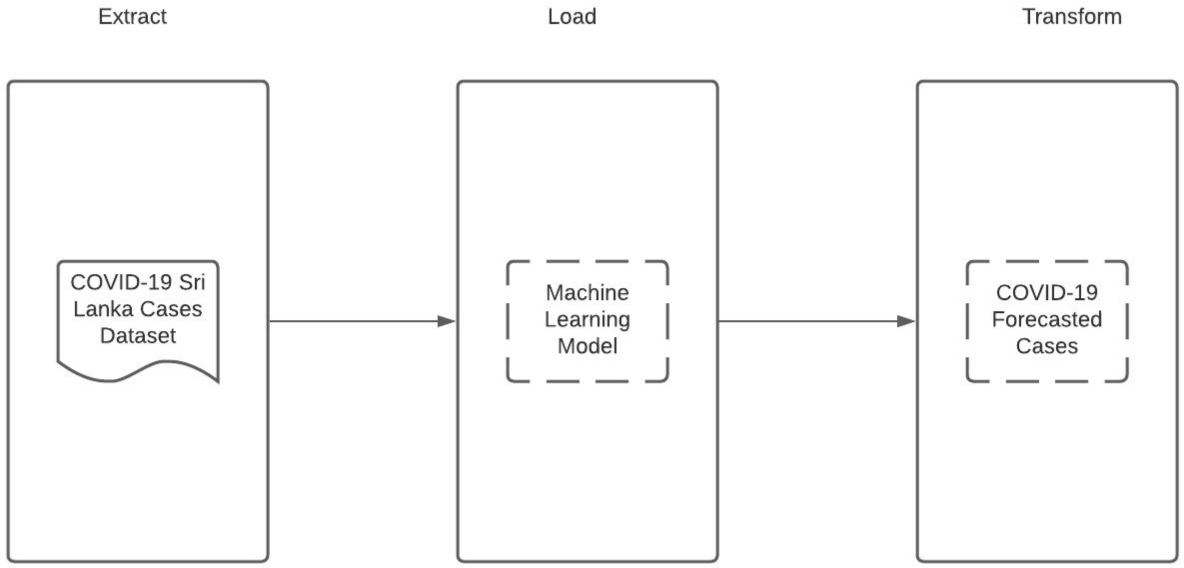
CodeBase

Extraction Method

The method used was the Long Short Term Memory algorithm. LSTMs are an improved version of rnns, rnns are used because they take into consideration the output from the previous layer.

Since time series analysis forecasting greatly depends on the previous layers output (the weather of yesterday greatly influences the weather of today or tomorrow), rnns were used. However, regular rnns suffer from the vanishing gradient problem, because an increase in the number of steps can greatly affect the gradient., since it updates itself each time step. LSTMs, on the other hand make use of a cell state, which only updates itself, when explicitly made to.

ETL Model



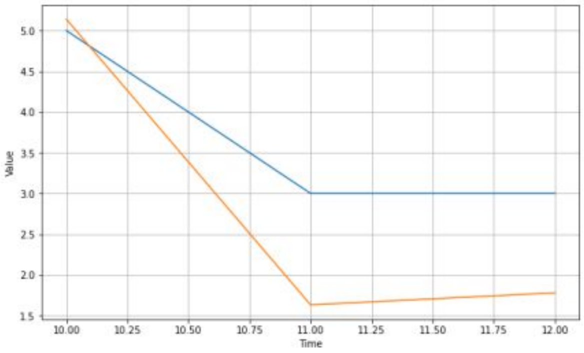
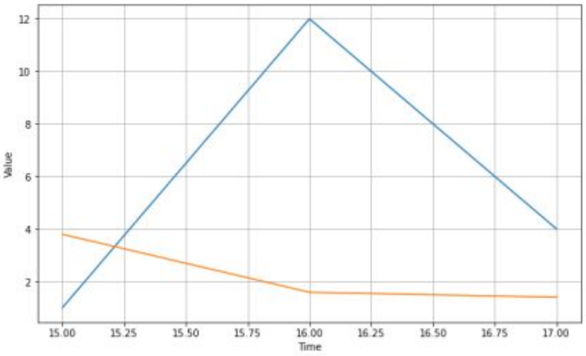
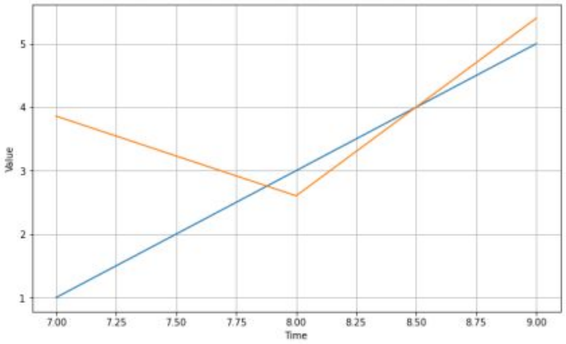
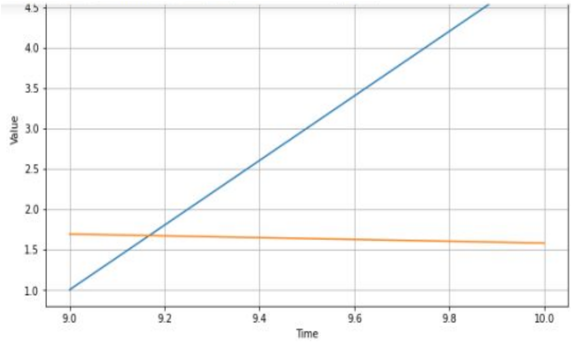
Data Science Analysis, Model Implementation and Evaluation

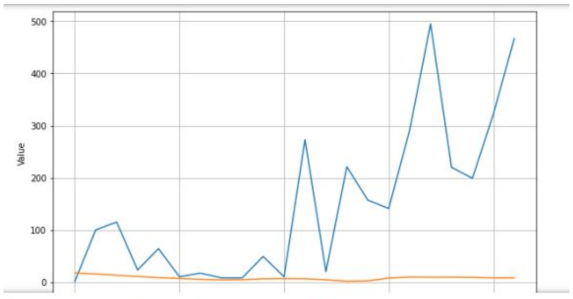
In the context of data analysis we had to inspect the data and perform cleaning of data, transforming and modelling the data.The dataset we found was a csv file so pandas library was used to read the data from the csv file and using the functions provided by pandas we investigated the dataset by checking the content of the head of the data, columns etc…

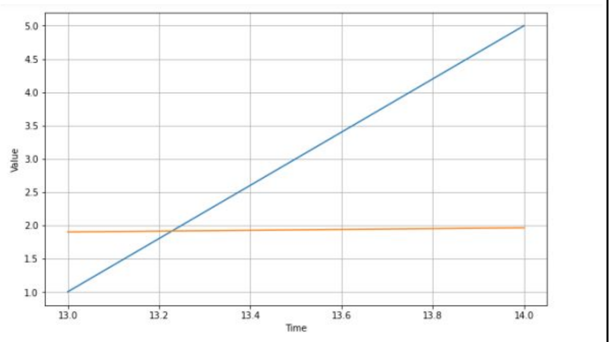
We had a lot of data cleaning process to execute; the dataset found had a lot of missing values for some columns and some useless data, so to clean all these and create a new cleaned dataset for model training we had to use a number of visualizations libraries such as sea born to visualize the missing data and we had to combine data from a number of columns into one. This is because some data was scattered into a number of fields, so instead of dropping the data we combined the columns with data related into one, therefore we won’t lose accuracy when training the model.

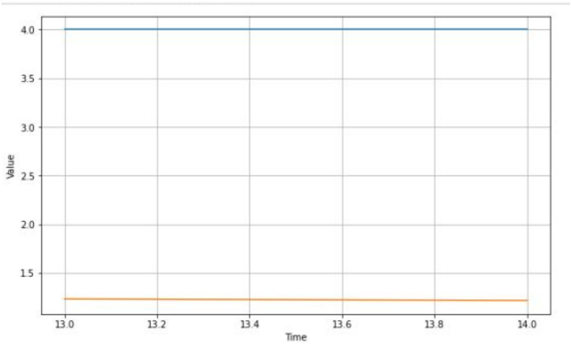
Visualizations

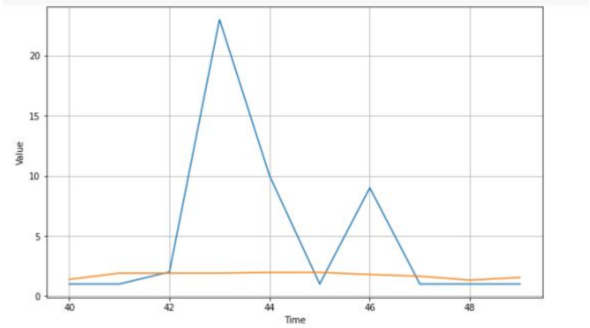
These are Time Series Graphs which display Forecasted COVID-19 Cases in Provinces Overtime. Below we have

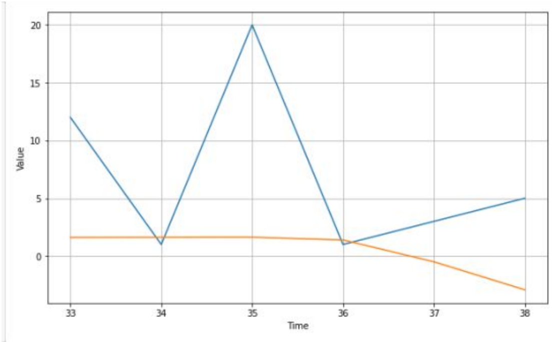
Northern	Eastern
	
Central	**Sabaragamuwa**
	

<table>
<tr><th>Western</th><th>Uva</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><th>Southern</th><th>Northern Central</th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><th colspan="2">North Western</th></tr>
<tr><td colspan="2">


</td></tr>
</table>