A TECHNICAL REPORT ON THE PREDICTIVE MODELLING FOR COVID-19 IN PUBLIC HEALTH

PROJECT

BY

DANIEL NICHOLAS DIBAL

Fellow ID
FE/23/67260109
Cohort 2

Applied Learning Cluster
Development Hub
Applied Learning Cluster ID:TP/23/26244738
Phone Number:'+234 803 044 5808
Email: Ababa@developmenthub.ng
Address: "No. 15B Behind Nursing Home,
Ibrahim Wadaima Road Off SS Kyari Crescent
Old GRA Maiduguri, Borno State"

Course
Data Science

INTRODUCTION

Hi everyone! In this project, I worked on building a system to predict and understand COVID-19 trends. The goal was to help public health organizations make informed decisions by analyzing past data, forecasting future cases, and identifying high-risk areas.

To tackle this, I cleaned a raw COVID-19 dataset, visualized key trends, and developed predictive models.

Here is the breakdown!

STEP 1: DATA PREPARATION

The raw dataset wasn't perfect—like most real-world data, it had issues. Some values were missing, and the **Date** column wasn't in a consistent format. Here's what I did:

- Filling in Missing Values: I forward-filled missing data to keep trends consistent and filled **Recovered** cases with 0 where data was unavailable.

- Adding Features: I created new columns like:

i. **Active Cases**: Confirmed - Deaths - Recovered

ii. **Daily Metrics**: Day-by-day changes in confirmed cases, deaths, and recoveries.

iii. **Rates**: Growth rate, mortality rate, and recovery rate to make comparisons easier.

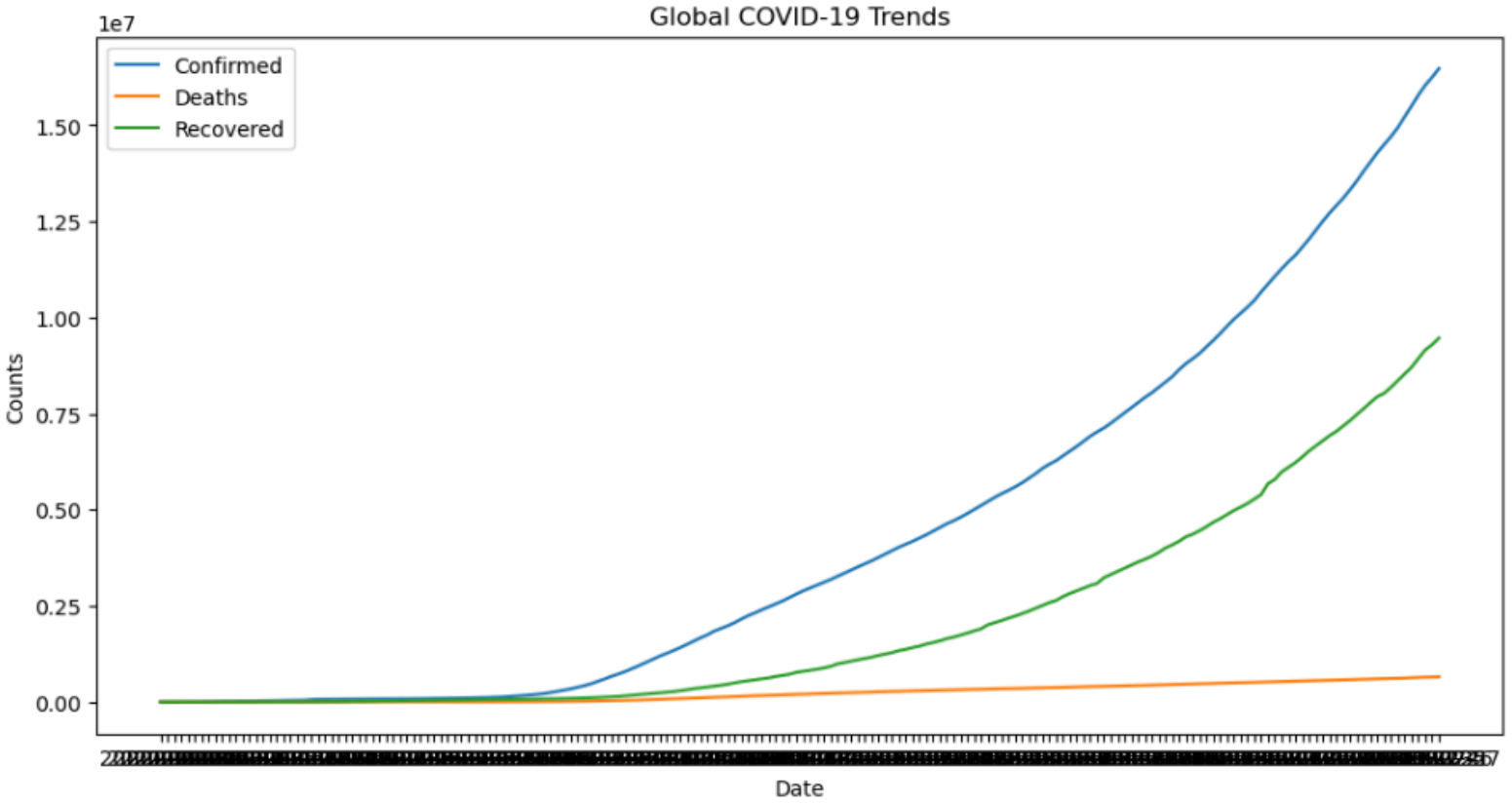
This cleaned dataset was saved as **final_covid_data.csv**, ready for analysis.

STEP 2: EXPLORING THE DATA

Once the data was clean, I visualized it to understand the story it told.

Global Trends

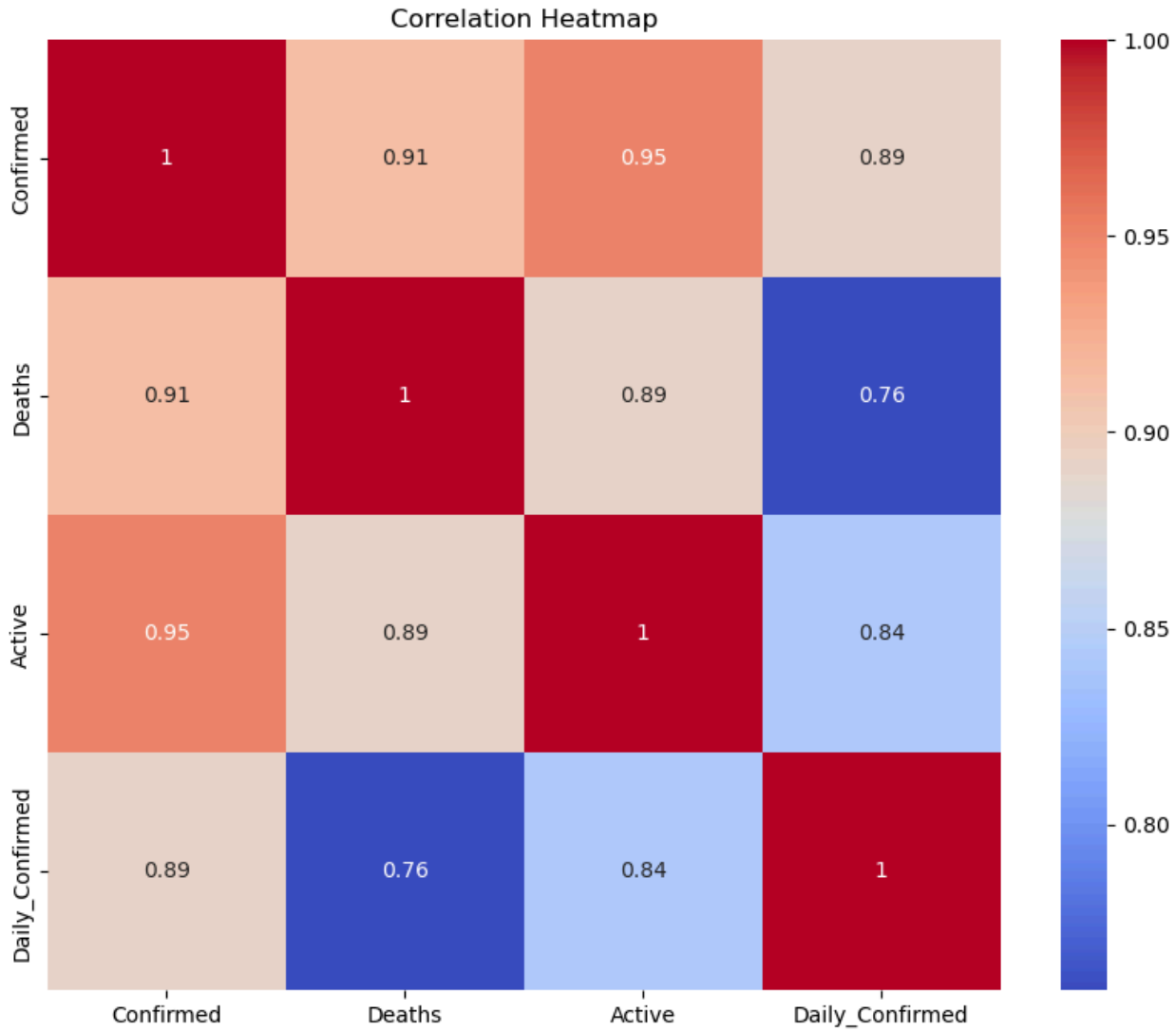
First, I plotted confirmed, recovered, and death trends over time. It was fascinating (and sobering) to see the spikes during various waves. Recoveries followed behind confirmed cases, and the death curve had a noticeable lag.



Correlations

Using a heatmap, I checked how different variables—like daily confirmed cases, deaths, and active cases—were related. Unsurprisingly, confirmed cases and active cases were closely tied. Interestingly, I also noticed moderate relationships between growth rates and deaths.

These insights helped frame how the pandemic evolved and highlighted the importance of early interventions.



STEP 3: PREDICTIVE MODELLING

After exploring, I moved to the fun part: building predictive models.

1. Time-Series Forecasting (ARIMA)

I used an ARIMA model to forecast the next 30 days of confirmed cases globally. Here's what went into it:

- **Stationarity Check:** The dataset needed differencing to stabilize trends.
- **Model Tuning:** I picked (5, 1, 0) for the ARIMA parameters after testing a few combinations.

Results: The model gave a 30-day prediction, which aligned well with past trends. It showed steady growth, emphasizing the ongoing challenge.

---

2. Classification (Random Forest)

For this part, I wanted to classify countries into high-risk or low-risk categories based on their mortality rates. Here's how I approached it:

- **Feature Selection:** I used columns like confirmed cases, deaths, active cases, and growth rate as predictors.

- **Handling Missing/Infinite Data:** Some values in the data were infinite (thanks, math!) or missing. I replaced them with averages using imputation.
- **Model:** A Random Forest Classifier was trained to predict whether a country's mortality rate was above 5%.
-

Results:
The model performed really well:

- **Precision:** 91%
- **Recall:** 90%
- **F1-Score:** 92%

This means the model was accurate in identifying high-risk countries while minimizing false predictions.

STEP 4: INSIGHTS AND VISUALIZATION

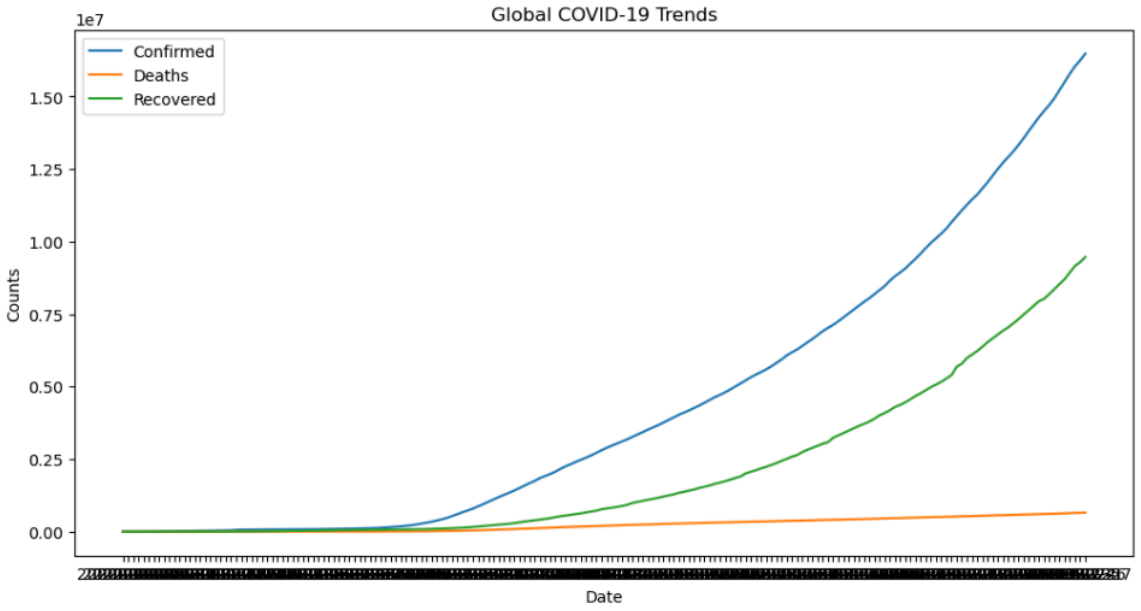
Here's what stood out from the analysis:

1. **Lockdowns Work:** The data clearly showed a drop in growth rates after strict lockdowns.
2. **High Mortality Regions:** Countries with limited healthcare saw significantly higher mortality rates.
3. **Future Outlook:** Forecasts showed continued case growth, stressing the need for ongoing vigilance.

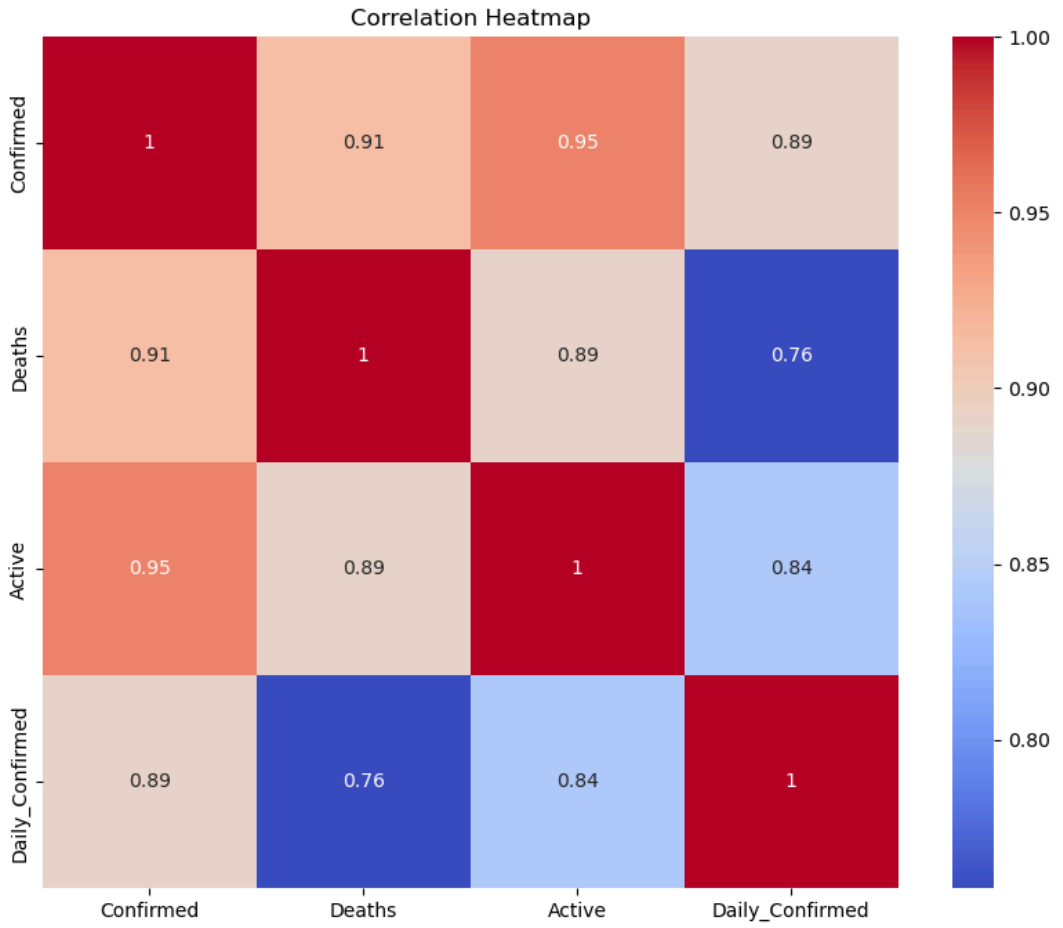
Visuals:

To make the data approachable, I created:

- **Line Charts:** Showing trends in confirmed, recovered, and death counts.



- **Heatmaps:** Visualizing correlations between variables.



- **Dashboards (Power BI):** Highlighting global trends, KPIs, and risk maps.