BUSA8031 BUSINESS ANALYTICS PROJECT REPORT

ASSIGNMENT 3

PREPARED BY:

AADHAR BAHETI, 46813292
IPSITA NATH, 46228136
PURUSHAARTH KUMAR, 46284702
SIDDHI VINOD KORGAONKAR, 46808043

MACQUARIE University
SYDNEY·AUSTRALIA

Introduction

Data has become an essential tool for informed decision-making and policy formation in the aftermath of the COVID-19 epidemic. The Agency for Clinical Innovation (ACI) recognised the enormous value inherent in the massive expanse of COVID-19 data and went on an exploratory journey to extract critical insights. While the data landscape provides multiple features and parameters for research, this report focuses solely on two critical dimensions: hospitalisation and policy analysis. Our study aims to uncover the complexities of pandemic waves, comprehend the hospitalisation load and excess mortality, assess the influence of vaccination on case numbers, and examine the interaction between policy implementation and case trends. We conducted a comparison analysis of Australia and Canada, two nations with similar population densities, to get deeper insights into the unique dynamics of the pandemic and to analyse the interplay between policy implementation and Hospitalisation.

We will share our findings, highlighting the power of data to influence health-related decisions in Australia and abroad. We hope to uncover crucial lessons that will guide our post-pandemic future by focusing on hospitalisation and policy analysis. This research emphasises the importance of data-driven enquiry in a changing global scene that has been dramatically influenced by the COVID-19 . We aim to shed light on the pandemic's intricate dynamics and enable more informed decision-making by comparing Australia and Canada, two nations with similar population densities. In addition, we will examine Australia's response to the pandemic, analysing the success of the rules put in place and obtaining insights into the techniques that were crucial in crisis management.

Background

The COVID-19 pandemic occurred in three waves, each with its own characteristics and global consequences:

First wave (early 2020): Original outbreak, spread quickly, overwhelmed healthcare systems.

Second wave (mid-2021): Driven by Delta variant, increased transmissibility, strain on hospitals, led to increased vaccination efforts.

Third wave (early 2022): Omicron variant, rapid transmission, milder symptoms, new concerns about immunization efficacy.

"Hospitals worldwide were overwhelmed, and governments launched various remedies, highlighting the dynamic interplay between countries and their distinct public health strategies. The pandemic shed light on the growing importance of vaccination in mitigating its impact."

DATA SOURCE AND METHODOLOGY:

Data Acquisition and Preparation:

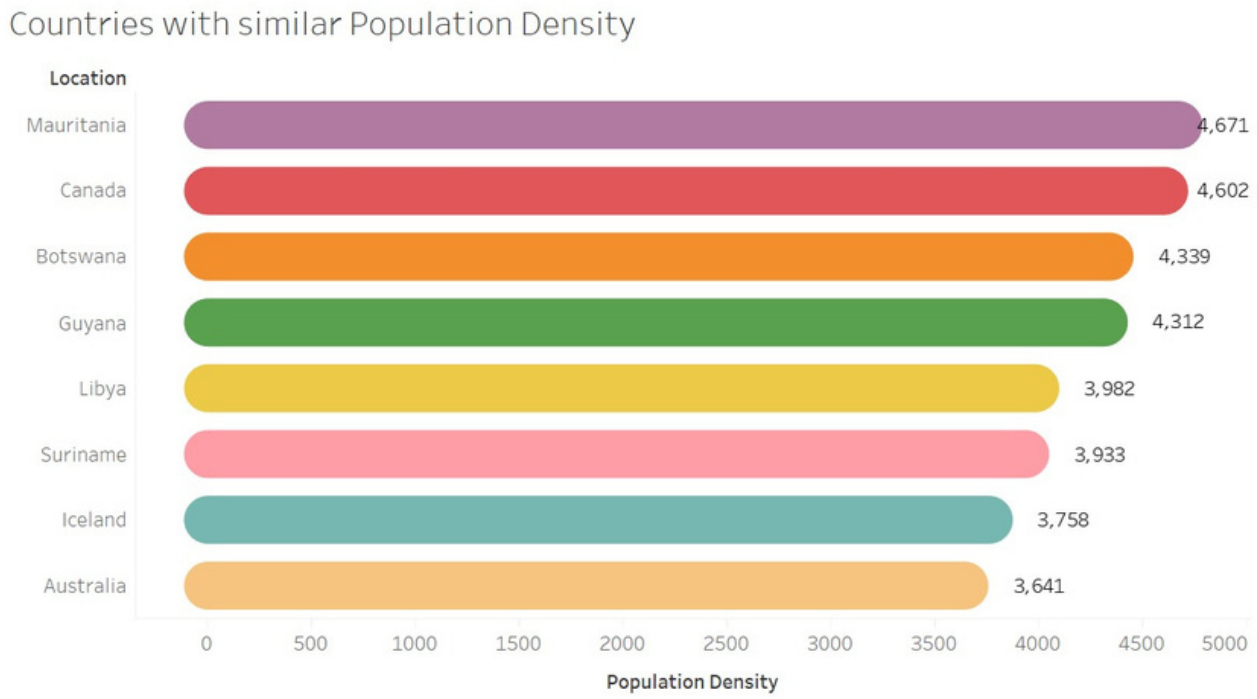
Data acquisition and preparation are crucial steps in conducting an in-depth analysis of the COVID-19 pandemic. To perform our study, we relied on the COVID-19 dataset provided by Our World in Data (OWID), which is available through their GitHub repository at https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data-old.csv. We chose this dataset because it covers all the essential parameters necessary for our analysis, such as location, date, total cases, new cases, total deaths, new deaths, reproduction rate, ICU admissions, hospitalizations, total tests, new tests, positive rates, total vaccinations, stringency index, and population density.

Data Preprocessing and Quality Assurance:

Data preprocessing is a critical stage to ensure data integrity. During this phase, we encountered several data quality issues that required our attention. We first removed irrelevant location values that did not pertain to specific countries, such as income categories ("upper middle income" or "lower middle income"). By doing so, we focused our analysis solely on individual countries. We also standardized the date formats to maintain data uniformity, as we found inconsistencies in this regard. Additionally, the dataset contained missing values, which is a common occurrence in COVID-19 data. We opted to omit rows with missing data, as these gaps were not conducive to our analysis.

Focus on Australia:

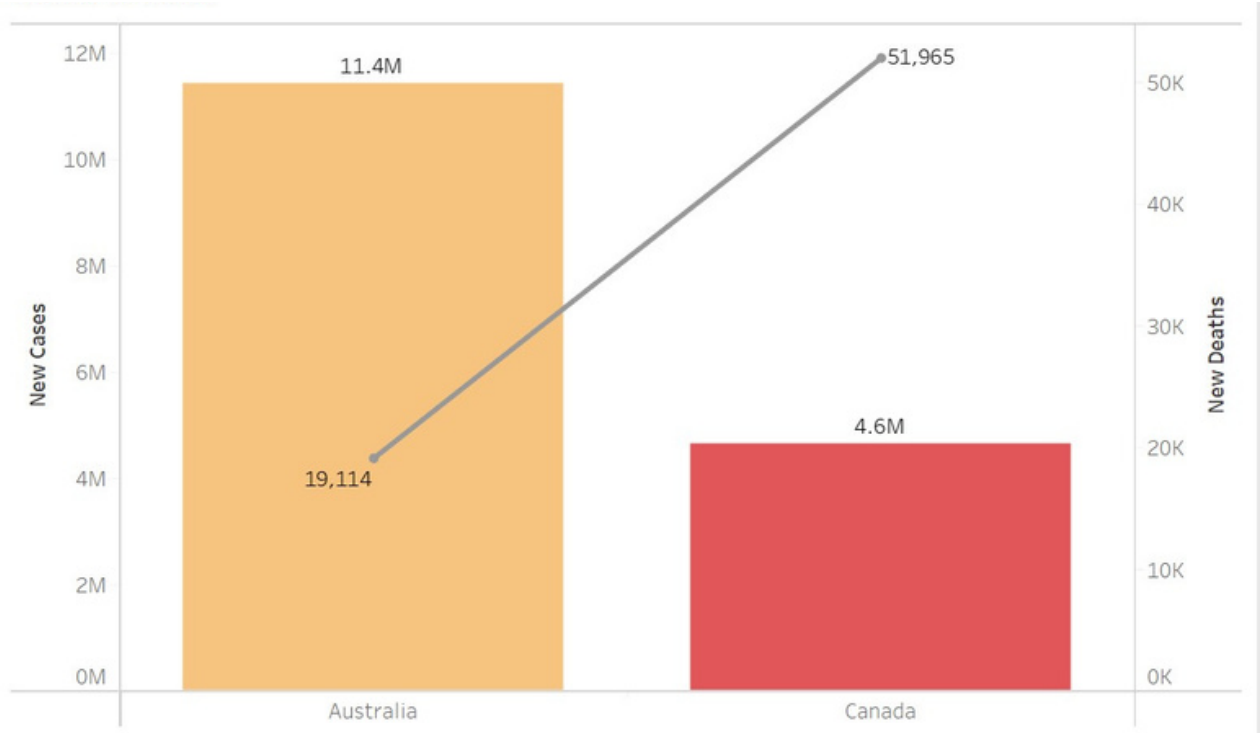
As our affiliation with the New South Wales (NSW) government through the ACI directed our primary focus towards Australia, we filtered countries based on population density to facilitate a meaningful comparative analysis. Australia's population density of approximately 3641 people per square kilometer became our logical criterion, and we filtered countries with population densities ranging from 3000 to 5000 people per square kilometer. This strategic filtration process yielded several candidates, including Mauritania, Canada, Botswana, Libya, Ireland, Suriname, and Australia.



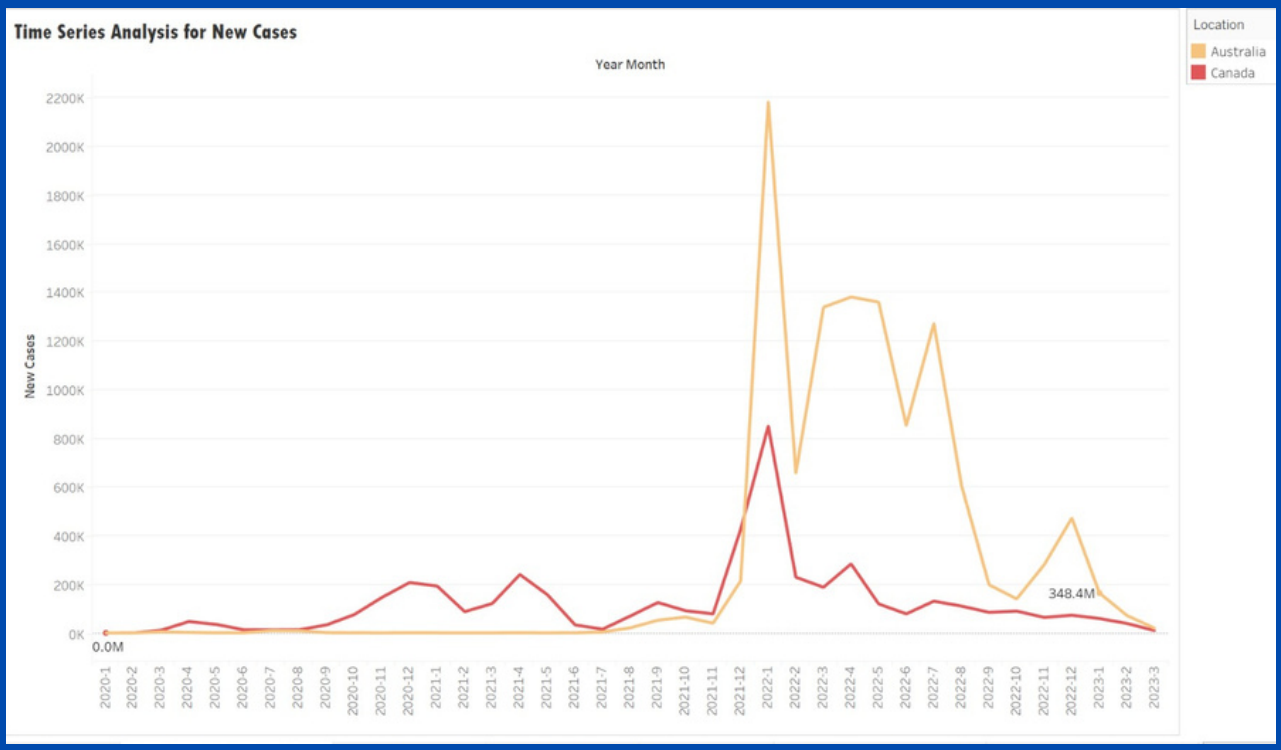
Selection of Comparative Benchmark & Data Availibilty:

To ensure an informed decision, we evaluated the availability of data for these countries. Our assessment indicated data availability ranging from 50% to 72.49% for the selected nations. Mauritania had data available for 55.11% of the time, followed by Canada at 66.12%, Botswana at 56.27%, Guyana at 52.32%, Libya at 50.62%, Suriname at 56.52%, Iceland at 62.12%, and Australia at 72.49%. Given the substantial data availability and the logical selection criteria, we concluded that Australia and Canada would be our focal points for a comprehensive comparative analysis. These two countries provide a compelling basis for assessing and understanding their distinct responses to the multifaceted challenges posed by the COVID-19 pandemic.

In our case study, as mentioned above, we have taken Australia and Canada for the analysis.



The graph above compares the total number of deaths to the total number of cases in Australia and Canada. The data clearly indicates that Australia had a significantly higher number of new cases (11.4 million) than Canada (4.6 million). However, in contrast to the number of cases, Canada had a surprisingly high death rate (~52,000) while Australia's death rate was much lower (19,000). This highlights the difference in how each country managed their medical situations. Australia closed its borders and implemented strict laws and policies, which helped save multiple lives. In contrast, Canada was unable to do the same, leading to a higher death rate.



The above graph is the Time series analysis for new cases in Australia and Canada for each month. As seen above, below are the predictions:

AUSTRALIA

First Wave:

Timeline: The First Wave of COVID-19 in Australia began in early 2020 and peaked around March and April of the same year with very few cases.

Impact: During this wave, Australia implemented strict lockdown measures and travel restrictions to contain the virus. The country saw a relatively low number of cases and deaths compared to many other nations. Quick government response and adherence to public health guidelines played a crucial role in controlling the spread.

Delta Wave:

Timeline: The Delta Wave of COVID-19 hit Australia in mid-2021, with cases increasing in various states, particularly New South Wales and Victoria.

Impact: The Delta variant caused a significant surge in cases, leading to localized lockdowns and restrictions in various parts of the country. Vaccination campaigns were accelerated to curb the spread of this highly transmissible variant. In New South Wales, the state's health system faced challenges with the rising number of hospitalizations.

Omicron Wave:

Timeline: The Omicron Wave reached Australia in late 2021, with cases surging toward the end of the year and into 2022.

Impact: The Omicron variant, known for its rapid transmission, led to a sharp increase in COVID-19 cases across Australia. While case numbers soared, the severity of illness appeared to be lower compared to earlier waves, largely due to high vaccination rates. Authorities encouraged booster shots and adapted public health measures to address the surge in cases.

CANADA

First Wave:

Timeline: The First Wave of COVID-19 in Canada began in early 2020 and reached its peak around April and May of the same year.

Impact: During the initial wave, Canada implemented widespread lockdowns and travel restrictions. The healthcare systems in provinces like Quebec and Ontario faced significant strain. The country recorded substantial cases and fatalities, particularly in long-term care facilities.

Delta Wave:

Timeline: The Delta Wave in Canada occurred in the summer and fall of 2021, with varying timelines across provinces.

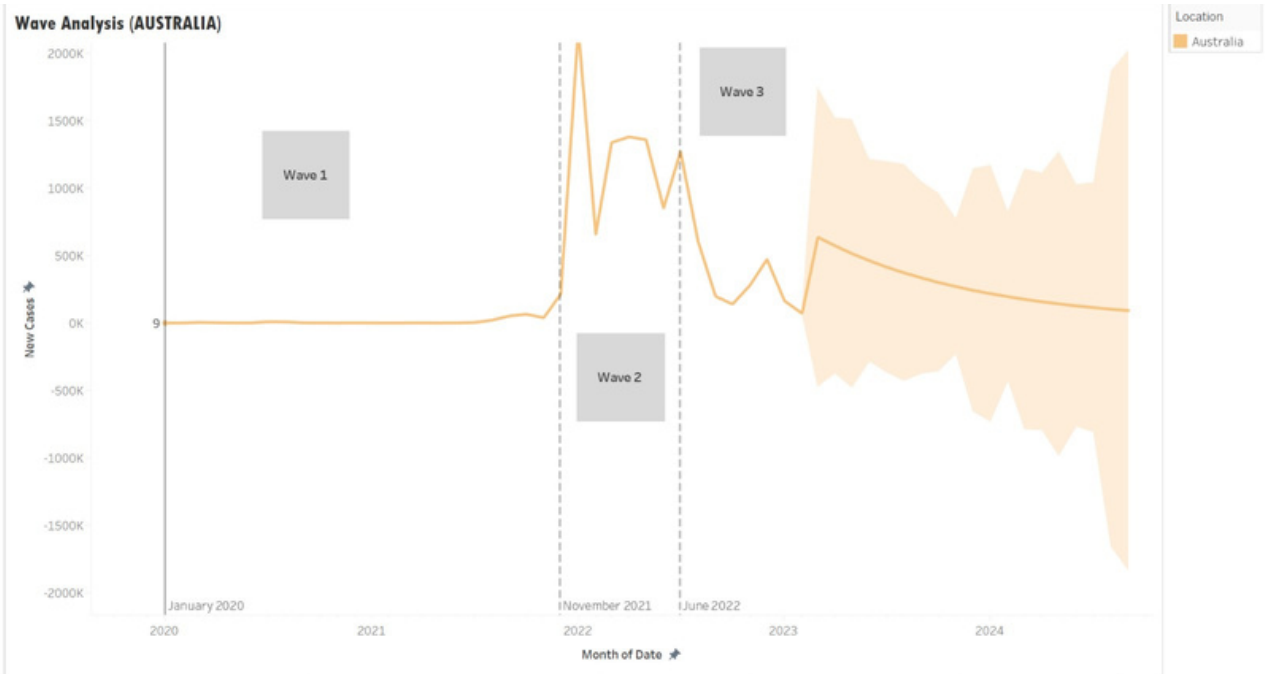
Impact: The Delta variant resulted in a surge in cases and hospitalizations in several provinces. Vaccine campaigns were intensified, and regions like British Columbia and Alberta imposed restrictions to manage the outbreak. Vaccination was identified as a crucial tool in controlling the spread.

Omicron Wave:

Timeline: The Omicron Wave reached Canada in late 2021 and continued into early 2022.

Impact: Omicron led to a rapid and substantial increase in COVID-19 cases, prompting provincial and territorial authorities to respond with a combination of public health measures, testing, and vaccination efforts. While case numbers rose significantly, the proportion of severe cases and hospitalizations was somewhat lower, which was attributed to high vaccination coverage.

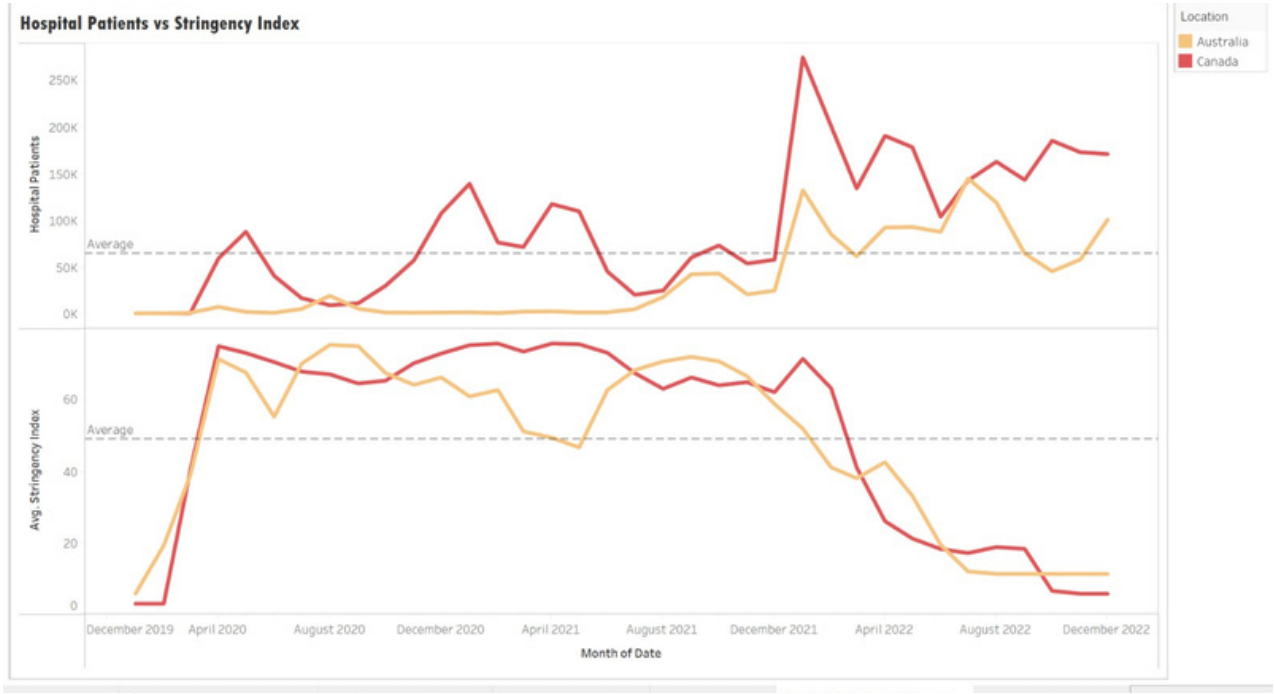
The graph below is the compilation of the waves of COVID-19 in Australia and also the predicted wave analysis based on the actuals of the waves.



Policy Analysis:

We used the "Stringency Index" as a crucial measure of the effectiveness of COVID-19 containment policies in our dataset. The index is a combination of nine different policy indicators such as workplace closures, school closures, and travel bans, and provides a numerical rating from 0 to 100. A higher number indicates stricter policies, while a lower number reflects more lenient policies. Our analysis aimed to understand the impact of these policies on several COVID-19 outcomes including new cases, deaths, hospitalization rates, and ICU admissions. We conducted a comprehensive analysis of these policy measures for Australia and Canada to gain valuable insights into their effectiveness in controlling the spread of the virus and reducing the strain on healthcare systems.

Graph 1: Impact of Stringency Index on Hospitalizations

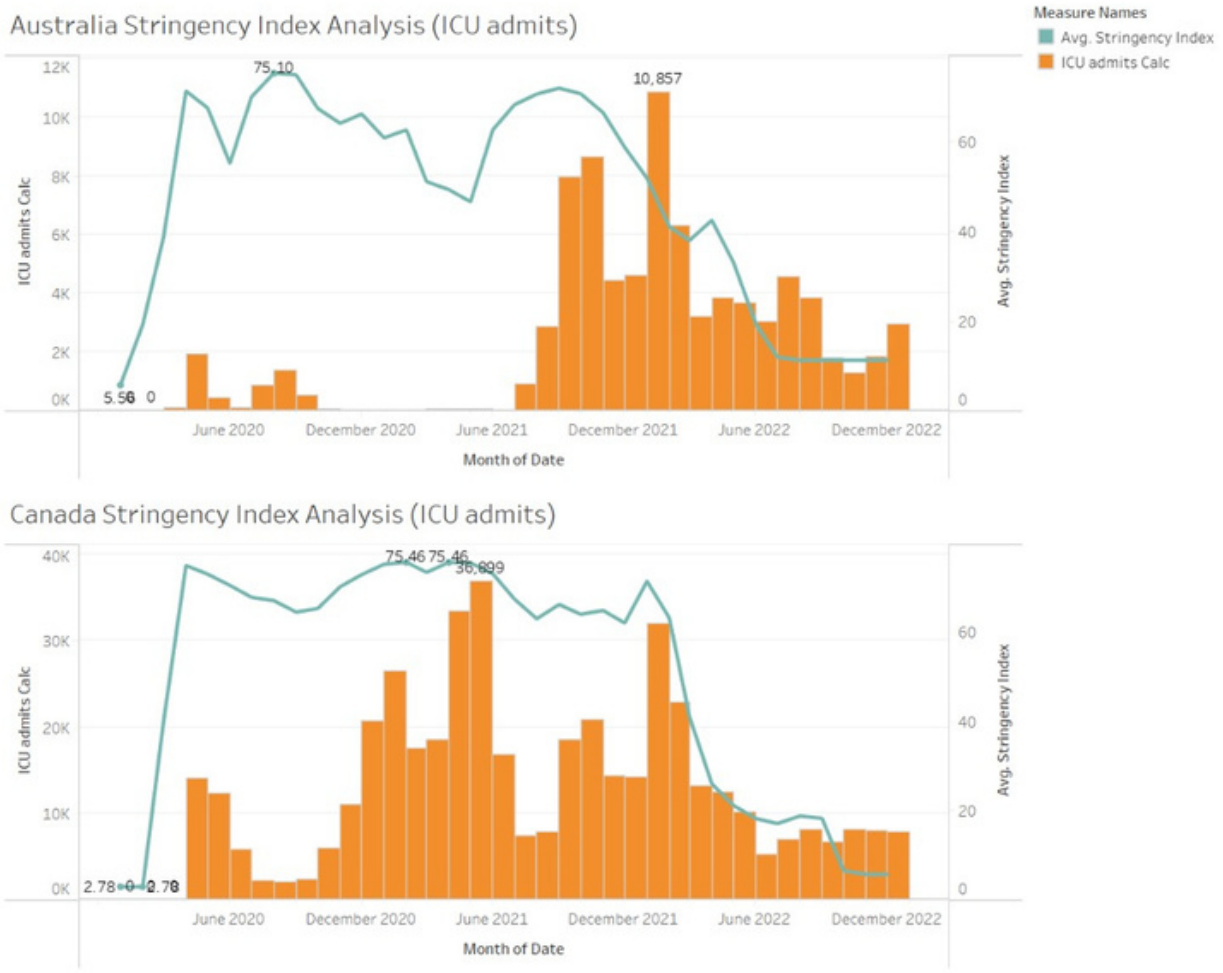


The above graph illustrates the relationship between the stringency index and the number of hospital patients over time. The analysis aims to determine whether stricter policies are associated with a reduction in hospital admissions.

We analyzed data from December 2019 to December 2021 to determine the impact of policy stringency on hospitalization rates. During this period, both Australia and Canada implemented relatively strict policies to control the pandemic. Despite Canada having slightly more stringent policies than Australia, both countries had below-average hospitalization rates, indicating the potential effectiveness of these policies in curbing the virus's spread and reducing hospitalizations.

However, after December 2021, both countries eased their policies, resulting in an increase in hospital patients. Canada experienced a more significant rise in hospital patient numbers than Australia, suggesting that relaxation of policies may lead to an increase in virus transmission and hospitalizations. This information is helpful for policymakers when determining the implementation and relaxation of COVID-19 measures.

Graph 2: Impact of Stringency Index on ICU Admissions



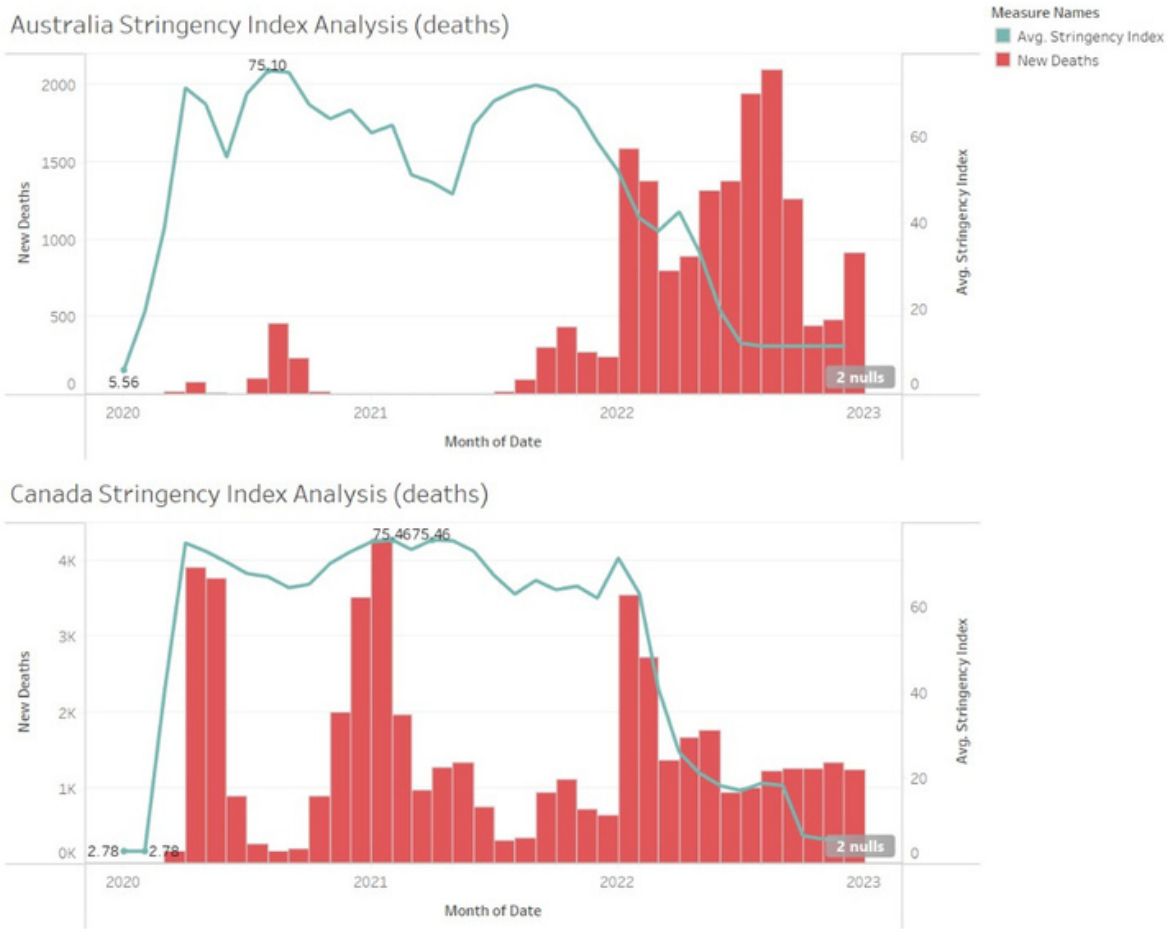
The graph shows the correlation between Stringency Index and ICU admissions in Australia and Canada.

Between June 2020 and September 2021, Australia implemented strict policies with an average Stringency Index of 75.10, which helped to effectively control the situation and resulted in a lower requirement for ICU support. However, starting from January 2022, when the Omicron variant emerged and policies were relaxed, there was a noticeable increase in the need for ICU admissions.

On the other hand, Canada had slightly higher Stringency Index levels (75.46) compared to Australia between June 2020 and December 2021, but the situation did not significantly improve, leading to a rising demand for ICU support. However, from February 2022 onwards, as the situation became more manageable, ICU requirements decreased alongside the easing of policies.

This graph is valuable for ACI in making data-informed policy decisions and allocating healthcare resources effectively by revealing the correlation between policy stringency and ICU admissions. It is crucial for managing the COVID-19 pandemic's impact on healthcare systems.

Graph 3: Impact of Stringency Index on Deaths

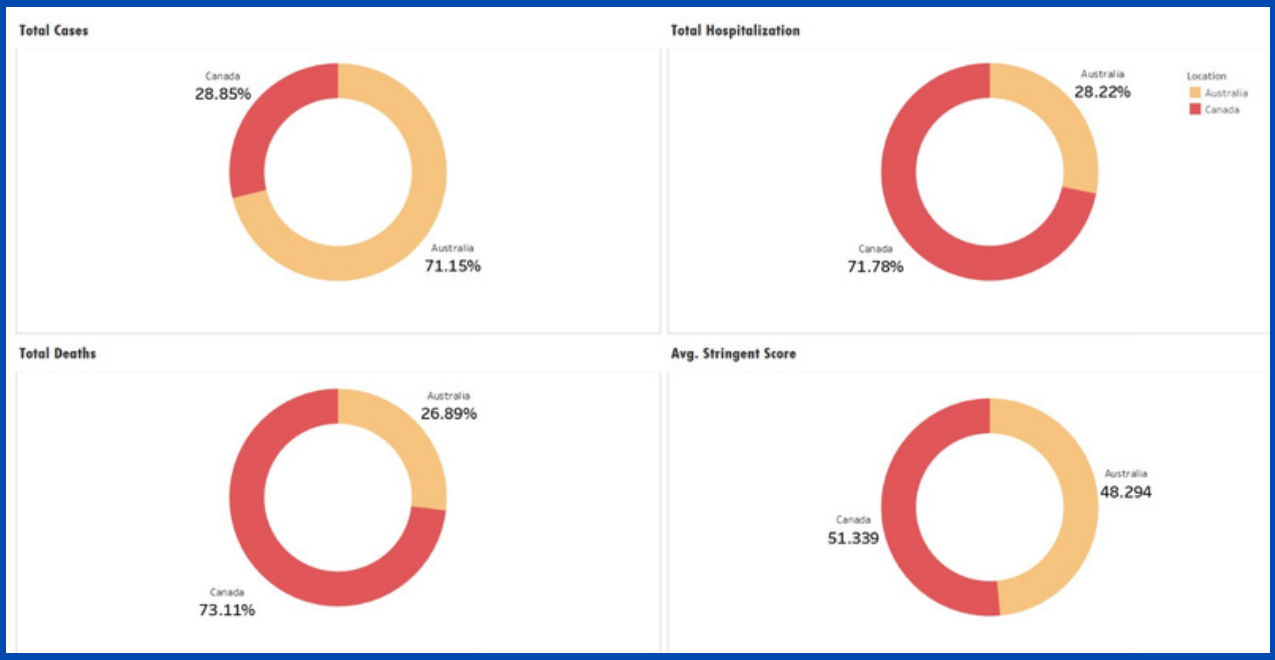


The graph presents an analysis of the impact of the stringency index on the number of deaths in Australia and Canada.

Between 2020 and October 2022, Australia had strict policies in place to control the spread of COVID-19, with a high stringency index. During this period, the number of deaths was relatively low, suggesting that the strict measures were effective in reducing fatalities. However, when the Omicron variant hit and policies were relaxed, the number of deaths increased. This suggests that the relaxation of policies may have led to an increase in virus transmission and consequently, fatalities.

In contrast, Canada had even stricter policies than Australia between 2020 and 2022, but the number of deaths remained high. This implies that the strict policies may not have been as effective in controlling the spread of the virus and reducing fatalities. However, as the number of deaths started to decrease, Canada eased its policies after January 2022.

These findings underscore the potential impact of policy measures on COVID-19 outcomes, especially deaths. Strict policies appeared to be effective in Australia until they were relaxed, while Canada saw high fatalities despite strict policies. However, as the situation improved, easing policies seemed to have a positive effect.



Insights of the graph:

- According to the data, Canada had a higher percentage of total cases (71.15%) compared to Australia (28.85%) despite having stricter policies. This trend was also observed in the total hospitalizations and deaths, with Canada having a higher percentage in both categories. Interestingly, even though Canada had a higher average stringency score (51.33) compared to Australia (46.94), which indicates stricter policies, it still faced more hospitalizations and deaths. This suggests that the effectiveness of policy measures is not solely determined by their strictness, but rather by other factors such as the timing of implementation, public compliance, healthcare capacity, and other societal factors.

- In contrast, Australia managed to keep the total cases, hospitalizations, and deaths relatively low, despite having less strict policies and a lower stringency score. This implies that Australia's response to the pandemic was more effective in controlling the spread of the virus and reducing its impact on the healthcare system. These findings highlight the complexity of pandemic management and emphasize the importance of taking a multifaceted approach that goes beyond policy strictness.

Building a Model to Get Predictions of Hospital Patients

Introduction :

Predictive modeling plays a crucial role in forecasting hospital patient counts and ensuring the effective management of healthcare infrastructure, especially during public health crises like the COVID-19 pandemic. Accurately predicting the number of hospital patients offers multiple benefits to healthcare providers, government entities, and related organizations. This section focuses on the development of a predictive model tailored to estimate hospital patient counts in the Australian context, enhancing our understanding of the pandemic's impact on hospitalizations in the country and providing valuable insights for resource allocation, healthcare capacity planning, and public health strategies. Predictive modeling is of utmost importance in this domain as it empowers healthcare systems and government bodies to proactively respond to the evolving healthcare needs of the population.

Model Development for Australia :

1. Feature Selection

In the development of our predictive model for estimating hospital patient counts in Australia, feature selection plays a crucial role. We carefully selected a subset of features that are highly relevant to the unique dynamics of the COVID-19 pandemic within the Australian context. Our selection was based on the potential impact and meaningful correlation of these features with hospital patient counts, providing valuable insights into the healthcare system's burden.

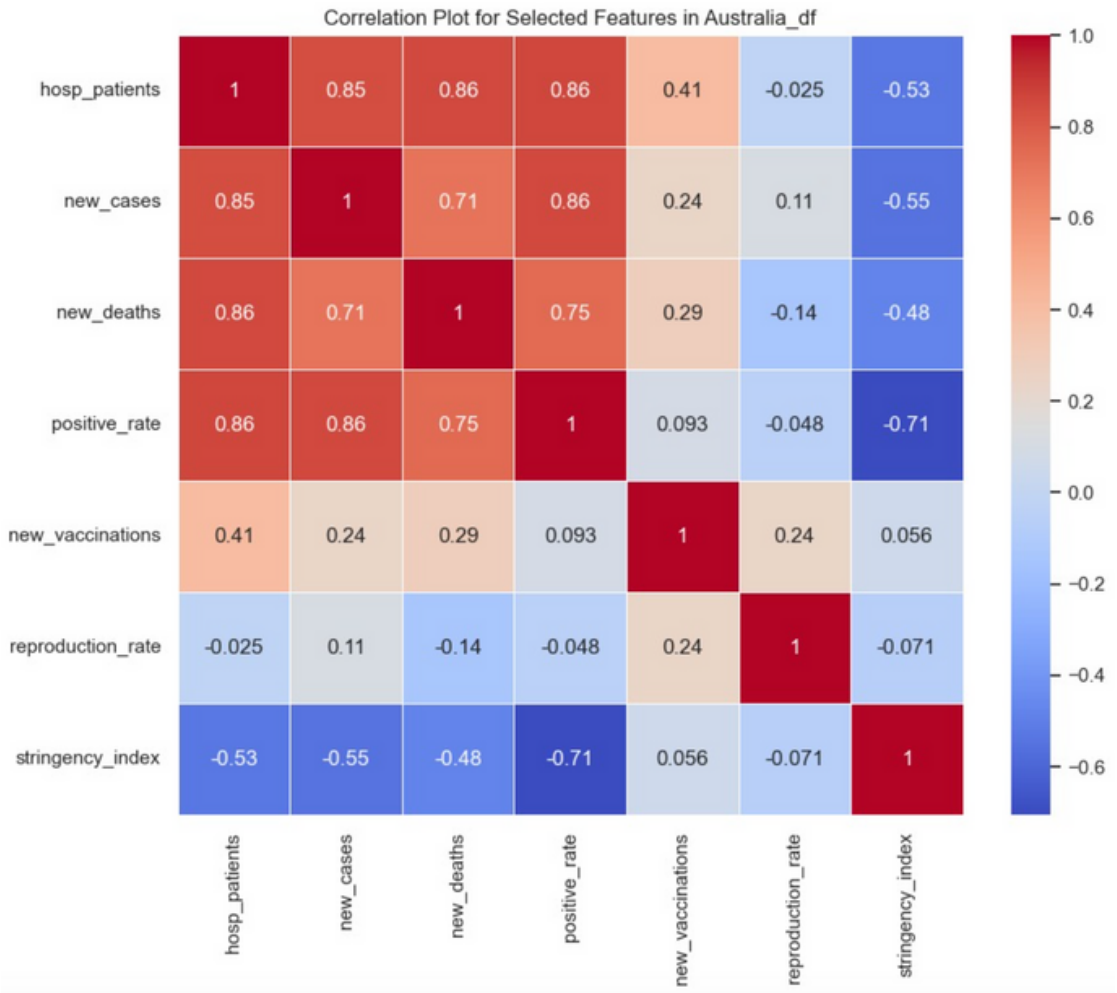


Fig 1 : Correlation Plot for Target Variable

We have conducted an in-depth feature correlation analysis to ensure the efficacy of our model. To visualize the relationships between the selected features, we used a Seaborn correlation plot. This method allowed us to carefully assess the relevance of each feature in predicting the hospital patient counts. The Seaborn correlation plot shows the strength and direction of these relationships, which helps us in the feature selection process.

The Seaborn correlation plot indicates that our target variable, hospital patient counts, has strong positive correlations with new cases, new deaths, and the positive rate. This means that an increase in these features leads to a rise in hospitalizations. On the other hand, there is a noticeable negative correlation with the stringency index. This indicates that a higher stringency index, representing more stringent pandemic control measures, is linked to a decrease in hospital patient counts.

The findings from this correlation analysis highlight the importance of the selected features in predicting hospital patient counts. It also provides valuable insights into the dynamics of the COVID-19 pandemic in Australia.

2. Model Type: Random Forest Regressor

We used a Random Forest Regressor to build our predictive model, optimizing its hyperparameters. The selected features were new cases, new deaths, stringency index, reproduction rate, positive rate, and new vaccinations, chosen for their correlations and domain relevance.

2.1 Model Training and Hyperparameter Tuning

We divided the data into two sets - training and testing, where 70% of the data was utilized for training, and the remaining 30% was used for testing. To ensure uniformity and improve the model's performance, we standardized the feature data. Our approach for hyperparameter tuning was GridSearchCV, which systematically explored different combinations of hyperparameters to find the best configuration. The grid of hyperparameters that we used included options for the number of estimators, maximum depth, minimum samples for splitting, and minimum samples per leaf.

The best hyperparameters for our Random Forest Regressor model were identified as follows:
'n_estimators': 200
'max_depth': 4
'min_samples_split': 2
'min_samples_leaf': 1

These hyperparameters were instrumental in achieving the model's optimal performance in predicting hospital patient counts.

2.2 Model Evaluation

After training and optimization, the model was evaluated using the testing dataset. To measure the accuracy of the model, Mean Squared Error (MSE) was computed. The model produced an MSE of approximately 28,132.75, which means the average squared difference between predicted and actual hospital patient counts.

Apart from MSE, we also obtained other evaluation metrics including R-squared (R2) score. The R2 score assesses the proportion of variance in the target variable that is predictable by the model. Our model achieved an R2 score of 0.9518, indicating that it accounts for a significant portion of the variation in hospital patient counts.

The mean residual value was also determined to be 27, which means that on average, the model's predictions deviated by 27 units from the actual hospital patient counts. It is crucial to note that the target variable, "hospital patients," exhibits a wide range from 0 to 5,500. Despite this range, our model demonstrates a high degree of accuracy as evidenced by the low MSE and high R2 score.

The figure below shows the distribution of residuals, which is slightly skewed towards the right, suggesting that our model may slightly be overfitting.

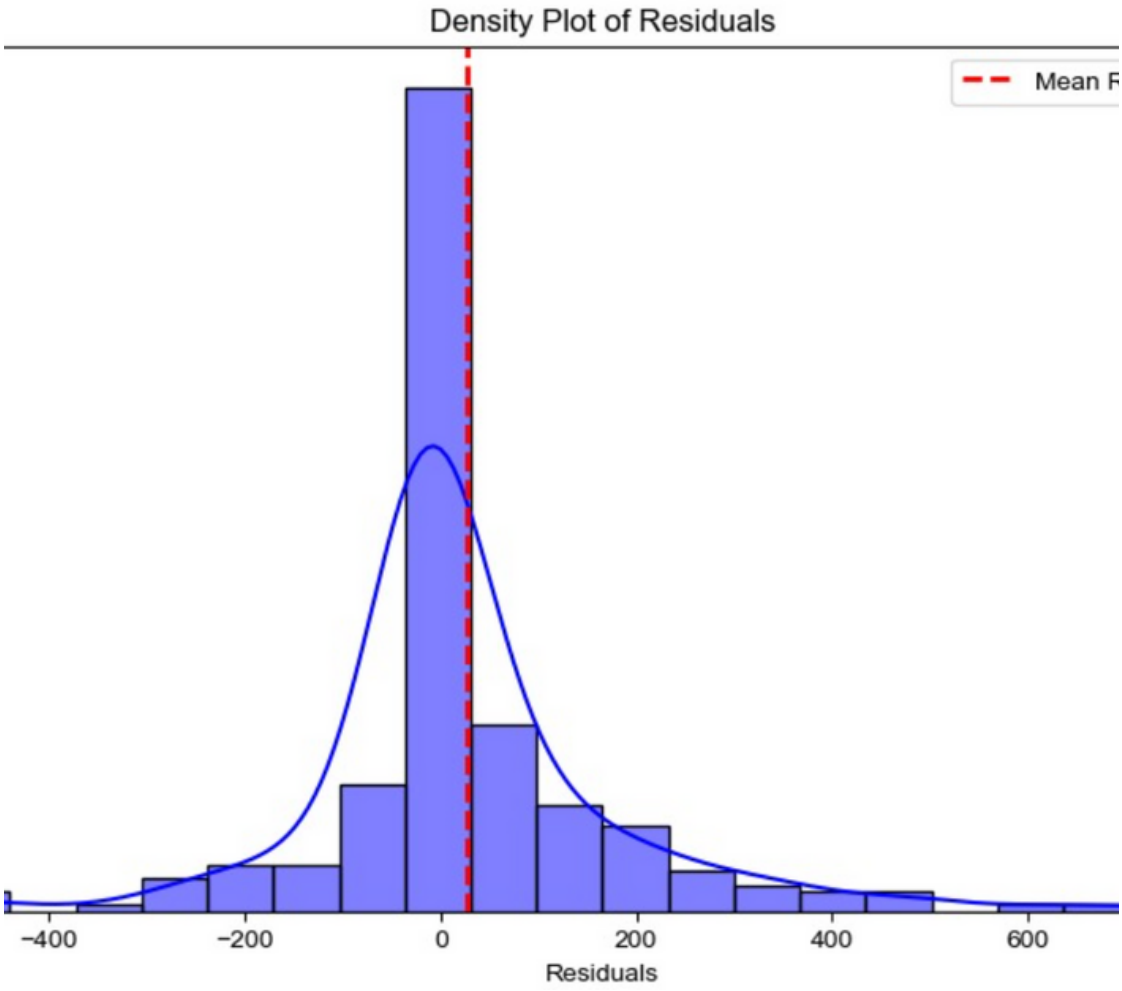


Fig 2 : Density Plot of Residual

The success of our Random Forest Regressor model in predicting hospital patient counts underscores its capability to provide valuable insights for healthcare planning and resource allocation.

3. Model Understanding Using SHAP

In this section, we delve into the world of SHAP (SHapley Additive exPlanations), a powerful tool that offers insights into how our predictive model arrives at specific predictions. SHAP provides a clear and interpretable framework for understanding the impact of each feature variable on individual predictions.

3.1 Understanding SHAP: A Brief Overview

SHAP values are a modern and comprehensive approach to explaining the output of machine learning models. They are based on cooperative game theory, which addresses the question: How should the "payout" be distributed among players in a game? In our context, the "players" are the features contributing to a model's prediction, and the "payout" represents the prediction for a specific instance.

3.2 SHAP Summary Plot

To initiate our journey of understanding our model, we first generated a SHAP summary plot. This visualization offers a global perspective of feature importance and impact across the entire dataset. Each dot on the plot represents a feature, and its position horizontally indicates whether the effect of that feature is associated with higher or lower predictions. The color intensity reveals the feature's value, with red indicating high values and blue representing low values.

The SHAP summary plot not only assists in identifying the most influential features but also offers a glimpse into the direction of their impact. By interpreting the SHAP summary plot, we can understand which features play a crucial role in predicting hospital patient counts and whether they have a positive or negative influence.

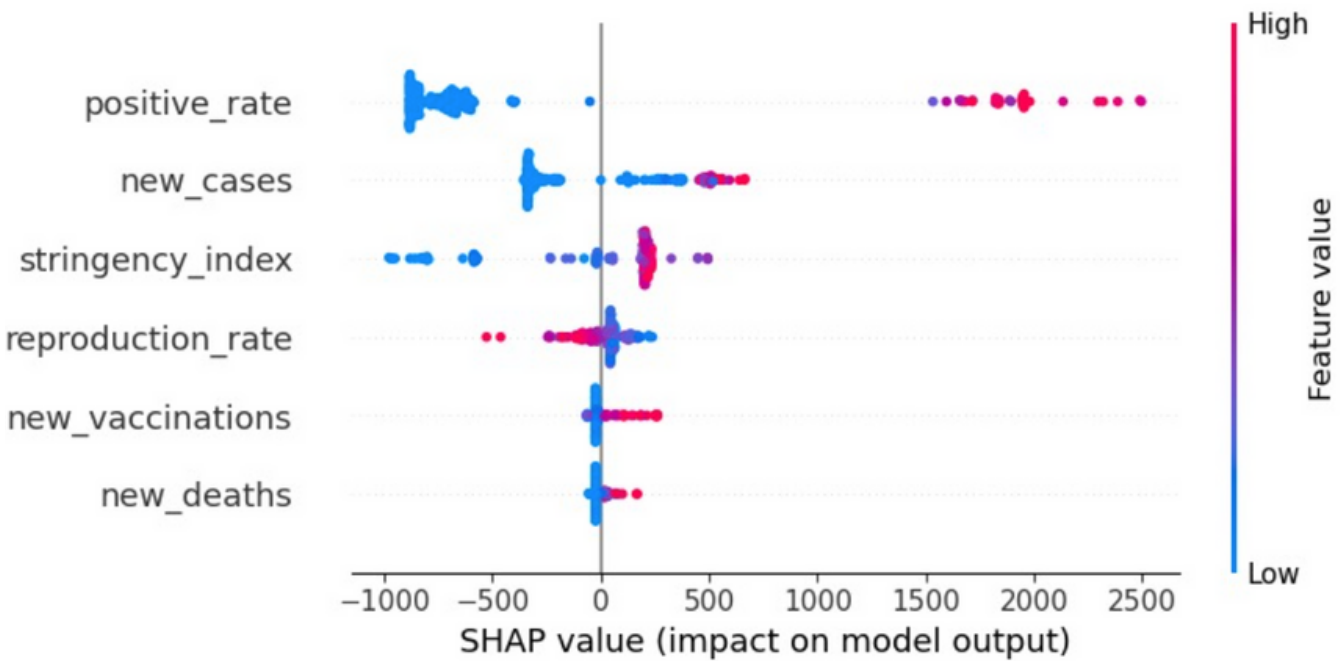


Fig 3 : SHAP summary plot

In Figure 3, the SHAP summary plot displays a summary of the feature effects. Features such as new cases and positive rate appear to exert a significant positive impact on the prediction, signifying that an increase in these feature values is associated with higher hospital patient counts. Conversely, stringency index exhibits a negative impact, indicating that stricter pandemic control measures are associated with lower hospital patient counts.

3.3 SHAP Force Plot

The SHAP summary plot offers a broad view, but for a deeper understanding of model behavior for specific instances, we turn to the SHAP force plot. A SHAP force plot dissects the prediction for a particular instance, unveiling how each feature contributes to that specific prediction. It effectively answers the question: “Why did the model predict this value for this instance?”

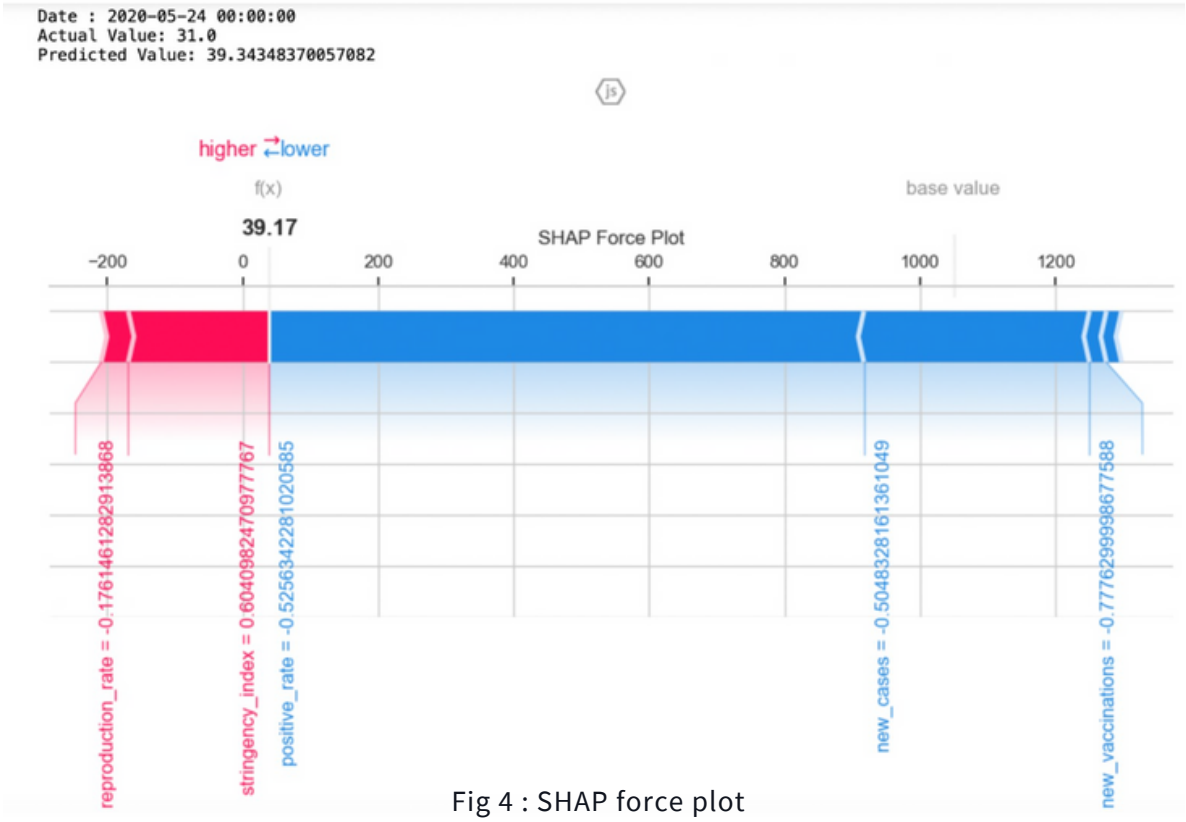


Fig 4 : SHAP force plot

In Figure 4, we present a SHAP force plot for 24th May 2020. The predicted hospital patient count for this instance is 39.13, while the actual count is 31. By analyzing this plot, we gain insights into how each feature behaves in producing this prediction:

Stringency Index: This feature has a magnitude of 0.60. A high stringency index tends to increase the predicted hospital patient count.

Reproduction Rate: high reproduction rate likely has a negative influence, as its magnitude is -0.17.

Positive Rate: This feature has a magnitude of -0.52, indicating a strong negative influence. A high positive rate is linked to a lower predicted hospital patient count.

New Vaccinations: New vaccinations have a magnitude of -0.77, suggesting that a higher number of new vaccinations corresponds to a lower predicted hospital patient count.

New Cases: New cases have a magnitude of -0.50. An increase in new cases is associated with a decrease in the predicted hospital patient count.

The Base Value in the SHAP force plot, which is 1060 in this instance, serves as the starting point from which the feature contributions are calculated. The combination of these features and their force respectively leads us to predict 39 patients. The feature in blue color pushes the prediction lower from baseline while features in red pushes the prediction higher.

The SHAP force plot provides a transparent window into the internal workings of our model, making it an invaluable tool for model interpretation and understanding.
By employing SHAP, we can unravel the intricate relationships between our chosen features and prediction outcomes, enhancing our ability to make informed and data-driven decisions in the context of healthcare and pandemic management.

CONCLUSION:

In summary, this comprehensive report delves into the intricate dynamics of the COVID-19 pandemic, focusing on two pivotal dimensions: hospitalization and policy analysis. It provides a compelling analysis of the different pandemic waves, demonstrating how Australia and Canada responded differently to these challenges. Australia, in the initial wave, implemented stringent policies that contributed to a well-controlled situation, while Canada faced a significant initial burden. As the pandemic evolved, the two nations adjusted their strategies, showing the flexibility required in pandemic management. Furthermore, the report introduces a predictive model for estimating hospital patient counts in Australia. This model equips healthcare systems and policymakers with valuable insights to efficiently allocate resources and plan for healthcare infrastructure. By understanding the interplay of various factors, including stringency index, new cases, and vaccination rates, the model offers a data-driven approach to proactively respond to the evolving healthcare needs. In a world marked by uncertainty, this report underscores the power of data and predictive modeling in guiding effective healthcare decisions and shaping future pandemic responses.

RECOMMENDATION:

To navigate the dynamic landscape of the COVID-19 pandemic, we recommend a flexible, data-driven approach to pandemic response. Governments must be ready to swiftly adjust policies based on the virus's severity and new variants. Effective communication and clear guidelines are vital. Employ predictive models to anticipate hospitalizations, proactively manage healthcare resources, and ensure sufficient ICU facilities, ventilators, and staff. Prioritize vaccination campaigns with clear public messaging. Collaboration between nations is essential, with a focus on data sharing and best practices. Continuously monitor the pandemic's progress, invest in research and development, and engage in scenario planning for unexpected surges. Embrace technology for monitoring and healthcare delivery. By following these recommendations, authorities can better protect their populations and enhance preparedness for future health crises.