TDS 3301
DATA MINING

A Love Story Between Covid-19 and Machine Learning

Prepared by

Sidharrth Nagappan, 1181102313, 0163132154
Tan Zhi Hang, 1191302698, 0182394458
Kan Eugene, 1191302380, 0182603233

Contents

1 Introduction **1**

2 Data Pre-Processing **1**
2.1 Conditional Feature Engineering . . . 1
2.2 Conditional Outlier Handling . . . 1

3 Exploratory Data Analysis **1**
3.1 Is there a correlation between the mean income of a state and the number of cases? . . . 1
3.2 Is there a correlation between vaccination and daily cases at a national level? . . . 2
3.3 Has vaccination helped reduced daily cases in Selangor, Sabah and Sarawak? . . . 2
3.4 If daily cases increases, does that also increase the number of people getting vaccinated on daily basis for Selangor, Sabah and Sarawak? Does the government put more effort into the vaccination campaign when cases spike? . . . 3
3.5 Which states have been most affected by Covid clusters? (e.g. Workplaces, Kluster Mahkamah, etc.) . . . 3
3.6 Which type of Covid-19 clusters are most prevalent? Are these clusters forming in workplaces, night markets, schools, etc.? . . . 4
3.7 How well is Malaysia's vaccination campaign doing compared to other countries in South-East Asia? . . . 4
3.8 Is there a correlation between individual casual contacts (contact tracing) and daily cases? . . . 4
3.9 How have vaccination numbers changed over time across states? . . . 5
3.10 How has the vaccination rate changed across the nation? . . . 5
3.11 Which vaccines have been used the most? . . . 5

4 Clustering **6**
4.1 How did clusters change over time with respect to cases and deaths? . . . 6
4.2 How did the clusters change over time with respect to cases, deaths and vaccinations? . . . 7
4.3 Which states require attention in terms of their vaccination campaign and deaths (relatively speaking)? . . . 7

5 Regression **7**
5.1 Can we predict the daily vaccination numbers? . . . 8
5.1.1 Boruta Feature Selection . . . 8
5.1.2 Multivariate Long Short Term Memory Time-Series Analysis . . . 8
5.1.3 Support Vector Regression . . . 8
5.1.4 Linear Regression . . . 8
5.1.5 Analysis of Different Models . . . 8
5.2 Does the current vaccination rate allow herd immunity to be achieved by 30 November 2021? . . . 9
5.2.1 Can we predict Covid-19 mortality numbers across the nation? . . . 9
5.2.2 Can we predict mortality numbers for Melaka, Negeri Sembilan, Perlis, Selangor and W.P. Putrajaya? . . . 9

6 Classification **9**
6.1 Can we classify individual location check-ins in Malaysia using a variety of variables into Low, Medium and High bins? . . . 9
6.2 Can we predict the type of vaccine based on the symptoms? . . . 10

7 Deployment and Conclusion **10**

1 Introduction

The intention of this project is to conduct a comprehensive analysis of Covid-19 in Malaysia and assess the performance of the nation across different metrics at the ASEAN, federal and state level. There are three key elements to this project: the Jupyter notebook, the interactive Streamlit application and this report.

Malaysia has been facing Covid-19 since March 2020, with a cumulative total of 2,460,809 cases as of 30 October 2021 [4]. Fluctuations in Covid cases, breakout clusters, by-elections, Covid-19 variants and the introduction of vaccines have led to interesting trends, that one can study to see how we really fared at the end of it all. Statistically speaking, the pandemic can be split into three waves: the first, second and third. The national immunisation program run by the Ministry of Health overlaps with the second and third.

Our study will focus on Covid-19 cases, vaccinations and deaths at different geographical levels, with appropriate exploratory analysis, visualisations and statistical modelling with both machine learning and neural time series models.

2 Data Pre-Processing

Datasets are sourced from the Ministry of Health's Github repository, the Covid Immunisation Taskforce open data [2], World Covid-19 dataset [3] and Department of Statistics Average Income dataset [1].

A crucial task is preprocessing all datasets used across the course of this study, although admittedly, the datasets released by the Ministry of Health and the Covid-19 Immunisation Taskforce are relatively clean. Key activities conducted here include dropping duplicate rows, filling missing values, converting the datatypes of columns as necessary and removing redundant columns that are irrelevant to our work to make dataframe handling easier in later stages. An example of the detection of null values is in Figure 1, using the isna().sum() method available in Pandas. Null or missing values are replaced with a 0, because the Ministry of Health itself has acknowledged that missing values are a result of a particular campaign having yet to start or lack of data for the day, instead of as a result of data entry errors. For example, the natural immunisation program started generating data only after January 2021. 0 is a safe assumption that does not alter the unpredictable Covid-19 trends.

2.1 Conditional Feature Engineering

While this was the general pipeline applied for most datasets, reading the metadata carefully allowed us to do further processing for the vaccine datasets, such as totalling up "pfizer1" and "pfizer2" to gauge the total number of Pfizer doses administered in the nation. Other than this, renaming lengthy column names to short and succinct ones made data manipulation faster, such as renaming the column "Mean Monthly Household Gross Income" to "income". Several cumulative totals of cases, vaccines, etc. were also done in cases where the original attribute was insufficient.

2.2 Conditional Outlier Handling

As for outliers, they are handled on a case by case basis. Most outliers related to Covid-19 statistics are inherently meaningful, so simply dropping them can be a dangerous practice. In cases where they obstruct the visualisation, such as in boxplots, they are purposely excluded. In other cases they are appropriately scaled using MinMaxScaler() and StandardScaler(). In proximity-based models such as K-Means clustering or Support Vector Machines, scaling helps normalise the distance between features on different scales.

3 Exploratory Data Analysis

Exploratory Data Analysis is an unstructured analysis of relationships, outliers and statistical distributions of the data. This section consists of several miscellaneous questions (with external datasets) and detailed questions that are used for the modelling phase later on.

3.1 Is there a correlation between the mean income of a state and the number of cases?

We wonder if there is a possibility that higher-income and more densely populated states are more susceptible to Covid-19 than smaller states. An additional dataset is brought in from the department of statistics Malaysia and the 2020 mean average household income is considered. Hence, this question is first conducted with (a)

```
date                        0
cases_new                   0
cases_import                0
cases_recovered             0
cases_active                0
cases_cluster               1
cases_pvax                  1
cases_fvax                  1
cases_child                 1
cases_adolescent            1
cases_adult                 1
cases_elderly               1
cluster_import            342
cluster_religious         342
cluster_community         342
cluster_highRisk          342
cluster_education         342
cluster_detentionCentre   342
cluster_workplace         342
dtype: int64
```

Figure 1: Analysis of Null Values in Cases

population as the third factor (do not account for population) and (b) scaling the cases to cases per 10000 people in the population (account for population). The findings of both are in Figure 2 and Figure 3.

	cases_new	income	pop
cases_new	1.000000	0.248051	0.931784
income	0.248051	1.000000	0.068793
pop	0.931784	0.068793	1.000000

Figure 2: Correlation table of cases new, income and pop

In Figure 3 plot 1, there appears to be a weak correlation. As the average income of states increase, the more populated it generally is, which would mean more cases. The population is a strong confounding variable, that can make or break the trend. Due to this suspicion, the second plot normalises cases based on state populations. With population accounted for in Figure 3 plot 2, the trend is broken, hinting that Covid-19 cases may not be a totally socio-economic one.

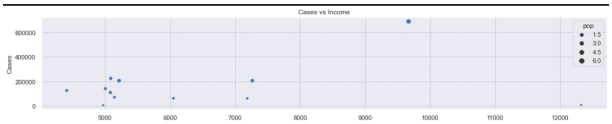
3.2 Is there a correlation between vaccination and daily cases at a national level?

A general regression plot, shown in Figure 4 subplot 4 between the cumulative vaccinations and the daily new cases indicate that the relationship isn't exactly linear; it is in fact **curvilinear**. It forms a parabolic trend towards the beginning and as the effects of vaccination kick in after the administration of around 1000000 doses, the number of daily cases are on a downward trend.

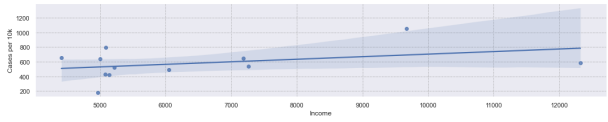
3.3 Has vaccination helped reduced daily cases in Selangor, Sabah and Sarawak?

The earlier question studied this question at a national level, but this question chooses to go state by state. The regression plots are displayed in subplots 1-3 of Figure 4. In Sabah and Selangor, the same **curvilinear** relationship between the variables can be observed. However, for Sarawak, there is an exponential increase in cases over time. We cannot conclude that vaccination has no effect here, because the curvilinear effect is attested to at both the national level and in Selangor and Sabah. We instead conclude that there are **confounding** variables at play in Sarawak.

Since the relationships are not linear, we do not proceed to conduct a Pearson's correlation test.



(a) Scatter plot of Income vs Cases with Population



(b) Scatter plot of Income vs Cases per 10k

Figure 3: Income and Cases Correlation Plots

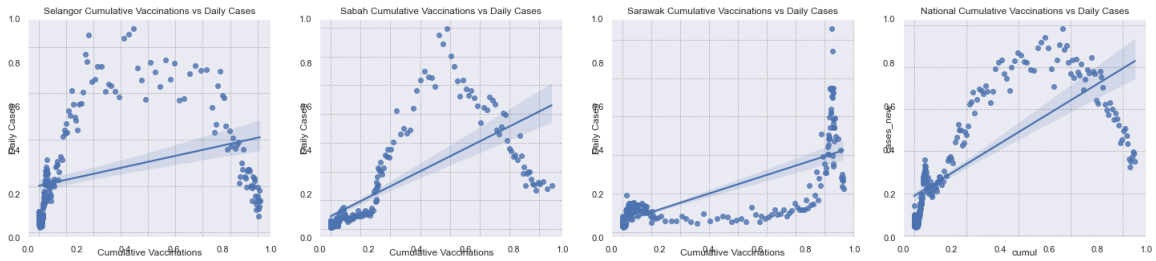


Figure 4: Selangor, Sabah and Sarawak Cumulative Vaccinations vs Cases

3.4 If daily cases increases, does that also increase the number of people getting vaccinated on daily basis for Selangor, Sabah and Sarawak? Does the government put more effort into the vaccination campaign when cases spike?

To reflect different periods of the vaccination campaign, we choose three thresholds: 5%, 10% and 15% state vaccination rates. After each of these milestones are reached, is the relationship between vaccination numbers and daily cases meaningful anymore?

	Full Period	When Vacinated Rate Over 0.05	When Vacinated Rate Over 0.10	When Vacinated Rate Over 0.15
Selangor	0.868888	0.149073	0.124465	0.130903
Sabah	0.887923	0.129022	0.148088	0.146401
Sarawak	-0.184678	-0.038726	-0.081525	-0.065913

Figure 5: Figure that show the correlation value for each Selangor, Sabah and Sarawak

Based on the Figure 9, we can see that the correlation between vaccination and daily cases changes drastically in different periods of the vaccination campaign, showing no noticeable pattern. Especially when the vaccination rate is over 5%, 10% and 15%, all 3 states also showed a weak correlation between vaccination and daily cases. Hence, we can conclude that the daily cases is not strongly correlated to the vaccination and the number of vaccines the government is administering is not dependent on the cases. **Whether cases go up or down, the government irrespectively continues with it's immunisation program**.

3.5 Which states have been most affected by Covid clusters? (e.g. Workplaces, Kluster Mahkamah, etc.)

Covid-19 clusters can lead to massive spikes. We set out to explore which states' cases are most affected by different Covid-19 clusters. Since some clusters span across states, we explode the dataframe based on the column with the list of states. For a cluster that spans across ['Selangor', 'Putrajaya' and 'Kuala Lumpur'], we create three separate rows.

From the original boxplot with outliers shown in Figure 6, we can see that the majority of Covid-19 clusters are moderately sized, often under 100 cases per cluster. However, there does exist unusually large clusters that appear as outliers. We remove these to show a new boxplot that visualises the distribution of the majority of Covid-19 clusters. Surprisingly, Negeri Sembilan and Perak are most affected by individual clusters, as opposed to more populated states like Selangor or Kuala Lumpur.

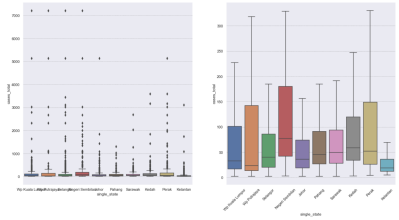


Figure 6: Boxplot of single state and total cases

3.6 Which type of Covid-19 clusters are most prevalent? Are these clusters forming in workplaces, night markets, schools, etc.?

To answer this question, Figure 7 shows the distribution of each cluster type. The detention centers generally have the largest Covid-19 clusters, with a strong right skew. The rest of the cluster categories consist of mostly small clusters.

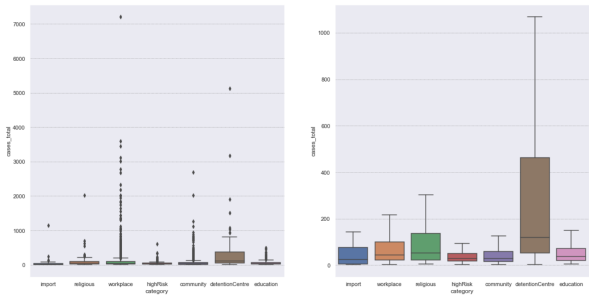


Figure 7: Boxplot showing the distribution of each cluster type, with and without outliers

An interesting example is workplace clusters. For instance, most companies/organisations in Malaysia are small, so there may be a **high frequency** of clusters, but only some are large enough to appear as an outlier. **We have heard of Top Glove in the news, but most organisations in Malaysia are not as large as Top Glove**.

3.7 How well is Malaysia's vaccination campaign doing compared to other countries in South-East Asia?

We use World Covid-19 datasets, with a focus on countries in South East Asia.

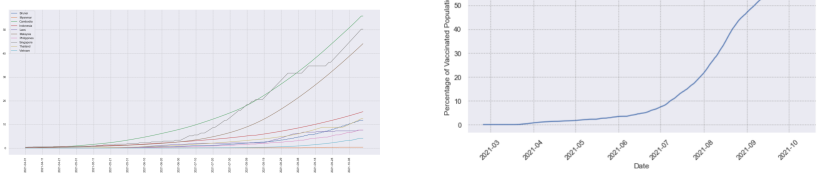
Figure 8 shows the vaccination rate for each country in South-East Asia. Based on the result, we can see that Cambodia has the highest vaccination rate compared to other countries. For Malaysia, we ranked top 3 in the graph and the vaccination rate is near 45%. Hence, we can conclude that Malaysia's vaccination campaign is doing better than the majority of South-East Asia's countries.

Percentages can be on the lower end because this dataset is a few weeks old, and when countries are vaccinating thousands a day, the original percentage can be higher.

3.8 Is there a correlation between individual casual contacts (contact tracing) and daily cases?

Correlating casual contacts for the day with the number of daily new cases. Since the relationship is linear (as shown in Figure 9, we proceed to calculate the Pearson's correlation between the two variables.

Based on our result, we can see that daily new cases are highly correlated with the number of casual contacts for the day. As in, if more people go out and come in close contact with infected people, the number of cases increases. The statistics fall in line with the science behind Covid-19.



(a) Vaccination Rates for South East Asia

(b) Vaccination Rate for Malaysia

Figure 8: 2 Figures side by side

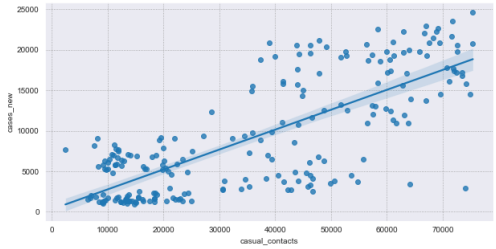


Figure 9: Casual contacts vs Daily cases

3.9 How have vaccination numbers changed over time across states?

We first normalize both the daily fully vaccinated number and the daily new cases of covid cases to examine the effect of the vaccination upon the number of daily cases. We look 5 random states(Melaka, Negeri Sembilan, Perlis, Selangor and W.P. Putrajaya) so that we can examine the effectiveness of vaccination without any bias.

When we refer to our results, it is clear that when the government puts more effort into getting the people vaccinated in that certain state, the daily cases would start to decline, except for Perlis. Perlis's daily cases spikes up a little in October 2021 but not drastically. We can safely assume that vaccination actually in a way help controlling the cluster cases in Perlis without going higher. Regardless, we can conclude that vaccination might be one of the contributing factor to reduce daily covid cases in Malaysia. However, more research is needed in order to conclude this finding. The figure is too large, please refer in either the notebook or Streamlit.

3.10 How has the vaccination rate changed across the nation?

A line graph is used to plot the cumulative vaccination percentage through time. The result is shown in Figure 8.

We have just crossed the 60% threshold in terms of vaccinations. It is observable that there is a slow start, but the speed of the campaign has admittedly picked up.

3.11 Which vaccines have been used the most?

Primarily due to the advent of social media, there has been a lot of controversy about the different vaccine brands administered. Malaysia is one of few countries that chooses to buy a variety of vaccines and even switch from Sinovac to Pfizer in the middle of the campaign, so it is interesting to see which vaccines were used to inoculate the majority of the population.

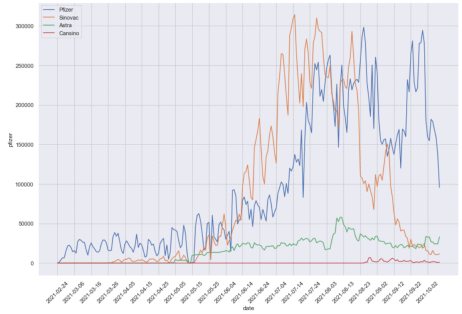
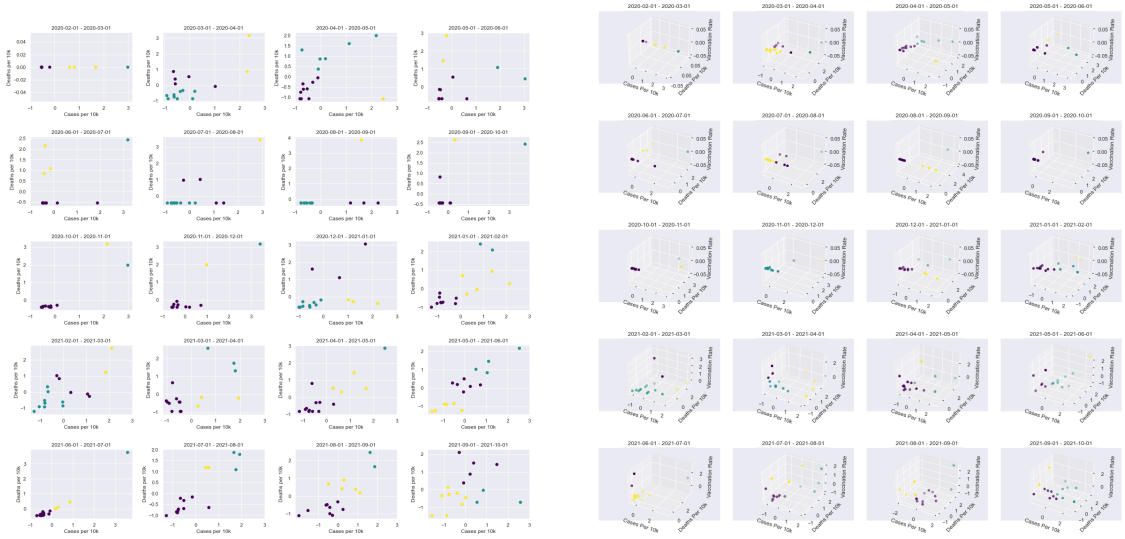


Figure 10: Line Graph of the number of vaccines that are being used daily

At the start of the campaign, MoH was primarily using Pfizer, but by June, Sinovac would rapidly overtake Pfizer to become the most used vaccine. Adoption of Sinovac however drops, but people still do come back for their second dose.

Pfizer is the most used vaccine in Malaysia, followed by Sinovac and then Astrazeneca. If you observe the usage of Astrazeneca, it flails in comparison to the other two because it was opened up for voluntary registrations. Furthermore, unlike Pfizer and Sinovac, Astrazeneca usage does not show an upward trend and only has a few spikes, which may be linked to the times the government opens up registrations.



(a) Cases per 10k vs Deaths per 10k

(b) Cases per 10k vs Deaths per 10k vs Vaccination Rate

Figure 11: 2-dimensional and 3-dimensional Temporal K-Means Clustering

4 Clustering

4.1 How did clusters change over time with respect to cases and deaths?

Some states reorganise into new clusters over time. Different states are of different sizes, so one has to consider the size of each state when clustering. Instead of taking the number as a whole, we find the cases per 10000 people in each state to account for this additional variable. We do this because we observed in the exploratory phase just how impactful population is as a confounding factor in any Covid-19 relationship.

Two clustering techniques are used here, namely temporal K-Means and temporal Density based clustering (DBSCAN). K-Means allows us to specify the number of clusters, but DBSCAN naturally forms clusters through density. K-Means produces clusters of mainly spherical shapes, while DBSCAN is more inconsistent in terms of cluster shapes. By visualising the clusters every month, we observe how the pattern changes with respect to time.

We can observe that in the beginning, the majority of the clusters were positioned towards the bottom left and they maintain a similar pattern until about August 2020. In August, the cases were still high but there were fewer deaths, which may signify that the situation was improving, besides the one state that is in the upper corner of the plot that stands out from the rest. Around December, the bottom-right cluster begins to break up and states start moving diagonally upwards in the graph, meaning higher number of deaths and more cases. By September 2021, the states fall in a sort of straight diagonal line, with the performance of states spread across the spectrum from mild to serious.

By October 2021, K-Means indicates the formation of 3 main clusters:

1. Low Cases and Low Deaths
2. High Cases and High Deaths
3. High Cases and Low Deaths

If there are low deaths despite high cases, then that is acceptable because vaccines have been known to reduce the seriousness of cases. However, the states in "high cases and high deaths" need to be looked at:

1. Johor
2. Kedah
3. Pulau Pinang
4. Sabah
5. Selangor

4.2 How did the clusters change over time with respect to cases, deaths and vaccinations?

For this question, we bring in a third variable: vaccinations. Temporal 3-dimensional K-Means clustering is once again used in this question.

We maintain that states can be grouped into 3 clusters. Throughout 2020, vaccinations are 0 and hence, the clusters slowly start to expand in terms of cases and deaths. By 2021, the states have spread reasonably wide throughout the 3 dimensions and cases and deaths are at an all time high.

Around this time, vaccination begins and clusters start moving higher in the "vaccination" dimension. As most states move vertically upward in the dimension, there are still a cluster of states with low vaccination rates.

Based on Figure 11, we can see that by October 2021, most states are on the upper end of the vaccination spectrum, the lower end of the deaths spectrum, but cases are still spread wide across. This may not be too consequential, because the fact that an increase in cases did not lead to an increase deaths shows the effect of the vaccination campaign.

4.3 Which states require attention in terms of their vaccination campaign and deaths (relatively speaking)?

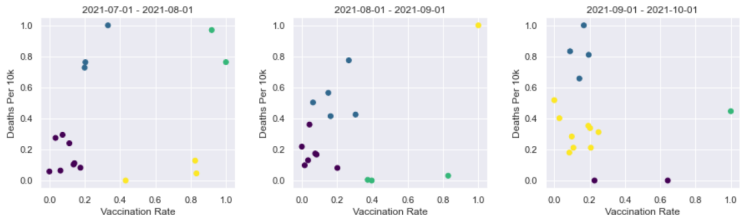


Figure 12: Scatter plot of Deaths Per 10k vs Vaccination Rate

This is an important question to answer because the states with low vaccinations and high deaths are of particular attention. Another interesting point to note that is that deaths are used here instead of cases, because Covid-19 has been known to reduce the intensity of cases rather than prevent them entirely. As in, vaccinated people can still get Covid-19 but may not go as far as die from the infection. Once again, we find the deaths per 10k people in the state to normalise for population differences in the states. We total up the deaths for every month and cluster it with the cumulative vaccinations. All values are scaled using the Standard Scaler. Then, we temporally visualise the movement of clusters through time. In Streamlit, this is done through an interactive visualisation, and for the notebook, we plot a grid of subplots. Only the last 3 months are shown here. For the full visualisation, please refer to the notebook or Streamlit.

We can see that throughout 2020, there are 0 vaccinations in all state since the vaccination campaign was yet to start. By February 2021, two states have begun their vaccination campaigns. It speeds up more rapidly by March and April, the vertical clusters start to spread out on the x-axis indicating higher vaccination numbers. In June 2021, there was a remarkable shoot where the y-axis scale completely changed. As of September 2021, the states that may require attention are those with low vaccination rates and high death rates (still very small), namely cluster 1, which contains the following states:

1. Johor
2. Kedah
3. Pulau Pinang
4. Selangor

5 Regression

Across regression questions, appropriate scaling is done before passing into the models.

5.1 Can we predict the daily vaccination numbers?

We answer this question using 3 steps, namely Boruta Feature Selection, Multivariate LSTM, Support Vector Regression and Linear Regression:

5.1.1 Boruta Feature Selection

We originally take features from the cases, tests and vaccination datasets and pass it into iterative Boruta Feature Selection to get an optimal feature set.

5.1.2 Multivariate Long Short Term Memory Time-Series Analysis

Since vaccination numbers change daily, it can be treated as a time series problem. A window of the past 100 days is chosen to consider past vaccination numbers and gauge a trend of how the values are changing.

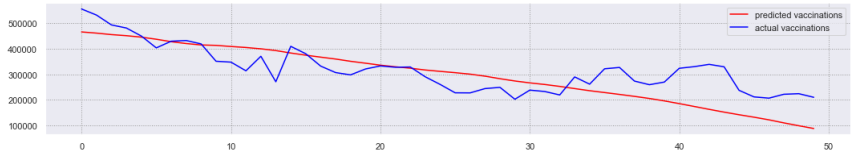


Figure 13: Vaccination Predictions from LSTM

Hyperparameter tuning is done by changing the number of layers and the number of units in each LSTM layer. The best model consists of LSTM(50), Dropout(0.2), LSTM(50), Dropout(0.2) and Dense(1) with a sigmoid output. We initially used a very deep network, but the dataset is very small, so a smaller network tended to work better in terms of the mean-squared-error performance. Early Stopping and adaptive learning rates are applied.

The best Tensorflow LSTM model returned a mean squared error of 0.017149. The difference between the predictions and the actual values are plotted in Figure 13. The model is then saved and imported into Streamlit for visualisation.

5.1.3 Support Vector Regression

Support Vector Regression does not treat this as a time-series problem, but instead as a complex relationship between the chosen features and the day's vaccination number. The best model returned a mean squared error of 0.00657. Difference between prediction and actual is displayed in Figure 14.

5.1.4 Linear Regression

Linear Regression is the simplest model for this suite. It simply plots a relationship between the features and the vaccination number. The best model returned a mean squared error of 0.0231. Difference between prediction and actual is displayed in Figure 14.

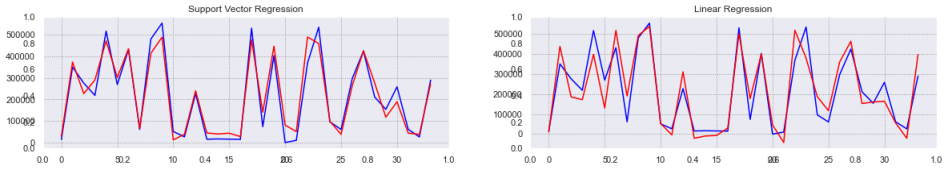


Figure 14: Vaccination Predictions from Support Vector Regression and Linear Regression

5.1.5 Analysis of Different Models

The unavailing performance of the LSTM models can be attributed to a small dataset that is insufficient to train a deep learning model. Support Vector Regression with appropriately tuned hyperparameters performs best in this case.

5.2 Does the current vaccination rate allow herd immunity to be achieved by 30 November 2021?

Considering that herd immunity means 80% of the population, we try to predict whether this threshold can be reached by the end of November. We originally planned to use the earlier question, but since it was multivariate, we do not have the other values for the future. We are therefore forced to use univariate methods and in this case, we choose ARIMA (Auto-Regressive Integrated Moving Average) to predict the future.

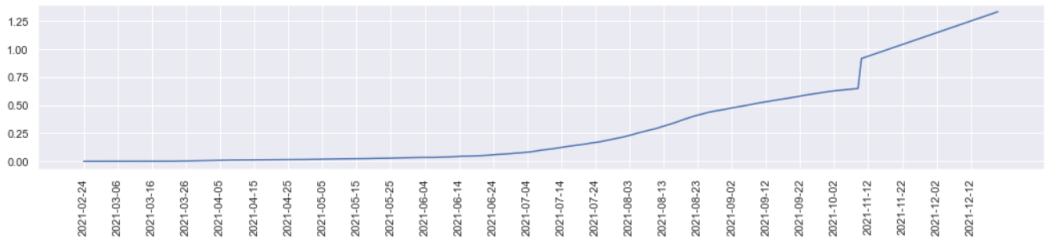


Figure 15: Line Graph of Predicting Malaysia Vaccination Rate

The best hyperparameters for the ARIMA model are choosen using the autoarima package. Once the model is tuned, we extrapolate the original graph using the values predicted by ARIMA. Based on the Figure 8, it is possible that herd immunity will be reached before 30 November if it continues at this rate.

5.2.1 Can we predict Covid-19 mortality numbers across the nation?

With vaccinations on the rise, the impact of case numbers on deaths are changing. Therefore, we try to predict mortality numbers based on the existing variables. Recursive Feature Elimination is first done to choose an appropriate feature set.

Both the mean squared error (MSE) and mean absolute error (MAE) are employed with a best performance of 0.057 and 0.0123 MSE and MAE respectively using the Decision Tree Regressor.

5.2.2 Can we predict mortality numbers for Melaka, Negeri Sembilan, Perlis, Selangor and W.P. Putrajaya?

Recursive Feature Elimination is also used here to identify the best features for each state. The Decision Tree Regressor is applied across the 5 states with performance reported in Table 1.

State	Mean Absolute Error	Mean Squared Error
Melaka	0.1051	0.0167
Negeri Sembilan	0.1333	0.0258
Perlis	0.15	0.0792
Selangor	0.0548	0.0067
W.P. Putrajaya	0.0333	0.0167

Table 1: Results of Each States

6 Classification

6.1 Can we classify individual location check-ins in Malaysia using a variety of variables into Low, Medium and High bins?

Feature selection is done using Boruta, data balancing is done using SMOTE, then the Random Forest Classifier, Logistic Regression and the Naive Bayes Classifier are evaluated. The classification report of the correctly and wrongly classified numbers are depicted in Table 2.

We find that the Random Forest Classifier performs best.

Model	Accuracy	F1-Score
Random Forest Classifier	0.93103	0.93159
Logistic Regression	0.77011	0.76823
Naive Bayes Classifier	0.60919	0.59438

Table 2: Results of Each Models

6.2 Can we predict the type of vaccine based on the symptoms?

Certain vaccines show more of a certain type of symptom than others and this information is reported by vaccinated individuals through the MySejahtera app. We set out to predict the vaccine used based on the reported symptoms.

Feature selection is done using Recursive Feature Elimination to identify the most telling symptoms. Then the dataset is SMOTEd to create additional synthetic data and balance the originally unbalanced dataset. We choose to only apply SMOTE on the training set so as to not distort the real-world distribution of Covid-19 symptoms. Classification models used include Logistic Regression and the Support Vector Classifier.

Hyperparameter tuning is done using GridSearchCV to identify the best configuration for Support Vector Machines, for example. The best parameters for the Support Vector Classifier is GridSearchCV(estimator=SVC(), param_grid='C': [0.1, 1, 10, 100, 1000], 'gamma': [1, 0.1, 0.01, 0.001, 0.0001], 'kernel': ['rbf'], verbose=3). Accuracy and weighted averaged F1-Score are used for evaluation. Since accuracy is not immune to unbalanced datasets, it is supplemented with the weighted F1-score that takes into account the occurrence frequency of the different classes in the dataset.

The performance of each model is shown in the Table 2.

Model	Accuracy	F1-Score
Logistic Regression	0.6923	0.7486
Support Vector Classification	0.92307	0.92896

Table 3: Results of Each Model

7 Deployment and Conclusion

The results of the exploration are visualised on Streamlit through an interactive dashboard with features such as being able to visualise the movement of clusters through time. Seaborn visualisations are converted into Plotly to add interactivity. Several time-consuming processes such as feature selection are omitted from the report. The application has been deployed and is available here.

Across this project, we studied the relationship between Covid-19 and various variables, revealing both expected and unexpected relationships. Generally, the impact of vaccination on the country is curvilinear and prediction of daily vaccination numbers is not a straightforward factor (as there is a lot of external influence from the government). Some vaccines exhibit more of a certain type of symptom than others and this is evident in the performance of the classification models in Table 3. Currently, we looked at the social impacts of Covid, but the economic ones are missing from this project; perhaps future work can study the relationship between Covid-19 and industrial variables, especially now that the immunisation program is rampant.

References

Malaysia, J. P. (2020). *Mean monthly household gross income by state, malaysia.* Retrieved 2020-07-29, from https://www.data.gov.my/data/en_US/dataset/mean-monthly-household-gross-income-by-state-malaysia

MoH-Malaysia. (2020). *Open data on covid-19 in malaysia.* Retrieved 2020-02-24, from https://github.com/MoH-Malaysia/covid19-public#readme

owidbot. (2020). *Data on covid-19 (coronavirus) by our world in data.* Retrieved 2020-02-24, from https://github.com/owid/covid-19-data/tree/master/public/data

worldometers. (2020). *Coronavirus cases in malaysia.* Retrieved 2020-02-15, from https://www.worldometers.info/coronavirus/country/malaysia/