TDS 3301

DATA MINING

GROUP PROJECT

Malaysia COVID-19 Cases and Vaccination

Prepared by

Koh Han Yi, 1181302907, 018-3758138
Lee Min Xuan, 1181302793, 012-2329158
Tan Jia Qi, 1191301879, 011-11446369

In this project we worked on Question 2: Malaysia COVID-19 Cases and Vaccination. This project focused on discovering meaningful insights for Malaysia COVID-19 situation. Throughout the whole project, we used data mining techniques such as feature selection, clustering, classification and regression models. The results of this project were deployed on Streamlit. We are using the datasets listed below for this project:

- Open data on COVID-19 in Malaysia (*GitHub - MoH-Malaysia/covid19-public: Official data on the COVID-19 epidemic in Malaysia. Powered by CPRC, CPRC Hospital System, MKAK, and MySejahtera.*, 2021)
  - Cases and Testing
  - Healthcare
  - Deaths
  - Vaccination Adverse Event Following Immunisation (AEFI)
- Open data on Malaysia's National Covid-19 Immunisation Programme (*GitHub - CITF-Malaysia/citf-public: Official data on Malaysia's National Covid-19 Immunisation Programme (PICK). Powered by MySejahtera.*, 2021)
  - Registration
  - Vaccination
- R-Naught Value in Malaysia (*Nilai R Malaysia*, 2021)
- Google Search Trends for COVID-19 Related Keywords (*Google Trends*, n.d.)

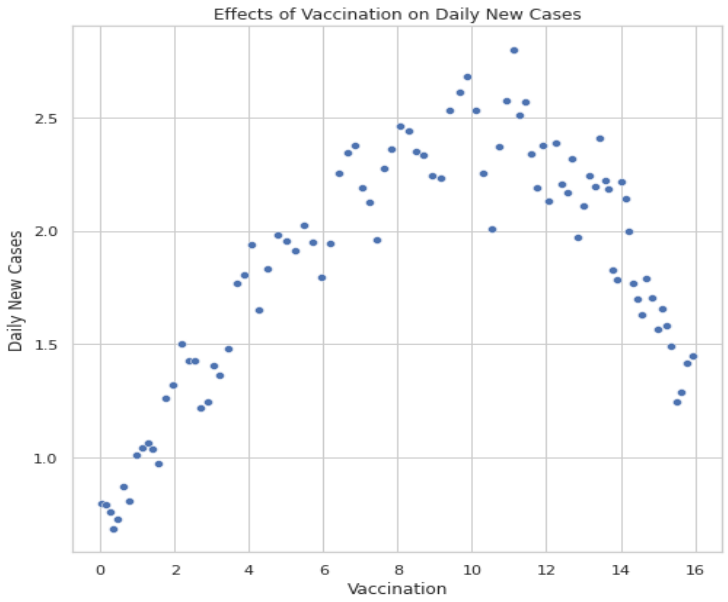
The datasets used are available and accessible on Github, Google Trends and Ministry of Health (MOH) COVID-19 websites, whereas the R-naught value data in csv format were collected manually through the information provided online. Different sets of data were used for each question formed in the project according to the appropriateness and relevance, and each question had a varied time frame.

1 Exploratory Data Analysis (EDA) and Data Pre-Processing

1.1 The Relationships between Covid-19 Vaccination and Daily New Cases

In the first question, we are curious about the relationships between the Covid-19 vaccination dataset and Covid-19 daily new cases dataset within July until September for each state in Malaysia. Before that, we look at the bigger picture of using whole Malaysia's data. We get to know the effect of vaccination on the daily new cases. The plot below has shown that the vaccination has been

helped reduce the daily new cases significantly in Malaysia. The scatter plot has a strong correlation, but it is not linear. The daily new cases increase, and the vaccination rate increases. After a certain point, the vaccination has played an important role to decrease the daily new cases.



In our findings, the states daily new cases that have been affected a lot by vaccination are Johor, Kedah, Pulau Penang, Sabah, Selangor and W.P. Kuala Lumpur. The states mentioned have a strong correlation between vaccination and daily new cases from July until September. This is probably related to the population density and the r-naught value of the states mentioned above.

1.2 Covid-19 Daily New Cases and Daily New Deaths for Each State

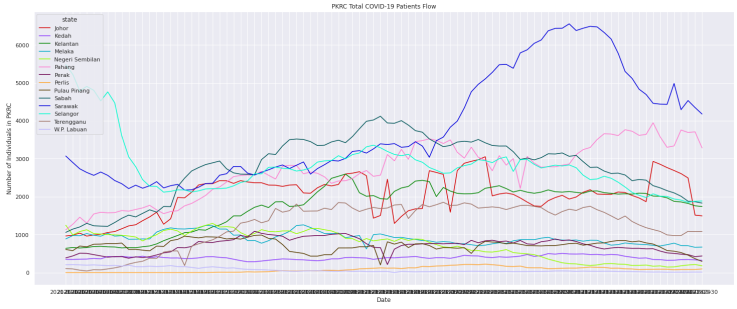
In this section, we are interested in finding which states have high daily cases and high deaths. The table below shows the average of daily new Covid-19 cases and daily new deaths due to Covid-19, which is higher than the median new cases and deaths in Malaysia. The result states are needed and advised to pay more attention by increasing the vaccination rate or reinforcing the Movement Control Order (MCO). The state residents also need to raise awareness of the current Covid-19 pandemic situation and infectious disease.

State	Deaths_cases	Cases_new
Johor	0.082657	0.154564
Kedah	0.049292	0.135457
Perak	0.022102	0.093466
Pulau Pinang	0.037726	0.118331
Sabah	0.050745	0.158899
Selangor	0.235800	0.530263
W.P. Kuala Lumpur	0.059041	0.133993

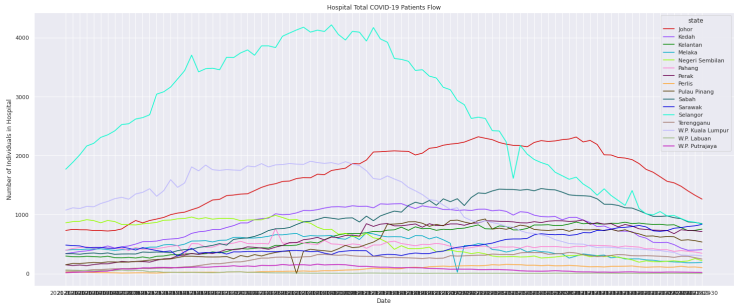
Table 1: States with High Daily New Cases and Deaths

1.3 The Admission and Discharge flow in PKRC, hospital, ICU of Each State

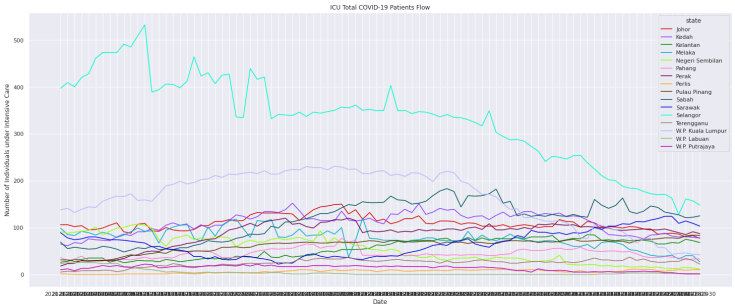
We used three datasets to know the flows of COVID-19 patients in PKRC (COVID-19 Quarantine and Treatment Centre), hospital, and icu. The *PKRC* data is only available for 14 states: Johor, Kedah, Kelantan, Melaka, Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, W.P. Labuan, *hospital* and *icu* data are available for 16 states with W.P. Kuala Lumpur and W.P. Labuan along with the other 14 states mentioned. We focused on the confirmed COVID-19 patients data only.



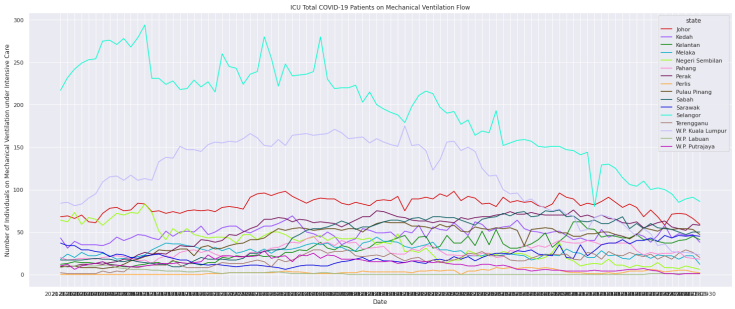
(a) PKRC Total COVID-19 Patients Flow.



(b) Hospital Total COVID-19 Patients Flow.



(a) ICU Total COVID-19 Patients Flow.



(b) ICU Total COVID-19 Patients on Mechanical Ventilation Flow.

According to the graphs above, the states that require more attention are Selangor,Johor, Sabah, W.P. Kuala Lumpus, Sarawak, and Pahang. Although the reasons behind them having more patients may be because they have more population, it is still obvious that they need more attention from government to put in efforts and works to improve the situation.

1.4 The Trend for Vaccinated and Cumulative Vaccination Registration for Each State

We used state registration dataset to understand the registration rate for each state in Malaysia. The dataset is only available for 16 states: Johor, Kedah, Kelantan, Melaka,Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, W.P. Labuan, W.P. Kuala Lumpur and W.P. Putrajaya.

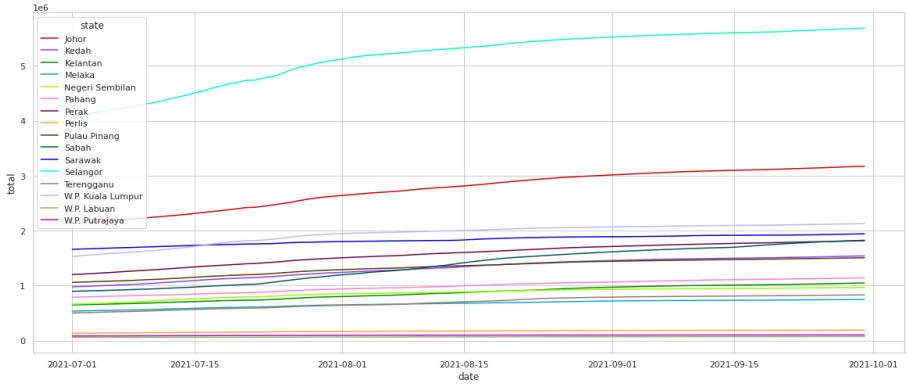


Figure 3: registration rate line plot.

Based on the findings, it shows that most of the citizens who registered for vaccination is from Selangor. Its population who registered for vaccination has been the highest throughout all times.

1.5 The Trend of R-Naught Index Value for Each State.

In this section, we are analyzing the r-naught index to understand its relationship with the Covid-19 spread. To obtain insights from the R-Naught Dataset, we have self-collected the data from the government website and manually key in the data into one csv file. Based on the plot, we find out that most of the Malaysia r-naught were having a spike in August 2021, that is probably because of the intrusion of the new variant of virus. However, the r-naught index keep decreasing after August. The decrease of r-naught index might be caused by the high vaccination rate of our population.

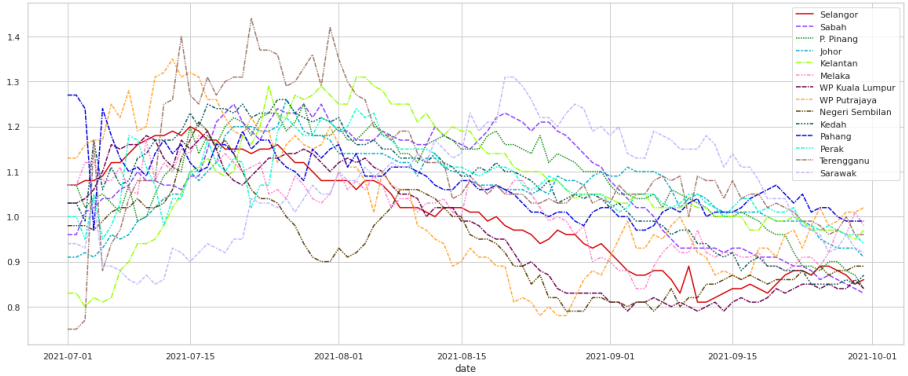


Figure 4: registration rate line plot.

1.6 The Interest in COVID-19 keywords of Each State from Google Trends Data.

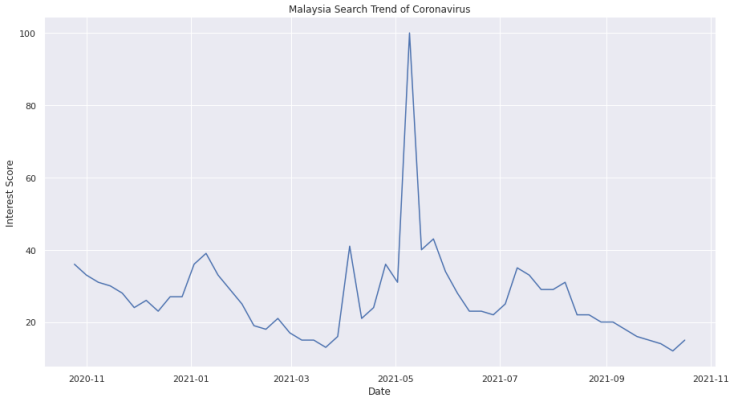


Figure 5: Malaysia Search Trend of Coronavirus.

State	Total Cases	Total Interest Score
Selangor	697659	561
Sarawak	232205	418
Johor	211098	477
Sabah	203170	439
W.P. Kuala Lumpur	187723	502
Kedah	142184	333
Pulau Pinang	140138	600
Kelantan	132427	276
Perak	112042	426
Negeri Sembilan	99543	425
Pahang	72942	318
Terengganu	65532	270
Melaka	62468	400
W.P. Labuan	9850	176
W.P. Putrajaya	6572	321
Perlis	4950	183

Table 2: Total Cases and Total Interest Score of States

In short, this section demonstrated basic exploration and analyzation on google trends data in Malaysia. The search trend of "coronavirus" and other keywords fluctuated over the one year span (19th Oct 2020 - 19th Oct 2021) due to various reasons. It is noticeable that interest score for "coronavirus" topped around the May of 2021 because of the pandemic situation started to worsen. The correlation between COVID-19 cases and people's interest towards COVID-19 is moderate positive. People in different states have different interest levels for different keywords, whichever is more related to themselves will be searched more. For example, Sabah topped the "Cansino" interest score while other states showed lower interest, because Cansino vaccines are mostly available and vaccinated in Sabah. This section demonstrated basic exploration and analyzation on google trends data.

2 Techniques and Models

2.1 Clustering Technique

2.1.1 Pattern of Different Vaccines Doses

We used K-means clustering to figure out the hidden pattern, because it suitable for general-purpose, not too many clusters, inductive use cases. The four types of vaccines with different total doses over the 3 months are classified into two groups. One group is the group with higher total doses, another group is the lower total doses group. We can observe that there are some vaccines having lower total doses compared to other vaccines, although it is already September 2021. This could be indicating a supply and demand problem for specific type of vaccine. The reasons behind it could be the differences in the vaccines' popularity and publicity effect. For example, there were more people talking good about Pfizer and the government recently used Pfizer the most for vaccination whereas AstraZeneca was said to be more likely to cause side effects and blood clots, so less people wanted to take it.

2.1.2 Pattern between Cummulative Vaccine Doses and Daily Cases

Affinity propagation algorithm is used because it is suitable for many clusters, uneven cluster size, inductive, and doesn't need to input clusters.By applying affinity propagation algorithm on the cummulative vaccine doses and daily cases in Malaysia over the three months, the estimated number of clusters is 6. With the graph shown above is a concave shape of points plotted, we can induced that the six clusters are: very low cummulative vaccine doses & very low daily cases (Purple), low cummulative vaccine doses & low daily cases (Blue, light blue), moderate cummulative vaccine doses & moderate daily cases (Aqua blue), high cummulative vaccine doses & very high daily cases (Green), high cummulative vaccine doses & moderate daily cases (Yellow, orange), very high cummulative vaccine doses & low daily cases (Red). The first cluster indicated a pandemic situation where the daily cases remained low and steady although the cummulative vaccine doses is not high. The second, third and fourth cluster indicated the climbing trend of both cummulative vaccine doses and daily cases, which means the vaccines effects have not shown yet and there might be outbreaks of COVID-19 happening at that time. The fifth cluster indicated the start of the daily cases decline trend where the vaccine can be said to be effective and have suppressed the pandemic a bit. The sixth cluster indicated the drastic drop of daily cases from its peak as the cummulative vaccine doses is very high. In short, the vaccines are effective but require sometime to show its affects. Hence, we should continue the vaccinations.

(a) Daily Total Vaccine Doses in Malaysia

(b) Cummulative Vaccine Doses and Daily Cases

2.2 Classification Model

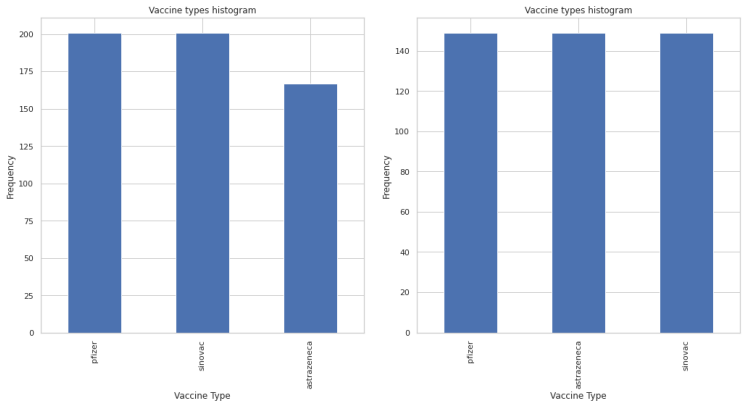
We have built two classification models to classify vaccine type based on side-effects and R-naught level based on R-naught values in Malaysia.

2.2.1 Classify Vaccine Type

The datasets used to classify vaccine type is *aefi.csv* under the MOH vaccination directory. The dataset has recorded vaccine type and its side effects. We wanted to build a model that can classify the vaccine type based on their side effects.

Feature Selection The feature selection technique used in classifying vaccine type is Recursive Feature Elimination (RFE). It is a wrapper method but internally running filtering methods. After trying different features, we have decided to select 15 important features from 30 features using RFE to increase the classifier's performance.
Feature Selected : d1_site_pain , d1_site_swelling, d1_joint_pain, d1_weakness, d1_fever, d1_vomiting, d1_rash, d2_site_pain, d2_site_swelling, d2_site_redness, d2_tiredness, d2_headache, d2_weakness, d2_fever.

Synthetic Minority Over-sampling Technique (SMOTE) Before oversampling the data, we first check on the data value counts to see whether it is balanced or imbalanced. Below is the data distribution for all three vaccine types: Sinovac, Pfizer, and Astrazeneca before oversampling technique using SMOTE and after balancing the dataset. In this stage, the data has split into train and test set in a ratio of 7:3 and is ready for modelling the classifier and making a prediction.



(a) Before SMOTE (b) After SMOTE

Modelling and Evaluation on Vaccine Type Classifier The classifier chosen is the logistic regression classifier. Using Logistic regression, the model can classify the test dataset and achieve 0.96 accuracies on the predicted results. The accuracy is better than Decision Tree (0.83), K-nearest neighbours (KNN) (0.94) and Naive Bayes Classifier (0.65). Overall, the Logistic Regression classifier did an excellent job predicting vaccine type based on the side effects.

	Score
Precision	0.96
Recall	0.96
F1	0.96
Accuracy	0.96

Table 3: Evaluation Metrics for Vaccine Type Classifier

2.2.2 Classify R-naught Levels in Malaysia

We used the self-collected dataset containing R-naught values for each state in Malaysia. The classifier is expected to predict R-naught levels for Malaysia based on the R-naught value of each state. The classified results are in three categories: ' Low', 'Medium', 'High'. Within the three classifiers, Logistic Regression, Gaussian Naive Bayes and Decision Tree, Decision Tree classifies the most accurate R-naught level on the test set. We compared the accuracy with and without oversampling using SMOTE, and the accuracy after performing SMOTE is reached up to 87% while before SMOTE is lower than 0.82.

	Score
Precision	0.88
Recall	0.88
F1	0.88
Accuracy	0.87

Table 4: Evaluation Metrics for R-naught Level Classifier

2.3 Regression Model

In this project, we used the regression model to analyze our datasets for predicting the covid-19 r-naught index value, etc. In this section will be using Top 3 most common metrics for evaluating predictions on regression model performance:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R2

2.3.1 R-naught Index Prediction

For r-naught index prediction problem, we have trained five algorithm for predicting the value. The inputs are the r-naught index of each state and the prediction is the Malaysia r-naught index of Malaysia. Among them, the random forest regression algorithm perform the best, it managed to reach the highest R2 value at 0.854 and getting the lowest value of error. The Lasso regression is the worst among all the algorithm being tested with only -0.192 R2 value and the mean absolute error is -0.066, which is the highest absolute error value in this experiment. The other algorithms except SVM regression are having similar error value and R2 value.

Algorithm	MSE	MAE	R2
Linear Regression	-0.001	-0.021	0.799
Decision Tree Regression	-0.001	-0.024	0.778
Lasso Regression	-0.006	-0.066	-0.192
SVM Regression	-0.002	-0.038	0.545
Random Forest Regression	-0.001	-0.021	0.854

Table 5: Evaluation of Regression Algorithm

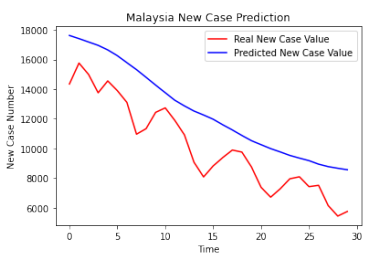
2.3.2 New Covid Case Trend Prediction

Corona Virus Disease 2019 (COVID-19) has spread swiftly to countries all over the world. Because there is no viable cure to this disease, hence, it is important to predict the future cases so that relevant departments can plan ahead and take action immediately. To predict the future trend of covid, we have predicted the trend with a LSTM model and the result shows the future cases in Malaysia will be decreasing.

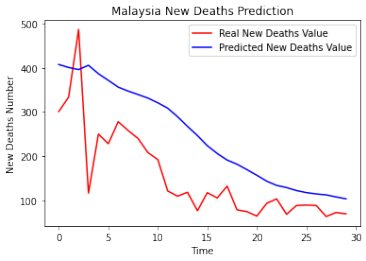
2.3.3 New Covid Deaths Trend Prediction

The increase in the number of confirmed cases of covid is not necessarily the only reason for the increase in the number of deaths. There may be many problems in country's healthcare system that have not yet been discovered. Therefore, we are also obligated to analyze and predict the number of deaths in the future in

order to discover potential problems and be prepared for it. For this purpose, we have trained a LSTM model for the prediction. As the result, the death number for Covid 19 is decreasing along with the decreasing of Covid 19 daily cases.



(a) New Covid Case Trend Prediction by LSTM



(b) New Covid Deaths Trend Prediction by LSTM

References

GitHub - CITF-Malaysia/citf-public: Official data on Malaysia's National Covid-19 Immunisation Programme (PICK). Powered by MySejahtera. (2021, 06). Retrieved from https://github.com/CITF-Malaysia/citf-public

GitHub - MoH-Malaysia/covid19-public: Official data on the COVID-19 epidemic in Malaysia. Powered by CPRC, CPRC Hospital System, MKAK, and MySejahtera. (2021, 07). Retrieved from https://github.com/MoH-Malaysia/covid19-public

Google Trends. (n.d.). Retrieved from https://trends.google.com/trends/?geo=MY

Nilai R Malaysia. (2021). Retrieved from https://covid-19.moh.gov.my/kajian-dan-penyelidikan/nilai-r-malaysia