TDS 3301 Data Mining

Trimester 1

GROUP PROJECT

REPORT

Prepared by :

Name	Student ID	Phone number
Chang Kai Boon	1181101282	011-29946989
Soe Zhao Hong	1181101614	018-9434610

Table of Content

1.0 Exploratory Data Analysis Questions

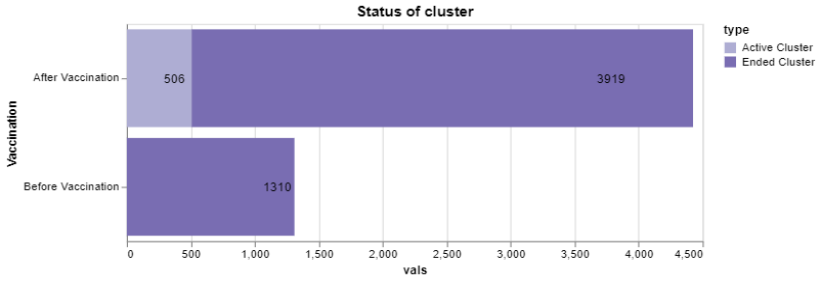
1.1 Any Correlation between vaccination and daily cases for every state of Malaysia?

Correlation is a statistical method used to assess a possible linear association between two continuous variables. So, we set up this question and find out the correlation between vaccination and daily cases in every state of Malaysia. The result will be shown with the highest correlation in vaccination, and daily cases column in the table below.

States	**Vaccination**	**Daily cases**
Johor	Kedah	Pulau Pinang
Perak	Pulau Pinang	Johor
Pahang	Perak	Terengganu
Perlis	Kedah	Terengganu
Terengganu	Johor	Kelantan
Kelantan	Kedah	Terengganu
Kedah	Johor	Sabah
Pulau Pinang	Johor	Perak
Melaka	Pahang	Selangor
Negeri Sembilan	Selangor	W.P. Kuala Lumpur
Selangor	W.P. Kuala Lumpur	W.P. Kuala Lumpur
W.P. Labuan	Sarawak	Negeri Sembilan
W.P. Kuala Lumpur	Selangor	Selangor
W.P. Putrajaya	Selangor	Selangor
Sabah	Johor	Pulau Pinang
Sarawak	W.P. Labuan	Terengganu

1.2 Have the clusters reduced after vaccination is started in Malaysia?

In this question, we separate the timeline of this pandemic into two: before vaccination, and after vaccination. The after vaccination starts 1 month after the vaccination started since the effectiveness of vaccination is able to be viewed after 1 to 2 months. Then, before vaccination is started from the beginning of this pandemic until 1 month after the vaccination started.



Unfortunately, the number of clusters does not decrease when vaccination is started in Malaysia, but the good news is although there are a lot of clusters after the vaccination start, there are only about 500 clusters still active until now which is just 0.11% of the clusters after vaccination started. However, we had explored that only the community and workplace which are the category of the cluster had increased rapidly. We guess this situation happened because the movement control order in Malaysia is not as strict as before the vaccination started. Most of the parents are going out for work. So, this causes the number of clusters to increase.

1.3 What states require attention now?

To explore this question, we have set up 1 sub-questions which is “Has vaccination helped to reduce the daily cases?” “Which states have shown the effectiveness of vaccination?” and “What is the admitted and discharged rate of hospitals and pkrc in every state during this pandemic?”

1.3.1 Has vaccination helped to reduce the daily cases? Which states have shown the effectiveness of vaccination?

Based on our exploration of this question, we found that the vaccination started in February 2021 and increased rapidly in July and August, but the daily cases increase rapidly at the same time unfortunately. As we know, a complete vaccination needs about 1 to 2 months depending on the types of vaccine received. So, the daily cases in September and October of 2021 start to decrease as we predicted.

Besides, we have found that Selangor shows the effectiveness of vaccination. We can approve this conclusion with our exploration. Selangor has the highest number of vaccinations, especially in July and August of 2021 and its daily cases are also the highest at the same time. Then, the daily cases of Selangor decreased in September and October of 2021 with exaggerated ups and downs. So, the explanation above approves that Selangor has shown the effectiveness of vaccination.

1.3.2 What is the admitted and discharged rate of hospitals and pkrc in every state during this pandemic?

In this question, the admitted rate is calculated with admitted cases/ daily cases, the discharged rate is admitted cases/ discharged cases. Based on the result of this question, the admitted rate of hospitals has decreased rapidly, but the discharged rate of hospitals remains the same. In addition, the results of PKRC are slightly different from hospitals. The discharged rate and admitted rate of PKRC rise and fall at the beginning of the pandemic which is in the year 2020, but the discharged rate remains the same and the admitted rate keeps decreasing in the year 2021.

Apart from this, some states had a high admitted rate in hospitals at the beginning of the pandemic which is Kelantan, Sarawak, and Terengganu. Terengganu has the highest admission rate which is about 47.0. But the admission rate in all states of Malaysia does not have a very high value which is just between 0 to 5.0 in the year 2021. At the same time, the discharged rate of hospitals in Negeri Sembilan had the highest value at the beginning of the pandemic which is 9.0 and all states have a better discharged rate which is different from the admitted rate in the year 2021. The discharged rate remains at 0 to 2.0 which is slightly higher than the admitted rate in all states. Therefore, the highest admitted rate in the states of Malaysia is between 35 to 40 which is from W.P. Labuan at the beginning of this pandemic, but the highest discharge rate at the beginning of the pandemic is about 65.0 from Johor. Based on the observation from us, the discharged rate and admitted rate of PKRC in every state are not very high, only in some specific states like Johor and W.P. Labuan will have a very high rate, but all the states have a stable admitted and discharged rate in between 0 to 5.0 in the year 2021.

A short conclusion can be made based on the 2 sub-questions above, Sarawak requires attention right now because the daily cases of Sarawak from July until September keep increasing and the vaccination of Sarawak in June and July has a good number of receiving which is 1 or 2 months before July until September. So, this shows that cases of Sarawak do not affect since the vaccination is started. Another reason is the admission rate of Sarawak has a very high value at the beginning of the pandemic but it does not have a high discharge rate during the middle of this pandemic which means that there are still have a lot of patients in the hospital of Sarawak because of Covid-19 pandemic.

1.4 Does vaccination affect the recovery and death cases?

Since there is a lot of vaccination received in July and August, and the effectiveness of the vaccination is able to show in 1 to 2 months, we will focus on September and October for the recovered cases and death cases. Based on our observation, the recovered cases increase in September but decrease in October. The reason that the recovered cases have decreased in October is that the daily cases have also decreased at that time. Besides, the death cases also increase in September unfortunately, but decrease in October. So, we can make a conclusion that the death cases and recovered cases have the same trend. To explore more detail on death cases, we have explored it with the death cases with partial and fully vaccinated. Based on this exploration, we found that the death cases with vaccination have a high percentage especially in October which contains 50% and above. To summarize, we can conclude that the vaccination can affect the recovered cases but it is not working in death cases. We surmise that death cases

do not decrease because hidden aspects like complications will happen to the patients and cause the effectiveness of vaccines to become very low.

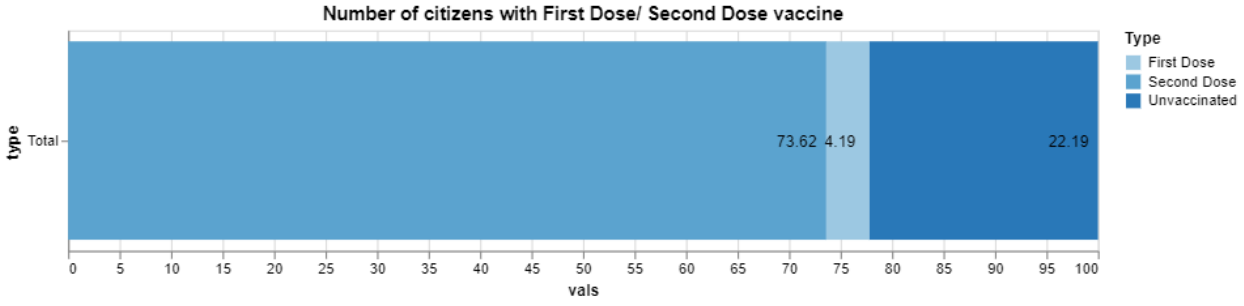
1.5 How is the progress of vaccination in Malaysia?

In this question, we have created a few sub-questions to help us to explore it. For example, “What is the vaccination progress in every state?”, “How many citizens of Malaysia have the first dose of vaccine and how many of them have completed the vaccination?”, and “Which type of vaccine has mostly been received by citizens of Malaysia?”

1.5.1 What is the vaccination progress in every state?

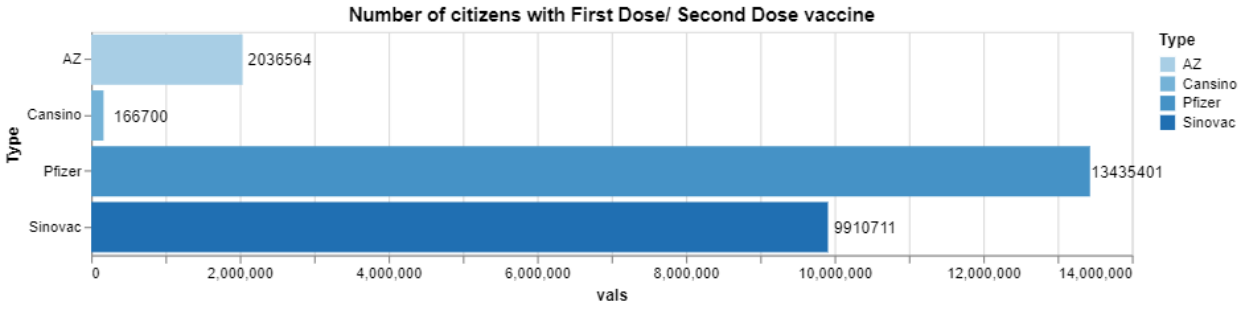
In this question, we have found that there are few states that have done more than 100% of vaccination such as W.P. Kuala Lumpur and W.P. Putrajaya. We are trying to find out the reason that the progress in both states is more than 100% of the population because there are a lot of foreign workers that are working in Malaysia especially in W.P. Kuala Lumpur. So, both states have more than 100% of vaccination progress. Besides, we can observe that partial vaccination in every state is mostly in nearly 70% and above, only a few states like Sabah and Kelantan only have 60.7% and 61.4% of partial vaccination. Then, we can also find that the full vaccination progress and partial vaccination progress only differ within 5%.

1.5.2 How many citizens of Malaysia have the first dose of vaccine and how many of them have completed the vaccination?



Based on the stacked bar chart above, we can see that the overall vaccination progress in Malaysia is almost 80%. The citizens who received the first dose vaccination is about 77.81%, and the second dose vaccination is about 73.62%. Only 22.19% of citizens in Malaysia have not received the vaccination yet.

1.5.3 Which type of vaccine has mostly been received by citizens of Malaysia?

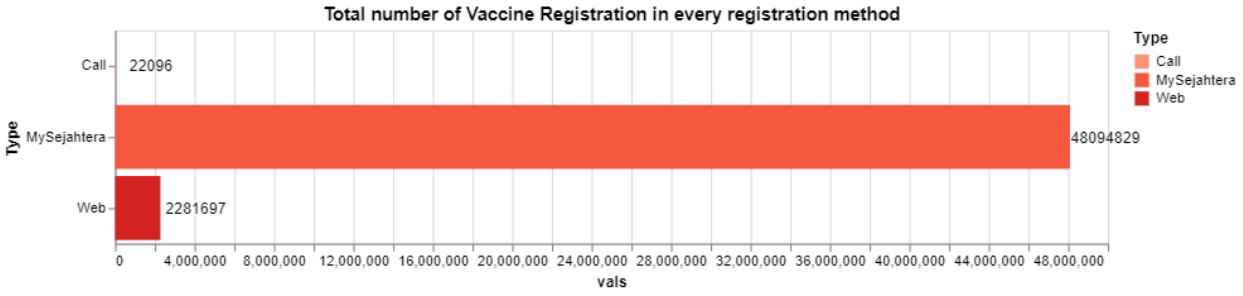


Based on the bar chart above, we can observe that Pfizer is the most popular vaccine in Malaysia and followed by Sinovac, AZ, and Cansino. Although Cansino has the least number of vaccinations in Malaysia, we can see that it has just started to be provided in June 2021 until now and it is increasing. Then, AZ starts to provide in May until now but the number of providing from AZ keeps decreasing. Besides, Pfizer does not always have the highest number of received every month. For example, Sinovac has received more than Pfizer in June and July 2021. But the main trend of vaccination in Malaysia is still Pfizer.

1.6 How is the registration of vaccination in Malaysia?

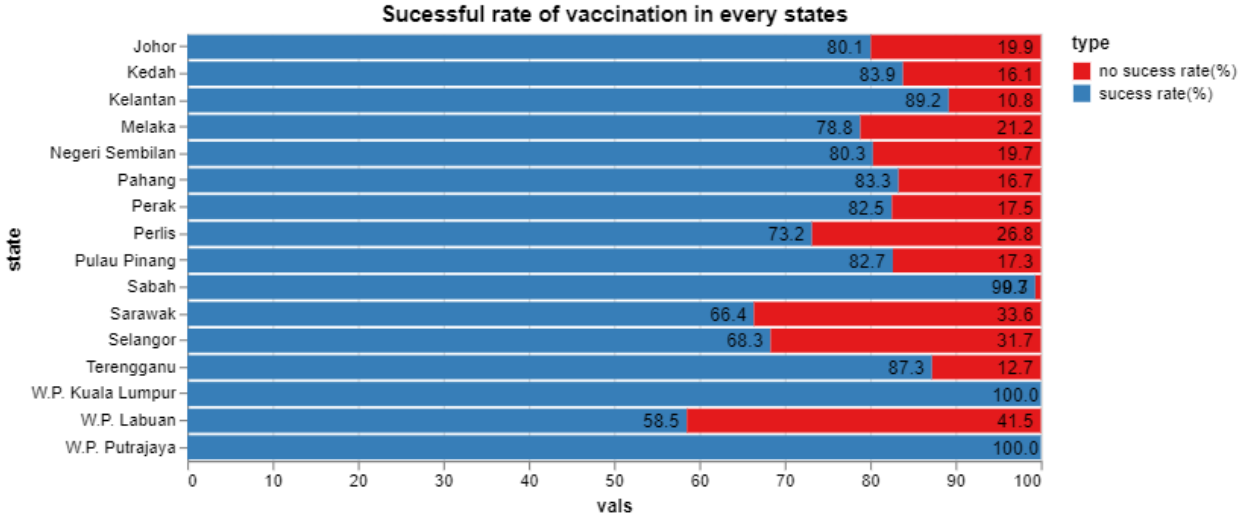
This question is built because we want to know the condition of the citizens while registering for the vaccination in Malaysia. So, there are 2 sub-question is built such as “How do the citizens of Malaysia register for the vaccination, and which has the highest number of registration?” and “What is the success rate of registering for the vaccination in every state?”

1.6.1 How do the citizens of Malaysia register for the vaccination and which has the highest number of registration?



Based on the graph above, we can easily conclude that the most popular way that Malaysians register for the vaccination is by registering through MySejahtera. Only a few Malaysians use the call and web for registering their vaccination especially in the first three months: June, July, and August.

1.6.2 What is the success rate of registering for the vaccination in every state?



Based on the stacked bar graph above, we can see that W.P. Kuala Lumpur and W.P. Putrajaya has 100% of successful registration. Besides, Sabah also has a very high percentage of success rate which is 99.7%. Unfortunately, W.P. Labuan has the lowest success rate on registering vaccination which is only 58.5%. Based on the graph above, we also found out that Sabah has a very high success rate on vaccination but it does not have high vaccination progress which is shown in the section before. In this situation, we had made a conclusion that Sabah has a very low number of registers for vaccination.

2.0 Feature Selection

In this project, we are going to predict 3 scenarios using various machine learning techniques and study their performances. The 3 scenarios are: i) daily covid cases of Malaysia, ii) daily admitted covid cases to hospital across Malaysia and iii) daily ICU cases across Malaysia. We combined some datasets and performed some preprocessing to get the variable we needed. For example, to predict the daily covid cases of Malaysia, we combine the data of daily Malaysia's cases, daily states' cases, daily vaccine, daily death cases, and perform data preprocessing to calculate the covid cases before one day to two weeks.Our aim here to is create as many features as we could, and in the next step perform feature extraction to find the insight and filter out unimportant features. Next, we decided to use Boruta algorithm and Recursive Feature Elimination (RFE) as our feature selector. For each scenario, we will select the top 30, top 50 and top 100 features separately. By doing this, we can reduce the features number from 200-300++ to 30,50 and 100.

As we compare the output generated by Boruta and RFE algorithms, we can conclude that RFE is more likely to be overfitted as it produces a lot of features which have the same ranking of 1. This results in a situation where we cannot clearly differentiate which features are more important than others. Thus, we decided to use the top 30, 50 and 100 features generated by Boruta as it can differentiate the features better. These features will be used to feed into our machine learning model to predict the scenarios.

3.0 Regression

Type	Metrics	Linear Regression	Decision Tree Regression	Random Forest Regression	Bayesian Linear Regression	Support Vector Regression
No Feature Selection	R2	0.998	0.985	0.994	0.997	0.981
	MAE	218.175	579.516	334.68	218.175	673.451
Top 30 Features	R2	0.979	0.961	0.967	0.98	0.539
	MAE	671.663	952.354	901.558	642.206	3453.937
Top 50 Features	R2	0.999	0.992	0.996	0.999	0.832
	MAE	132.51	431.693	276.316	133.67	1990.152
Top 100 Features	R2	0.999	0.99	0.996	0.998	0.9685
	MAE	154.465	433.77	276.088	154.76	863.507

From the table above, we can see that the best performing model for predicting daily covid cases of Malaysia is the linear regression model with R2 of 0.999 and MAE of 132.5. Next, it is followed by the Bayesian linear regression model with R2 of 0.999 and MAE of 133.67. Both of the results are obtained using the dataset of top 50 features.

Type	Metrics	Linear Regression	Decision Tree Regression	Random Forest Regression	Bayesian Linear Regression	Support Vector Regression
No Feature Selection	R2	0.981	0.781	0.887	0.981	0.936
	MAE	51.644	188.697	125.334	51.046	88.531
Top 30 Features	R2	0.929	0.744	0.862	0.928	0.828
	MAE	103.126	188.65	142.499	105.626	157.92
Top 50 Features	R2	0.982	0.796	0.899	0.982	0.904
	MAE	53.918	152.139	112.78	52.797	109.55
Top 100 Features	R2	0.975	0.768	0.861	0.984	0.926
	MAE	57.72	188.65	128.965	48.531	94.726

Moreover, we can see that the best performing model for predicting daily admitted covid cases to hospital across Malaysia is Bayesian linear regression model with R2 of 0.984 and MAE of 48.531 using the dataset of top 100 features, followed by the same model with R2 of 0.981 and MAE of 51.046 using the dataset that do not undergo any feature selection.

Type	Metrics	Linear Regression	Decision Tree Regression	Random Forest Regression	Bayesian Linear Regression	Support Vector Regression
No Feature Selection	R2	0.989	0.965	0.983	0.989	0.987
	MAE	18.8	39.39	26.3	18.88	20.897
Top 30 Features	R2	0.969	0.943	0.974	0.966	0.959
	MAE	37.673	47.511	34.59	40.415	43.07
Top 50 Features	R2	0.957	0.959	0.968	0.96	0.952
	MAE	45.896	44.04	38.316	44.39	46.7
Top 100 Features	R2	0.942	0.852	0.97	0.978	0.969
	MAE	50.8	61.65	37.441	32.09	37.635

Furthermore, the table shows us that the best performing model for predicting daily ICU cases across Malaysia is the Linear Regression and Bayesian Linear Regression model. Both of them have the same R2 and MAE which are 0.989 and 18.8 respectively using the dataset that does not undergo any feature selection. In general, we can conclude that for the regression model, Linear Regression and Bayesian Linear Regression outperforms other models.

4.0 Classification

Category	Type	Metrics	Decision Tree Classifier	Random Forest Classifier	Support Vector Machine	Naive Bayes	KNN
No Feature Selection	No SMOTE	F1 Score	0.77	0.78	0.75	0.72	0.72
		Accuracy	0.77	0.79	0.76	0.73	0.73
	SMOTE	F1 Score	0.88	0.87	0.85	0.81	0.82
		Accuracy	0.89	0.87	0.85	0.81	0.82
Top 30 Features	No SMOTE	F1 Score	0.73	0.74	0.72	0.7	0.7
		Accuracy	0.73	0.74	0.73	0.71	0.71
	SMOTE	F1 Score	0.79	0.8	0.8	0.77	0.71
		Accuracy	0.79	0.81	0.81	0.77	0.73
Top 50 Features	No SMOTE	F1 Score	0.77	0.78	0.75	0.72	0.72
		Accuracy	0.77	0.79	0.76	0.73	0.73
	SMOTE	F1 Score	0.88	0.87	0.85	0.81	0.82
		Accuracy	0.89	0.87	0.85	0.81	0.82
Top 100 Features	No SMOTE	F1 Score	0.82	0.77	0.88	0.75	0.7
		Accuracy	0.84	0.77	0.89	0.76	0.71
	SMOTE	F1 Score	0.88	0.84	0.84	0.78	0.82
		Accuracy	0.89	0.84	0.84	0.79	0.82

To predict daily covid case severity , Decision Tree Classifier(DTC) model outperforms other models with the accuracy and F1-score of 0.89 and 0.88 respectively, testing on the dataset of top 50 features, top 100 features, not undergoing feature selection and SMOTE is applied to the dataset.

Category	Type	Metrics	Decision Tree Classifier	Random Forest Classifier	Support Vector Machine	Naive Bayes	KNN
No Feature Selection	No SMOTE	F1 Score	0.8	0.79	0.86	0.72	0.79
		Accuracy	0.79	0.79	0.86	0.72	0.79
	SMOTE	F1 Score	0.86	0.82	0.75	0.72	0.77
		Accuracy	0.86	0.81	0.74	0.72	0.77
Top 30 Features	No SMOTE	F1 Score	0.84	0.77	0.84	0.77	0.75
		Accuracy	0.84	0.77	0.84	0.77	0.74
	SMOTE	F1 Score	0.91	0.77	0.82	0.74	0.69
		Accuracy	0.91	0.77	0.81	0.74	0.7
Top 50 Features	No SMOTE	F1 Score	0.86	0.84	0.86	0.77	0.79
		Accuracy	0.86	0.84	0.86	0.77	0.79
	SMOTE	F1 Score	0.91	0.82	0.91	0.74	0.68
		Accuracy	0.91	0.81	0.91	0.74	0.67
Top 100 Features	No SMOTE	F1 Score	0.89	0.84	0.86	0.72	0.82
		Accuracy	0.88	0.84	0.86	0.72	0.81
	SMOTE	F1 Score	0.89	0.82	0.89	0.72	0.82
		Accuracy	0.88	0.81	0.88	0.72	0.81

The best performing models for predicting daily admitted covid cases to hospital severity are DTC and SVM model with the accuracy and F1-score of 0.91 and 0.91 respectively, testing on the dataset of top 30 and top 50 features which SMOTE is applied to the dataset.

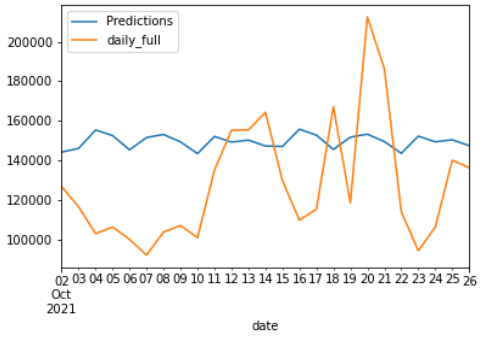
Category	Type	Metrics	Decision Tree Classifier	Random Forest Classifier	Support Vector Machine	Naive Bayes	KNN
No Feature Selection	No SMOTE	F1 Score	0.93	0.88	0.88	0.78	0.86
		Accuracy	0.93	0.88	0.88	0.81	0.86
	SMOTE	F1 Score	0.95	0.89	0.81	0.81	0.84
		Accuracy	0.95	0.88	0.81	0.81	0.84
Top 30 Features	No SMOTE	F1 Score	0.81	0.84	0.81	0.89	0.81
		Accuracy	0.81	0.84	0.81	0.88	0.81
	SMOTE	F1 Score	0.81	0.86	0.82	0.83	0.63
		Accuracy	0.81	0.86	0.81	0.84	0.61
Top 50 Features	No SMOTE	F1 Score	0.81	0.89	0.8	0.86	0.79
		Accuracy	0.81	0.88	0.81	0.86	0.79
	SMOTE	F1 Score	0.84	0.86	0.85	0.81	0.8
		Accuracy	0.84	0.86	0.84	0.81	0.79
Top 100 Features	No SMOTE	F1 Score	0.91	0.93	0.86	0.89	0.93
		Accuracy	0.91	0.93	0.86	0.88	0.93
	SMOTE	F1 Score	0.89	0.89	0.93	0.84	0.83
		Accuracy	0.88	0.88	0.93	0.84	0.81

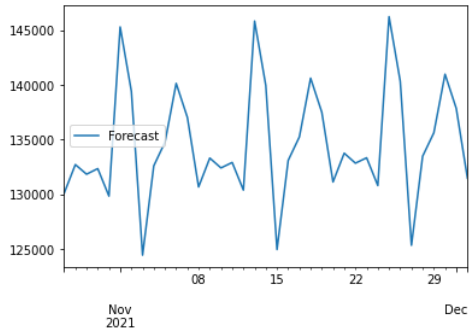
The best performing model for predicting the daily ICU cases severity is DTC with the accuracy and F1-score of 0.95 respectively, testing on the dataset which is not undergoing feature selection but SMOTE is applied. In conclusion, DTC and SVM performed the best among other classification models in 3 scenarios. We highly recommend you to apply SMOTE as it has been proved to increase the performance.

5.0 Herd Immunity Prediction

5.1 ARIMA

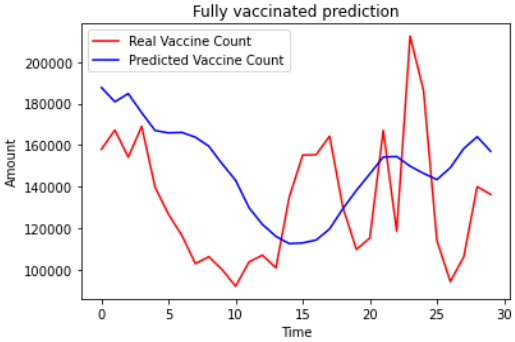
ARIMA stands for AutoRegression Integrated Moving Average. In order to use the ARIMA model, we need to specify the order in terms of (p,d,q). We have decided to use the auto-arima package to help us loop through a range of values and decide the best value for (p,d,q). By running through the auto-arima function with the vaccination data, the model suggested a SARIMAX model with order of (2,0,1) and seasonal order of (0,1,1,12). By using the suggested model and parameters, our SARIMAX model obtains a RMSE of 37344 as shown in the picture. By using the trained model for forecasting, the model predicts that the total predicted number of vaccines for the next 5 weeks is 4980740. With this amount of vaccine, we are able to reach 88.8% of herd immunity by the end of November 2021.

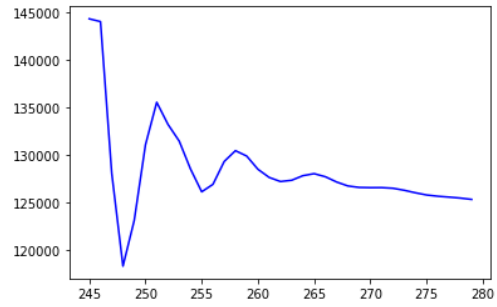




5.2 LSTM

LSTM stands for Long Short Term Memory, which is a type of Recurrent Neural Network (RNN) used in sequence prediction problems.The process can be divided into 2 parts: In the first part, we split the vaccine dataset into a train and test set to build the model. In the second part, with the trained model, we pass in the whole dataset as input to forecast the number of vaccines for the coming month. We set the lookup variable to 7, meaning we will use the last 7 days's data to predict the next day value. Besides, the model obtained the RMSE of 37396. Next, we fed the whole dataset into this model and used it to forecast the vaccine number given for the coming month. From the result, we can see that the total predicted number of vaccines for the next 5 weeks is 4493483. With this amount of vaccine, we are able to reach 87.38% of herd immunity by the end of November 2021.





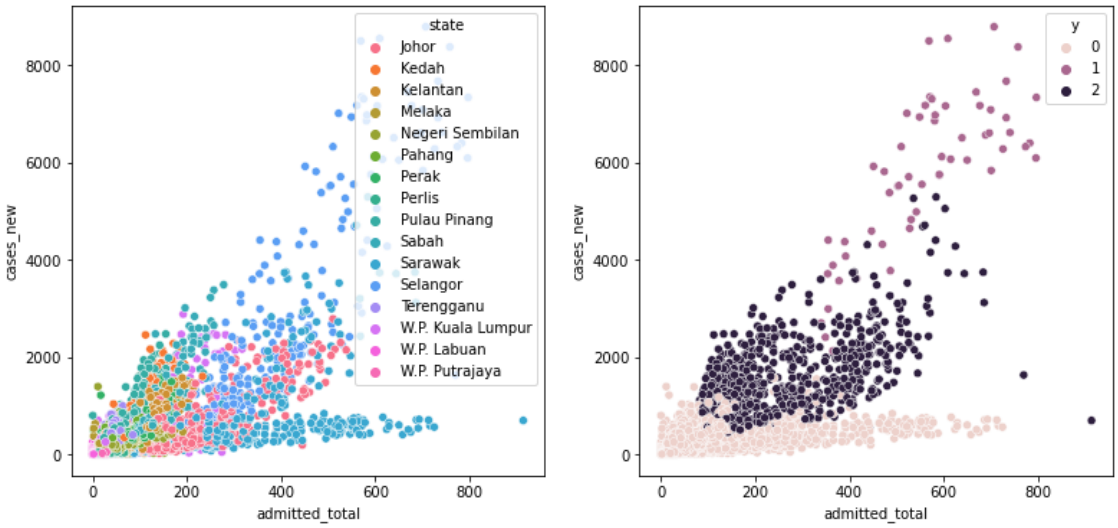
5.3 Basics Mathematics

In this method, we are grouping the vaccination data into a weekly manner. After that, we calculate the change of vaccine per week to obtain the weekly rate.From the current weekly rate, we predict the total vaccine given for next week until 30/11/2021.From the result, we can see that the total predicted number of vaccines for the next 5 weeks is 4013075. With this amount of vaccine, we are able to reach 85.9% of herd immunity by the end of November 2021. In conclusion, based on the three of the approaches we have tried, all three approaches have predicted that we can reach 80% of herd immunity by 30 November 2021.

6.0 Clustering

Clustering is a task of dividing the data points into a number of groups so that the data points in the same group are more similar to each other than those in the other group. In this section, we are going to cluster 2 scenarios, they are: daily covid cases and admitted to hospital cases in each state and daily covid cases and daily death cases in each state. Three clustering techniques(KMeans, DBSCAN and agglomerative clustering) are selected to be used in this section.

From the images below, we can see that the data points have been grouped into 3 groups with clear borders.After applying clustering, a new column named [‘y’] will be generated and we can use that variable to define the risk. For example, at the output of KMeans clustering, the data points are grouped into cluster 0, 1 and 2 while cluster 0 starts from the left bottom corner, followed by cluster 2 and cluster 1. For this situation, cluster 1 has higher risk than cluster 0 and 2 as it appears at the right corner of the graph, where the x and y variables (cases_new and admitted_total) are high. Thus, any data points that are categorized into cluster 1 have a higher risk compared to cluster 2 and 0. In addition, we also can use the [‘y’] variable generated to be the target variable and apply it to a classification problem to predict whether the new data fall in which level of risk.



Link to Heroku : https://bukancovidnowdua.herokuapp.com/