International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Implementation of Support Vector Machine (SVM) on Province Classification in Indonesia Based on Herd Immunity Rate

Ayunda Maudiatama[1], Aviarini Indrati[2]
[1]Master of Information Systems Management Department, Business Information System, Gunadarma University, Jakarta, Indonesia, 10440
[2]Master of Information Systems Management Department, Business Information System, Gunadarma University, Jakarta, Indonesia, 10440

Abstract**—** *One of the government's policies and efforts to prevent virus mutation and suppress the rate of increase in positive confirmed cases of COVID-19 is the existence of a vaccination program to achieve herd immunity. Therefore, this study carried out a classification that aims to classify provinces in Indonesia based on the level of achievement of the herd immunity target, to understand which provinces have reached herd immunity or not. The method used in this study is Support Vector Machine (SVM). This study used COVID-19 first and second doses vaccination data from January 2, 2021 to January 22, 2022 through Ministry of Health Republic of Indonesia's website https://vacsin.kemkes.go.id/#/detail_data. Implementation of SVM classification in this study is divided into three scenarios with different training data and testing data ratio, including 60%:40%, 70%:30%, and 80%:20%. Classification results shows that the performance of SVM works well in the distribution data of 70%:30% in the first and second vaccination data, which produce high accuracy values of 94.8% and 98.3% respectively because class labeling on the data can affect high outcome. Furthermore, the average precision value is 64.8% and 66.3%, the average recall value of 94.1% and 99.2%, the average f1-score value of 71.2% and 74.2%, so it can be concluded that the performance of SVM depends on the data used and data splitting ratio.*

Keywords**—** *Herd Immunity, Classification, Support Vector Machine.*

I. INTRODUCTION

COVID-19 spreading in Indonesia is still keep increasing, so the government issued policy and tried to stop that virus mutation so it will not spread any further and suppress COVID-19 case increasing. One of those policy or efforts is achieving herd immunity or group immunity. By achieving herd immunity, if most people in a group were immune toward a contagious plague, there is a higher chance for every person in that group will be protected, not only the person who immune to that plague. Herd immunity can be created by vaccination. Thus, government created COVID-19 vaccination program to achieve herd immunity. The government put on target 70% from entire Indonesia population or about 182 million population who has been vaccinated [1].

Up to 24th October 2021, total of first vaccination was 113.0 million or 54.3% from target, and second vaccination was 67.9 million or 32.6% from target [2]. Those vaccinated total were accumulation from each province in Indonesia. This means, vaccination program has yet reached its target so the government needs more effort to reach that target. This is also caused by vaccination program from each province that was not evenly. For that reason, to make the government easier to carry vaccination evenly for each province in Indonesia, information of province where it has yet to reach herd immunity target is needed. Furthermore, a system that can identify a province in reaching herd immunity is needed. One of those efforts to identifying is by using classification concept, by using Support Vector Machine (SVM).

In this study, a classification was carried out which aims to identify provinces based on the level of herd immunity target reached which is divided into two classes, not reaching the target and have reached the target. Classification uses the SVM method which trained to classify between the two classes by integrating vaccination target and total vaccinated recipient data. The results of the classification will be analyzed to find out how well the performance of SVM in conducting the classification so that it can be used to identify which provinces have not reached the herd immunity target. So that there can be an even distribution of vaccination programs to every province in Indonesia.

II. LITERATURE REVIEW

In several previous studies, SVM was widely used for sentiment analysis with good results, such as study [3] related to KAI tweets getting 80.59% accuracy, study [4] about the Zoom Cloud Meeting application review getting 91.61% accuracy, study [5] regarding twitter tweets related to Jakarta, Bandung, and Medan, obtained a precision and recall value of 88%, and study [6] on Twitter tweets containing the word "mudik" obtained an accuracy of 87%. Each of these studies got good classification results using SVM. Study [7] obtained an accuracy of 70% in classifying drunk drivers and normal drivers using SVM. Study [8] in the classification of tweets containing cyberbullying obtained an accuracy of 76.66%. Study [9] in classifying student learning outcomes using SVM obtained higher accuracy result of 90.91%. Study [11] in the classification of the lungs with tuberculosis or not was getting an accuracy of 79%. Based on these studies, SVM can be used for various classification fields.

There have been several previous studies that compared the performance of SVM with other classification methods such as research [11] which compared the SVM method with K-Nearest Neighbor (KNN) in the classification of big cats based on animal skin patterns with the results showing that the

Ayunda Maudiatama and Aviarini Indrati, "Implementation of Support Vector Machine (SVM) on Province Classification in Indonesia Based on Herd Immunity Rate," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 226-230, 2022.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

SVM classification got an accuracy of 91 ,7% which is better than KNN which gets an accuracy of 69.47%. In addition, studies [12] and [6] compared the performance of SVM with the Naive Bayes method in the classification of sentiment analysis of Customer Review and Homecoming.

The classification of Customer Review sentiment analysis by [12] aims to divide the reviews obtained from TripAdvisor, Booking.com, Expedia, Agoda and Pegi-Pegi into five aspects of location, room, food, price, and service where each aspect is divided into two sentiment, positive and negative which can be concluded that in every aspect SVM is always superior to Naive Bayes with an average accuracy of 84.6%, while Naive Bayes got an average accuracy of 81.4%. Study [6] compared the SVM method with Naive Bayes in classifying tweets containing the word "mudik" into three sentiments, positive, negative, and neutral. The results shows that the Naive Bayes method got an accuracy of 82% while SVM got a higher accuracy result of 87%. Based on these studies, SVM is a superior method to be used in the classification process.

III. Research Method

This study begins with collecting data on COVID-19 vaccinations. The second step is processing the data that will be used in this study. After the data is processed, the next step is labeling data. After all data has been labeled, the next step is to implement the SVM method classification. After that, perform an analysis based on the results of the SVM classification performance that has been carried out. The last stage is to test the SVM model that has been created. The stages of research that will be carried out in this study are shown in Figure 1.

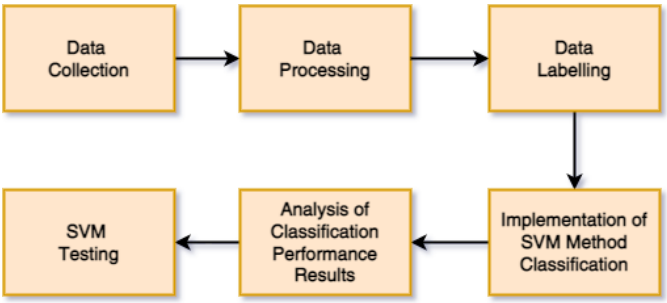

Fig. 1. Research Stages

A. Data Collection

The data collected will be used as input data into SVM model. The data taken is vaccine data which can be used as an indicator to determine the level of herd immunity target reached. The indicators or parameters used in this study are the number of vaccine recipients and vaccine targets for each province in Indonesia. The collected vaccine data are divided into two parts, the first dose of vaccine and the second dose of vaccine. The data collected was taken from Indonesian Ministry of Health by downloading an excel file through the website https://vacsin.kemkes.go.id/#/detail_data. The data collected is vaccine data starting from January 2, 2021 to January 22, 2022.

B. Data Preprocessing

At this stage, there are three processes that are carried out manually, data cleaning, data transformation, and changing field names. All three processes can be seen in Figure 2.

Fig. 2. Data Preprocessing Stages

The data cleaning process is carried out to remove data that is not used in this study, as if data has no value, data of different format, and duplicated data. After that, data transformation process is carried out by improving the data structure so this data can be used as input data in the SVM classification process. Furthermore, field name changes were made to make it more informative. The results of this data preprocessing stage are vaccine data that has been preprocessed according to the needs of the SVM model input data so that it can be used in the next stage.

C. Data Labelling

At data labeling stage, data that has gone through the data preprocessing stages is processed by giving them a class label to the data which aims to determine the class for each data. In this study, labeling was done manually which consisted of two classes, the class that had reached the target which denoted as 1 and the class had not reached the target which denoted as 0. If the vaccinated value was greater than the target value, it was assigned a value of 1, otherwise it will be given a value of 0. The results of this data labeling stage are vaccine data that has been labeled.

D. Implementation of SVM Method Classification

The next step is to classify the SVM method implemented by Python programming language with the help of the scikit-learn library as a tool that can perform SVM classification. This stage consists of four processes, data splitting, data normalization, SVM training, and SVM testing. The four processes can be seen in Figure 3.

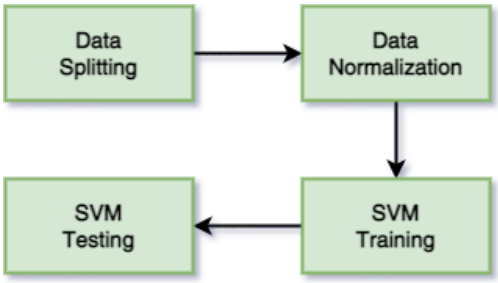

Fig. 3. Stages of SVM Classification Implementation

In data splitting process, data that has been labeled will be separated. The data will be divided into two parts, training data and testing data. Training data is used as input data in the SVM at training process, while testing data is used as input data in the SVM at testing process. In this study, a comparison of data distribution between training data and testing data will

Ayunda Maudiatama and Aviarini Indrati, "Implementation of Support Vector Machine (SVM) on Province Classification in Indonesia Based on Herd Immunity Rate," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 226-230, 2022.

be carried out with three scenarios which can be seen in Table I.

TABLE I. Splitting Data Scenario

Data	Scenario Name	Training Data	Testing Data
First Vaccination	Scenario 1	60%	40%
	Scenario 2	70%	30%
	Scenario 3	80%	20%
Second Vaccination	Scenario 1	60%	40%
	Scenario 2	70%	30%
	Scenario 3	80%	20%

The distribution ratio of data separation in Table I refers to several studies [10],[4], and [8]. Each scenario in Table I will go through the process of data normalization, SVM training, and SVM testing.

After training data and testing data obtained, the next process is data normalization. The training data and testing data must be normalized so that they have the same value range so that the SVM training and testing process will produce stable data. Next is the SVM training and testing process. The training process is carried out using training data to train the SVM model so that the SVM model used can classify properly. After the SVM is trained, then the SVM is tested in the testing process to test the previously trained SVM model using test data. The results of SVM testing process are in the form of confusion matrix table where the results are obtained based on the data from class prediction results compared with the labeled class. Based on this table, the accuracy, precision, recall, and f1-score values will be obtained.

E. Analysis of Classification Performance Results

At this stage, an analysis will be carried out based on the results of the program output so that it can be seen how well the SVM performs in classifying data to get a conclusion.

F. SVM Testing

After the SVM model has succeeded in classifying the data, the SVM have to be tested to identify new data. In other words, the SVM model will be inputted with new data, then the data will be identified by the SVM model to issue the correct classification class results. To get the classification results of class 0, SVM must be able to identify the new data entry classified as class 0. Thus, the test will be carried out by inputting the vaccinated number that is lower than the target number. On the other hand, to obtain results of class 1, the SVM must be able to identify the new data classified as class 1. Thus, the test will be carried out by inputting the number of vaccinated that are greater than the target number.

IV. Results and Discussions

A. Data Collecting Result

Data was collected from website with .xlsx extension named as “Time Series Pelaksanaan Vaksinasi Pertama Per Provinsi (2)” as first vaccination data; “Time Series Pelaksanaan Vaksinasi Kedua Per Provinsi 2” as second vaccination data which each data has 34 records and 384 columns in total.

B. Data Preprocessing Result

After first vaccination data and second vaccination data were processed, 13056 records were collected from each data. As for processed first vaccination data result could be seen as Table II below.

TABLE II. Preprocessing Result of First Vaccination Data

No	Province	Date	Vaccinated	Target
1	Aceh	01/02/2021	0	4028891
2	Aceh	01/03/2021	0	4028891
3	Aceh	01/04/2021	0	4028891
4	Aceh	01/05/2021	0	4028891
...	...	...	...	...
13057	Yogyakarta	1/22/2022	3127367	2879699

Following processed second vaccination data result as Table III below.

TABLE III. Preprocessing Result of Second Vaccination Data

No	Province	Date	Vaccinated	Target
1	Aceh	01/02/2021	0	4028891
2	Aceh	01/03/2021	0	4028891
3	Aceh	01/04/2021	0	4028891
4	Aceh	01/05/2021	0	4028891
...	...	...	...	...
13057	Yogyakarta	1/22/2022	2647926	2879699

C. Data Labelling Result

Prior to be used as SVM input model data, collected data must be classified beforehand. As for first vaccination data labelling result could be seen as Table IV below.

TABLE IV. First Vaccination Data Labelling Result

No	Province	Date	Vaccinated	Target	Class
1	Aceh	01/02/2021	0	4028891	0
2	Aceh	01/03/2021	0	4028891	0
3	Aceh	01/04/2021	0	4028891	0
4	Aceh	01/05/2021	0	4028891	0
...	...	...	...	...	...
13057	Yogyakarta	1/22/2022	3127367	2879699	1

Following second vaccination data labelling result as Table V below.

TABLE V. Second Vaccination Data Labelling Result

No	Province	Date	Vaccinated	Target	Class
1	Aceh	01/02/2021	0	4028891	0
2	Aceh	01/03/2021	0	4028891	0
3	Aceh	01/04/2021	0	4028891	0
4	Aceh	01/05/2021	0	4028891	0
...	...	...	...	...	...
13057	Yogyakarta	1/22/2022	2647926	2879699	0

D. SVM Classification Method Implementation Result

In this part will be showing the SVM classification implementation result on conducted program to each scenario by using first vaccination and second vaccination input data. As for SVM classification implementation summary on first vaccination data could be seen as Table VI below.

Ayunda Maudiatama and Aviarini Indrati, “Implementation of Support Vector Machine (SVM) on Province Classification in Indonesia Based on Herd Immunity Rate,” *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 226-230, 2022.

TABLE VI. First Vaccination Data Result Summary

Scenario	Class	Accuracy	Precision	Recall	F1-Sore
Scenario 1 60% : 40%	0	94.2%	99.8%	94.3%	97%
	1		29.3%	90.9%	44.4%
	$\bar{x}$		64.5%	92.6%	70.7%
Scenario 2 70% : 30%	0	94.8%	99.8%	94.9%	97.3%
	1		29.7%	93.3%	45.1%
	$\bar{x}$		64.8%	94.1%	71.2%
Scenario 3 80% : 20%	0	92%	100%	91.8%	95.7%
	1		26.9%	98.7%	42.3%
	$\bar{x}$		63.4%	95.2%	69%

Following, second vaccination data program outcome result summary as Table VII below.

TABLE VII. Second Vaccination Data Result Summary

Scenario	Class	Accuracy	Precision	Recall	F1-Sore
Scenario 1 60% : 40%	0	98.2%	100%	98.2%	99.1%
	1		28.1%	100%	43.9%
	$\bar{x}$		64.1%	99.1%	71.5%
Scenario 2 70% : 30%	0	98.3%	100%	98.3%	99.1%
	1		32.7%	100%	49.2%
	$\bar{x}$		66.3%	99.2%	74.2%
Scenario 3 80% : 20%	0	97.9%	100%	97.8%	98.9%
	1		22.2%	100%	36.4%
	$\bar{x}$		61.1%	98.9%	67.6%

E. SVM Classification Performation Result Analysis

Based on Table VI and Tabel VII, it could be concluded that all scenario has >90% accuracy score, as class 0 label has more data than class 1 label. This means, class label on data affects higher accuracy result. Highest accuracy score of first and second vaccination data is on scenario 2 which each data has 94.6% and 98.3%. Highest overall score of first and second vaccination data is on scenario 2 which each data has 64.8% and 66.3%. Highest recall score of first vaccination data is on scenario 3 which has 95.2%, while second vaccination data is on scenario 2 which has 99.2%. Highest f1-score of first and second vaccination data is on scenario 2 which each data has 71.2% and 74.2%. Those results show that SVM performance in classification depends on used data, also separation data between training data and testing data. Higher training data than testing data separation ratio does not affect accuracy which get higher and vice versa, because splitting data process is carried randomly. Thus, it can be concluded that created SVM model could classify provinces in Indonesia according to herd immunity target progress with a good performance on training data 70% and testing data 30%.

F. SVM Testing Result

In this section, will be shown the program's outcome of SVM testing process. The implementation of this SVM testing will use the SVM model with data splitting ratio between training data and testing data of 70%:30%. The results of the SVM model testing on the first vaccination data can be seen in Figure 4.

```
======== TESTING ========
Input Province Name: Jawa Timur
Data Target: 31826206
Input Vaccinated Data: 2899506
Classification Class: [0]

Input Province Name: Bali
Data Target: 3405130
Input Vaccinated Data: 3899702
Classification Class: [1]
```

Fig. 4. SVM Testing Result on First Vaccination Data

Following, the results of the SVM model testing on the second vaccine data are shown in Figure 5.

```
======== TESTING ========
Input Province Name: DKI Jakarta
Data Target: 8395427
Input Vaccinated Data: 18395427
Classification Class: [1]

Input Province Name: Jambi
Data Target: 2686193
Input Vaccinated Data: 1630108
Classification Class: [0]
```

Fig. 5. SVM Testing Result on Second Vaccination Data

Figure 4 and Figure 5 show that the SVM model produces the right class results in classifying provinces in Indonesia based on the level of herd immunity target reached.

V. CONCLUSION AND RECOMMENDATION

A. Conclusion

Based on the results of study conducted, it can be concluded as below:

1. The SVM method can be applied in the classification of provinces in Indonesia based on the level of herd immunity target reached.
2. The performance of SVM works well in the classification of provinces in Indonesia based on the level of herd immunity target reached at the ratio of 70%:30% in data separation between training data and testing data of the first and second vaccination data. The resulting accuracy values are 94.8% and 98.3% respectively, the average precision values are 64.8% and 66.3%, the recall averages are 94.1% and 99.2%, the f1-score averages are 71.2% and 74.2%
3. High accuracy result can be affected by class labeling on the data.
4. The performance of SVM in classifying depends on the data used and the separation ratio between training data and testing data.

B. Recommendation

This study was limited to the first and second vaccinations. However, since this study was made, the government has announced the third vaccination program, so that further development can be carried out by adding new data for the

Ayunda Maudiatama and Aviarini Indrati, "Implementation of Support Vector Machine (SVM) on Province Classification in Indonesia Based on Herd Immunity Rate," *International Research Journal of Advanced Engineering and Science*, Volume 7, Issue 1, pp. 226-230, 2022.

third vaccination in the classification. In addition, SVM has the ability to process data at a higher dimension, so that further development can be carried out by adding new parameters that can determine the herd immunity target in a province.

REFERENCES

[1] CNN Indonesia, "PR Berat Pemerintah Capai Herd Immunity Lewat Vaksinasi Covid," 2021 Available: https://www.cnnindonesia.com/nasional/20210113160954-20-593247/pr-berat-pemerintah-capai-herd-immunity-lewat-vaksinasi-covid, [Accessed 24 Oktober 2021].

[2] BeritaSatu, "Data Penerima Vaksin Covid-19 sampai 24 Oktober 2021," 2021. Available: https://www.beritasatu.com/berita-grafik/845023/data-penerima-vaksin-covid19-sampai-24-oktober-2021, [Accessed 24 Oktober 2021].

[3] Fitriana, Dhina Nur. & Sibaroni, Yuliant., "Sentiment Analysis on KAI Twitter Post Using Multiclass Support Vector Machine (SVM)," *RESTI Journal (System Engineering and Information Technology)*, Vol. 4, No. 5, 846 – 853, 2020.

[4] Wardhana, Janu Akrama., and Sibaroni, Yuliant., "Aspect Level Sentiment Analysis on Zoom Cloud Meetings App Review using LDA," *RESTI Journal (System Engineering and Information Technology)*, Vol. 5, No. 4, 631 – 638, 2021.

[5] Putri, Tansa Trisna Astono., Mendoza, Mhd. Dominique., and Alie, Muhammad Fadhiel., "Sentiment Analysis On Twitter Using The Target-dependent Approach And The Support Vector Machine (SVM) Method," *Jurnal Mantik*, Volume 4, Number 1, pp. 20-26, 2020.

[6] Sautomo, Sabar., Hafidz, Noor., Achyani, Yuni Eka., and Gata, Windu., "Sentiment Analysis Due to "Mudik" Prohibited of COVID-19 Through Twitter," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, VOL. 6, NO. 1, P-ISSN: 2685-8223 | E-ISSN: 2527-4864, 2020.

[7] Chen, Huiqin., and Chen, Lei., "Support Vector Machine Classification of Drunk Driving Behaviour," *International Journal of Environmental Research and Public Health*, 14, 108, 2017.

[8] Purnamasari, Ni Made Gita Dwi., Fauzi, M. Ali., Indriati., and Dewi, Liana Shinta., "Cyberbullying Identification in Twitter using Support Vector Machine and Information Gain Based Feature Selection," *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 18, No. 3, pp. 1494~1500, ISSN: 2502-4752, 2020.

[9] Saputra, Elin Panca., Sugiono., Indriyanti., Supriatiningsih., and Nurdin, Hafis., "Grouping of Success Levels in E-Learning Learning Factors: Approaches with Machine Learning Algorithm," *Jurnal Mantik*, Volume 5, Number 1, pp. 78-85, 2021.

[10] Sudirman, Ira Farenda., Giawa, Winda Hartati., Sarumaha Intan Permatasari., Ndraha, Sukurman., and Fawwaz, Insidini., "Linear Kernel and Polynomial Analysis in Recognizing Tuberculosis Image Using HOG Feature Extraction," *Jurnal Mantik*, Volume 4, Number 3, pp. 1693-1698, 2020.

[11] Pratama, Fernanda Januar., Al Maki, Wikky Fawwaz., and Sthevanie, Febryanti., "Big Cats Classification Based on Body Covering," *RESTI Journal (System Engineering and Information Technology)*, Vol. 5, No. 5, 984 – 991, 2021.

[12] Bachtiar, Fitra A., Paulina, Wirdhayanti., and Rusydi, Alfi Nur., "Text Mining for Aspect Based Sentiment Analysis on Customer Review: A Case Study in The Hotel Industry," in *Conference: 5th International Workshop on Innovations in Information and Communication Science and Technology*, Malang, Indonesia, 2020.