Air Quality Analysis in Yogyakarta during the COVID-19 Pandemic Using Naive Bayes and Support Vector Machine

Alisha Zahra Saadiya
Statistics Department,
School of
Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
alisha.saadiya@binus.ac.id

Edrick Setiawan
Statistics Department,
School of
Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
edrick.setiawan@binus.ac.id

Gregory Nicolla
Statistics Department,
School of
Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
gregory.nicolla@binus.ac.id

Abstract – Air quality is essential for humans and the environment. Living things need clean air to minimize the probabilities of getting a disease. However during Covid-19, the virus has spread to various cities which polluted the air making people infected by the coronavirus. Yogyakarta is one of the biggest cities that has been affected by air pollution before and during the pandemic. This study aimed to use Naive Bayes and Support Vector Machine on the data of air quality in Yogyakarta. Using Naive Bayes and SVM this study wanted to predict if the air quality is good or bad by comparing the different amounts of components in the air. The best method to use for predicting the air quality in Yogyakarta is Support Machine Vector with an accuracy on data training is 1.00 and an accuracy on data testing is 0.95. This is based on the f1-score that has been calculated in the testing dataset.

Keywords – Naive Bayes, Support Vector Machine, Air Quality

I. Introduction

COVID-19 first entered Indonesia in March 2020 in Depok, West Java. The spread of the COVID-19 virus continues to increase in Indonesia and in a short time has spread to various cities and districts with a high number of cases. Data from the covid19.go.id page shows that confirmed positive cases in Indonesia reached 882,148 as of January 2021. Various efforts have been made by the government to prevent the increase in COVID-19 cases, one of which is the implementation of Large-Scale Social Restrictions (PSBB).

The restriction of human activities during the COVID-19 pandemic provides an opportunity to observe the changes that occur in the environment around us. One of the aspects affected is the air quality in various regions. Based on the Air Pollution Standard Index (ISPU) from the Air Quality Monitoring System (AQMS), it is recorded that the ISPU value tends to decrease and tends to be good due to limited motor vehicle activity (DPLH, 2021 in Mubarak, 2023). Sulistiani et al. (2021) also found a decrease in SO2, NO2, NH3, CO, TSP, and H2S concentrations during the COVID-19

pandemic and the new normal, while O3 concentrations increased compared to normal periods.

Yogyakarta is one of the cities in Indonesia that cannot be separated from the impact of this pandemic. Mobility restrictions, the closure of a number of activities, and changes in human lifestyle have a direct impact on the air quality in the region. Poor air quality can potentially endanger human health and the surrounding environment. Conversely, the better the quality, the air that is inhaled will not harm the health of the body.

Therefore, this paper aims to analyze the air quality in Yogyakarta during the COVID-19 pandemic. The approaches used in this analysis are the Naive Bayes method and the Support Vector Machine (SVM). The Naive Bayes method is a classification method based on simple probability theorems, while SVM is a machine learning algorithm that is popular in classification and regression. This analysis will use air quality data collected during the COVID-19 pandemic in Yogyakarta. This data will include parameters such as Particulate Matter (PM10), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Ozone (O3), and Sodium Dioxide (NO2) particles. Using Naive Bayes and SVM methods, we will analyze this data to understand the pattern of air quality changes in Yogyakarta during the pandemic.

II. Naive Bayes

Naive Bayes is a classification method using probability and statistical methods. The Naive Bayes method assumes that the attributes in the data have no correlation with each other, which makes them independent. The theory of Naive Bayes was discovered by Thomas Bayes in the 18th century and then applied to a simple algorithm called the Naive Bayes Classifier (NBC). NBC aims to predict class objects whose labels are unknown or can predict data that appears in the future (Ginting & Trinanda, 2014). The equation used in Naive Bayes is as follows:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

When:
Y = Particular class
X = Data of an undefined class
P(X|Y) = likelihood of a condition-dependent hypothesis
P(Y|X) = Probability of a class based on a hypothetical condition
P(Y) = Probability of Y
P(X) = Probability of X

III. Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning method in machine learning algorithms that use classification and regression. The SVM algorithm was discovered by Vladimir N. Vapnik in the 1970s, SVM is often used in machine learning applications, especially in image classification. The basic form of SVM is a binary linear classification that wants to identify a boundary between two classes. SVM has a goal to solve problems on a global scale by processing data to find a hyperplane.

Hyperplane is the best separator function (decision boundary) that can be found in the area between the distance

between two hyperplanes in that class. To optimize the search for hyperplanes, the dataset will be split into several training data. The training sample that is closest to the feature space will be used as support vectors to become the boundaries of the hyperplane. Part of the hyperplane will be the assumption that the greater the distance between the two class hyperplanes makes the data have more generalization errors than the classifier.

IV. Evaluation Metrics

In machine learning, one method of the evaluation metrics is based on the confusion matrix. The confusion matrix consists of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).A few terms associated with the confusion matrix are :

- True positive: An instance for which both predicted and actual values are positive.
- True negative: An instance for which both predicted and actual values are negative.
- False Positive: An instance for which predicted value is positive but actual value is negative.
- False Negative: An instance for which predicted value is negative but actual value is positive.

	Predicted 0	**Predicted 1**
Actual 0	**TN**	**FP**
Actual 1	**FN**	**TP**

Table 4.1 Confusion Matrix

All of those data points are going to be used to show the result of accuracy, precision, recall, and f-1 score with the formula stated below :

- $Accuracy = \frac{(TP + TN)}{(TP+TN+FP+FN)}$
- $Precision = \frac{TP}{(TP+FP)}$
- $Recall = \frac{TP}{(TP+FN)}$
- $f - 1\ score = \frac{2\ x\ precision\ x\ recall}{precision + recall}$

A. Data Training

[[54 0]
[10 224]]

Table 4.2 Confusion Matrix for Data Training with Naive Bayes method

From the confusion matrix above, several points can be concluded as follows :

- In the first class (Moderate), there are 54 data that are correctly classified into the second class (True Positives).
- In the second class (Good), there are 224 data that are

correctly classified into the third class (True Positives).

[45 1]
[0 206]]

Table 4.3 Confusion Matrix for Data Training with Support Vector Machine method

From the confusion matrix above, several points can be concluded as follows :

- In the first class (Moderate), there are 45 data that are correctly classified into the second class (True Positives).
- In the second class (Good), there are 206 data that are correctly classified into the third class (True Positives).

B. Data Testing

[13 0]
[3 56]]

Table 4.4 Confusion Matrix for Data Testing with Naive Bayes method

From the confusion matrix above, several points can be concluded as follows :

- In the first class (Moderate), there are 13 data that are correctly classified into the second class (True Positives).
- In the second class (Good), there are 56 data that are correctly classified into the third class (True Positives).

[[17 4]
[1 86]]

Table 4.5 Confusion Matrix for Data Testing with Support Vector Machine method

From the confusion matrix above, several points can be concluded as follows :

- In the first class (Moderate), there are 17 data that are correctly classified into the second class (True Positives).
- In the second class (Good), there are 86 data that are correctly classified into the third class (True Positives).

V. Data and Research Methods

A. Data

The data used in this paper is air quality in Yogyakarta during 2020 period, sourced from Kaggle. There are total of 9 variables in the data, but we only take total of 6 variables with the following details :

- Particulate Matter (PM10) as predictor variable (X variable)
- Sulfur Dioksida (SO2) as predictor variable (X variable)
- Karbon Monoksida (CO) as predictor variable (X variable)
- Ozon (O3) as predictor variable (X variable)
- Natrium Dioksida (NO2) as predictor variable (X variable)
- *Category* as response variable (Y variable)

The response variable or Y variable will determine which

level/category of air quality in Yogyakarta during 2020 period. There are at least 2 categories used in this dataset, including :

- Moderate which is interpreted as number 0
- Good which is interpreted as number 1

B. Research Methods

The research method is divided into several steps, such as data preprocessing, data preparation, Exploratory Data Analysis (EDA), naive bayes modeling, support vector machine modeling, and comparing models. The steps can be explained as follows:

- Preprocessing

  Preprocessing is the initial stage of processing input data before entering the main stage of the process. The data used in the mining process is not always in ideal condition for processing. Sometimes in the data, there are various problems that can interfere with the results of the mining process itself, such as missing values, redundant data, outliers, or data formats that are not compatible with the system. At this stage, all the columns in the dataset are indicated "False", this indicates that there is no missing value in the dataset "Air Quality in Yogyakarta during 2020 period". In addition, we also change categorical variables into numeric variables because the classification method can only process data that is numeric in nature. Therefore, the 2 classes in the "Category" are changed to numbers 0, and 1.

- Data Preparation

  After that, we pre-processed the data where we divided the variables into 2, namely predictor variables and response variables. If the predictor variables and response variables have been set, then the next step is to divide the data into data training and data testing. Training data is used to build a model or classifier. Meanwhile, the testing data is used to test the classifier that has been built to see how accurate the classification results are. Data division is done using the Pareto principle, namely the 80:20 principle. As much as 80% of the data is training data and the remaining 20% is testing data. In this case, the amount of data to be processed is 360 data, so 80% of them, namely 288 data as training data, and the remaining 20%, namely 72 data as test data.

- Exploratory Data Analysis (EDA)

  Exploratory Data Analysis is a method of exploring a lot of data with arithmetic techniques and graphical visuals in summarize the observed data. This method is used to find out patterns of distribution patterns, summarize and visualize data in various forms of graphs, plots and tables with the aim of presenting a

comprehensive statistical summary visually. With EDA we can identify errors by understanding a pattern in the data or anomalies, and finding relationships between variables.

- Naive Bayes
  1. Split the data into training and testing
  2. Predict the model using the training and testing data
  3. Import the libraries of confusion matrix, classification report and naive bayes
  4. Create the confusion matrix for comparing between the training and testing data.
  5. Evaluate the model by calculating the accuracy, precision, recall, and f-1 score
- Support Vector Machine (SVM)
  1. Split the data into training and testing
  2. Predict the model using the training and testing data
  3. Import the libraries of confusion matrix, classification report and Support Vector Machine
  4. Create the confusion matrix for comparing between the training and testing data.
  5. Evaluate the model by calculating the accuracy, precision, recall, and f-1 score
- Comparison models

The result of the confusion matrix, namely accuracy, precision, recall, and f-1 score are going too be compared between the naive bayes and support vector machine to determine the best model of the training and testing data.

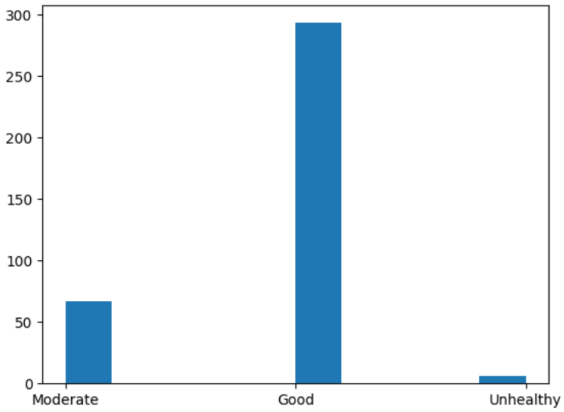
VI. Results and Discussions

Classification reports reveal performance evaluations of the models we work on datasets, focusing on the accuracy of data training and data testing.

A. Exploratory Data Analysis (EDA)

a) Check the Imbalance

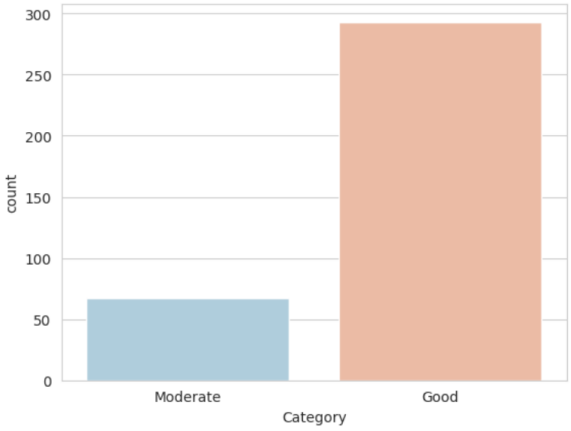
One way to check the balance of data is to use a histogram. Data is declared imbalance if there is a ratio difference of more than 1: 10 in each category.



Picture 6.1 Histogram for check the imbalance of data

From the results of the histogram above, it is obtained that the difference in the number between categories is not significant, namely with a ratio of around 1: 10: 50 . So, we need to remove the category that has a small number, namely the "unhealthy" category.

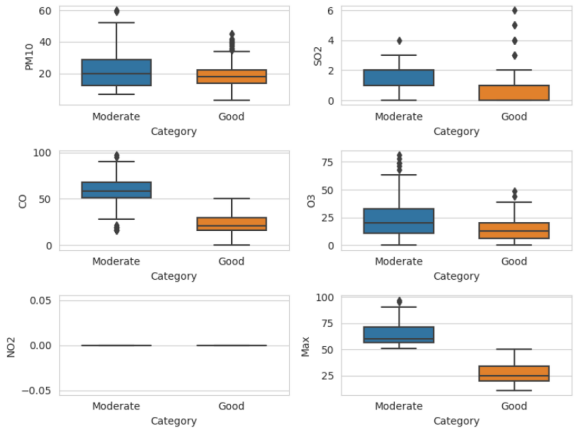
Picture 6.2 Histogram for check the imbalance of data



Based on the result, after we remove the "unhealthy" category, the data is balanced with a ratio of about 1: 5.

b) Check the Outliers

An outlier is a value or point that differs substantially from the rest of the data. We can identify outliers by utilizing Box-Plot visualization. Through Boxplot visualization, we can objectively see which data is an outlier or not. Boxplot utilizes the quartile values of the data, as well as the minimum and maximum values.



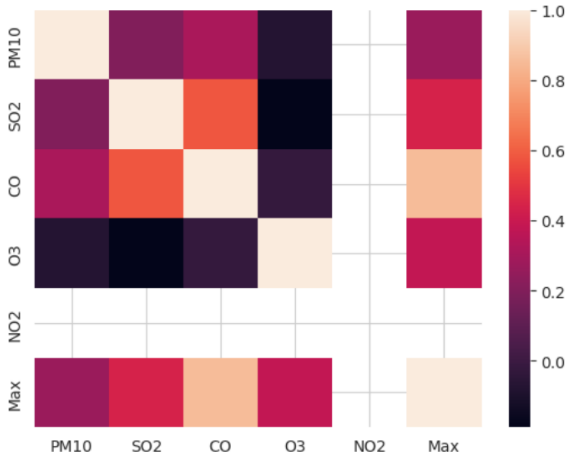
Picture 6.3 Boxplot for check the outliers of data

Based on the results, there are outliers in our data. However, because the data used air quality data, these outliers can be said to be natural outliers. Natural outliers are outliers that occur due to natural factors or real events that are not caused by data processing errors. Therefore, outliers in the data do not need to be removed.

c) Check the Correlations

A heatmap is a visualization or mapping that displays data with different color representations. Usually, the higher the number of a data group, the darker the

color. With heatmaps, we can easily see the correlation between existing parameters.



Picture 6.4 Heatmap for check the correlations of data

Based on the results above, it can be concluded that 03 has a poor correlation with the variables PM10, SO2, and CO.

B. Naive Bayes Model

The classification report presents an analysis of the performance of our Naive Bayes model on the dataset, with a specific focus on the level of accuracy achieved by the model on data training and data testing.

	Precision	Recall	f1-score	Support
0	1.00	0.84	0.92	64
1	0.96	1.00	0.98	224
Accuracy			0.97	288
Macro Avg	0.98	0.92	0.95	288
Weighted Avg	0.97	0.97	0.96	288

Table 6.1 Classification Report for Data Training

	Precision	Recall	f1-score	Support
0	1.00	0.81	0.90	16
1	0.95	1.00	0.97	56
Accuracy			0.96	72
Macro Avg	0.97	0.91	0.94	72
Weighted Avg	0.96	0.96	0.96	72

Table 6.2 Classification Report for Data Testing

The aspect being discussed is about the accuracy between data training and data testing. From the data that have been presented, it showed that the accuracy between data training and data testing has the same value. The level of accuracy obtained for data training is 0.97 and

data testing is 0.96. This demonstrates the model's ability to make accurate predictions on both sets.

The results of this high accuracy score show the level of effectiveness of our model in correctly classifying sample data into their respective categories. This means that the model has successfully learned and generalized patterns from data training to make accurate predictions on new events that are not visible in data testing. In addition, balanced accuracy across all classes indicates the model's ability to handle different classes equally well. Our Naive Bayes model shows consistency and reliability in its classification performance, which is evidenced by the results of a high accuracy score.

The achieved accuracy score of 0.97 and 0.96 also demonstrates the robustness and reliability of our Naive Bayes model, giving us confidence in its ability to make accurate predictions in real-world scenarios.

C. Support Vector Machine Model

	Precision	Recall	f1-score	Support
0	1.00	0.98	0.99	46
1	1.00	1.00	1.00	206
Accuracy			1.00	252
Macro Avg	1.00	0.99	0.99	252
Weighted Avg	1.00	1.00	1.00	252

Table 6.3 Classification Report for Data Training

	Precision	Recall	f1-score	Support
0	0.94	0.81	0.87	21
1	0.96	0.99	0.97	87
Accuracy			0.95	108
Macro Avg	0.95	0.90	0.92	108
Weighted Avg	0.95	0.95	0.95	108

Table 6.4 Classification Report for Data Testing

The aspect that is specifically discussed is about the accuracy between data training and data testing. From the data that have been presented, it showed that the accuracy between data training and data testing has the same value. The level of accuracy obtained for data training is 1.00 and data testing is 0.95.

The high level of accuracy achieved in data training and data

testing demonstrates the effectiveness of the SVM algorithm in capturing and studying complex patterns and decision constraints from data training. This acquired knowledge enables our SVM models to make precise predictions on unseen data testing instances with a high degree of accuracy. Balanced accuracy across all classes indicates the model's ability to handle different classes equally well. This shows that our SVM model performs well across different classes, ensuring unbiased and accurate predictions in various scenarios.

The achieved accuracy score of 0.95 and 1.00 also demonstrates the robustness and reliability of our Naive Bayes model, giving us confidence in its ability to make accurate predictions in real-world scenarios.

D. Comparison of Naive Bayes and Support Vector Machine

Comparison	Training			Testing		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Naive Bayes	0.97	1.00	0.84	0,96	1.00	0.81
		0.96	1.00		0.95	1.00
Support Vector Machine	1.00	1.00	0.98	0.95	0.94	0.81
		1.00	1.00		0.96	0.99

Table 6.5 Comparison of models

The high accuracy achieved in data training and data testing indicates that our Naive Bayes model exhibits strong predictive capabilities and has the potential for successful application in practical applications requiring accurate classification.

On the other hand, our SVM model has shown great accuracy in both phases. This demonstrates a very strong predictive ability that allows our SVM model to be applied with confidence in practical applications where accurate classification is required.

VII. Conclusion

Based on the results we achieve from the dataset of Air Quality in Yogyakarta during the Covid-19 pandemic using Naive Bayes and Support Vector Machine. We can conclude that the best classification method to use on determining the predicted air quality in Yogyakarta is the Support Vector Machine model. The SVM model has a more accurate result than the Naive Bayes model using the same dataset. We can see the SVM model has almost a perfect f1-score for each category of air quality with an accuracy on data training is 1.00 and an accuracy on data testing is 0.95 which is higher than the Naive Bayes model.

References

1. Novianto, H., Azis, M. M., & Arini, H. M. (2022). Analisis Perubahan Sistem Kualitas Udara Kota Yogyakarta pada Masa Pandemi COVID-19. *Jurnal Rekayasa Proses*.
2. Rushayati, S. B., Hermawan, R., Setiawan, Y., Wijayanto, A. K., Prasetyo, L. B., & Permatasari, P. A. (2020). Pengaruh Pola Pemanfaatan Ruang Terbuka Hijau terhadap Dinamika Perubahan Kualitas Udara Akibat Pandemi Covid-19. *Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan (Journal of Natural Resources and Environmental Management)*, *10*(4), 559-567.
3. Mubarak, A. M. N. (2023). Pengaruh Pemberlakuan Pembatasan Kegiatan Masyarakat Pada Masa Pandemi Covid-19 Terhadap Konsentrasi So2, No2, Dan Pm2, 5 Di Kabupaten Bantul, Daerah Istimewa Yogyakarta.
4. Sulistiani I, Partama IGY, Kaler Surata SP, Sumantra IK. 2021. Dinamika kualitas udara ambien selama masa pandemi covid-19 di kawasan indonesia tourism development corporation nusa dua bali. ECOTROPHIC: Jurnal Ilmu Lingkungan (Journal of Environmental Science). 15(1):124. doi: 10.24843/EJES.2021.v15.i01.p11.
5. Pristiyono, Ritonga, M., Ihsan, M. A. A., Anjar, A., & Rambe, F. H. (2021, February). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012045). IOP Publishing.
6. Loelianto, I., Thayf, M. S. S., & Angriani, H. (2020). Implementasi Teori Naive Bayes Dalam Klasifikasi Calon Mahasiswa Baru Stmik Kharisma Makassar. *SINTECH (Science and Information Technology) Journal*, *3*(2), 110-117.