The 1st International Conference On Mathematics Education And Technology (ICOMET) 2022
pp. 115-123 (2022)
Penerbit: Universitas Islam Malang

CLASSIFICATION OF HUMAN DEVELOPMENT INDEX (HDI) ON THE ISLAND OF SUMATRA IN MULTI-CLASS DATASET WITH ARTIFICIAL NEURAL NETWORK (ANN) METHODS

Muhammad Fathurrahman[1*]

[1]Informatics Engineering Study Program, Faculty of Computer Science, University of Dian Nuswantoro
Jl. Imam Bonjol No. 207, Semarang, Indonesia

ABSTRACT

Uneven development in Indonesia as indicated by the Human Development Index (HDI) in very different districts/cities. High to very high HDI categories are only available in major provinces/cities in Indonesia, because these large agencies/cities have complete health, education, and economic facilities. This study aims to compare 2 classification methods namely, Support Vector Machine (SVM) and Artificial Neural Network (ANN) to obtain the best of all for HDI data classification method in Sumatra Island regencies/cities in 2020. The study gave the result that the precision values of the Artificial Neural Network (ANN) and the Support Vector Machine (SVM) method were 97.4% and 53.25%, respectively. Therefore, an Artificial Neural Network method can be used to rank the human development index (HDI) on the island of Sumatra, because the accuracy values of Artificial Neural Network (ANN) methods outperform those of Support Vector Machine (SVM) methods.

Keywords: HDI, Artificial Neural Network, Support Vector Machine, Accuracy

1. Introduction

According to [1], man is the real wealth of the country. Therefore, the main goal of development is to create an environment where people can enjoy a long, healthy and productive life. Simple, but often forgotten by various short-term busyness to collect treasures and money [1].

The concept of human development is very different from the concept of development. The concept of human development includes a very wide range of dimensions. According to [1], human development is one of the processes of expanding choices for society (a process of enlarging people's choices). Development policies that do not pay attention to improving people's living standards will be left behind compared from one region to another. Among the 17 Sustainable Development Goals (SDGs) there are various goals related to human development, and through the Sustainable Development Goals (SDGs), human development goals and targets continue to evolve. In an increase, of course, a measure is needed in the calculation. The United Nations Development Programme (UNDP) uses three dimensions to measure human development: a long and healthy life, knowledge and a decent standard of living. Longevity and healthy living are the dimensions of health, knowledge is the dimension of education, and a decent standard of living is the economic dimension.

Uneven development in Indonesia as indicated by the highly diverse Human Development Index (HDI) of the region/city [2]. According to [3], HDI is divided into 4 categories: low is less than 60, medium is less than equal to 60 and less than 70, high is less than equal to 70 and less than 80, and very high is greater than equal to 80. High to very high HDI indices are only available in major districts/cities in Indonesia. Because, these big districts/cities have adequate health, education, and financial facilities.

Data mining is the manual discovery of added value in the form of unknown knowledge from a very large database. Data mining has a classification algorithm that is an operation that separates several entities into different classes [4][5]. This study aims to compare the accuracy of Human Development Index (HDI) data between two classification methods, namely Artificial Neural Network (ANN) and Support Vector

Machine (SVM) to obtain the best classification method.

Some previous studies related to the "Classification of Human Development Index (HDI) in Central Java Province using the k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) Methods" by [6] resulted in a comparison of classification accuracy of 91.64% in k-Nearest Neighbour (k-NN) method and 95.36% in Support Vector Machine method (SVM), it is concluded that the right method for HDI Classification in Central Java Province is a Support Vector Machine (SVM) with an accuracy of up to 95.36%.

In addition, another study on "Comparison of Support Vector Machine (SVM) and Neural Network (NN) Methods for the Highest Prediction Accuracy Rate of Stock Price" by [7] obtained prediction accuracy using the RMSE method on Neural Network (NN) model of 0.503 +/- 0.009 (micro 503) and Support Vector Machine (SVM) model of 0.477 +/- 0.008 (micro 0477), then it is concluded that Neural Networks (NN) have a greater predictive trend than Support Vector Machine (SVM) models.

Then in the research conducted [8] on "Performance Classification of Regional Drinking Water Companies (PDAMs) in Indonesia using comparative Artificial Neural Network (ANN) and Support Vector Machine (SVM) models", it was explained that Artificial Neural Network (ANN) have an accuracy value of 83.93% and a precision value of 86.36% while the Support Vector Machine (SVM) has an accuracy value of 82.14% and a precision value of 80%. So it can be concluded that the Artificial Neural Network (ANN) model it can be used to predict the performance of Regional Drinking Water Companies (PDAMs) in Indonesia. However, there is no research on case studies of the Classification of Human Development Index (HDI) with the Artificial Neural Network (ANN) method and the Support Vector Machine (SVM), so this study will try to examine the classification of the Human Development Index (HDI) using the Artificial Neural Network (ANN) method and the Support Vector Machine (SVM). This study aims to compare the level of accuracy of the Artificial Neural Network (ANN) and Support Vector Machine (SVM) models.

2. Research Methodology

In this study, experiments were carried out with two kinds of classification methods, namely Artificial Neural Network (ANN) and Support Vector Machine (SVM). The classification is carried out using the RapidMiner Studio Version 9.3 software. The results of the classification of the two methods will be compared to obtain the best method based on the highest accuracy value.

The stages of this study are shown in Figure 1. Modeling using Datasets. This dataset is needed as training data and testing data for classification. Each sample dataset has attributes and classes or labels. The Artificial Neural Network (ANN) and Support Vector Machine (SVM) algorithms use the same Dataset with attributes and labels to generate the model to be used. The model will be tested using the k-Fold Cross Validation method 10 times. Verification is represented by the Confusion Matrix to calculate the accuracy of the model as a whole.

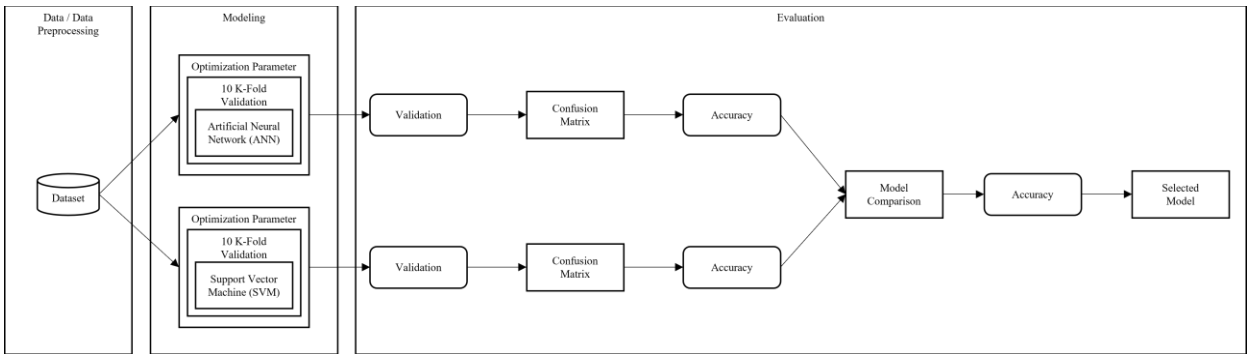

Figure 11. *Research Methodology*

A. Data

This study used secondary data sourced from various survey results conducted by the Central Statistics Agency (BPS) with an observation unit of 154 regencies/cities on the island of Sumatra. The time period used is 2020 [3]. In this study,

there is a variable HDI status as a label consisting of 3 classes and variables life expectancy (AHH), Expected Year School (EYS) or expectations for length of school, Mean Year School (MYS) or average length of schooling, and adjusted per capita expenditure as attributes [9].

B. Data Preparation

Data Preparation or often also called Data Preprocessing is a process of preparing data before being processed using analytical methods in data mining [7]. Data preparation can be in the form of selection, cleaning, or transformation of data into the desired form. A set of datasets in data mining consists of data objects referred to as samples / examples / tuples and data fields or attributes that represent the characteristics / features of the objects. Special attributes that are targets in classification analysis are referred to as labels.

C. Modelling

This process contains a classification algorithm from pre-prepared data. At this stage, the formation of a model that can distinguish data classes will be carried out. Data is divided into two types of datasets derived from preprocessing data, namely training data and testing data. Training data is a dataset used to build models. Testing data is a dataset that is used to calculate the performance of the formed model. The performance of the model is measured by comparing the actual data label with the predicted result data label [10]. The formation of both types of datasets can use the k-Fold Cross Validation method which is a form of resampling that takes several samples from the entire observation and makes it a training data for the model. After the two types of datasets are formed, classification is carried out using the Artificial Neural Network and Support Vector Machine algorithms.

- *Artificial Neural Network (ANN)*

Artificial Neural Network (ANN) is a type of artificial neural network that models the nervous system of the brain to carry out pattern recognition tasks, specifically classification [8]. ANN's Backpropagation algorithm is as follows:

1. Initialize all weights and define activation functions and learning rates.
2. Network activation by applying inputs and outputs.
3. Updates the weight by correcting the weight when the error is returned in the direction in which the output signal returns.
4. Iterations are performed until an error in the weight is less than the desired value.

Each neuron in the network calculates the weight of the *wij* sum of the input signal *p yi*, for *i* = 0.1,2, ... *n* on the hidden layer, and then apply a nonlinear activation function to generate a *uj* output signal. The form of this function is as follows:

$$uj = \sum_{i=0}^{n} w_{ij} y_i \qquad (1)$$

- *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a discriminatory grouping method using a hyperplane separator. SVM analyzed the most optimal hyperplane in describing the classification of groups [5][6]. As for the linear function of the SVM [6]:

$$g(x) = sign(f(x)) \qquad (2)$$

And

$$f(x) = w^T x + b \qquad (3)$$

Some of the parameters used in parameter optimization are as follows:

1. The kernel is a function used to form a separator hyperplane function.
2. C indicates how likely it is to avoid misclassification.
3. Gamma defines how close an affordable dataset is as a support vector to its hyperplane.

Using kernel functions as basic functions is one of the methods used to address nonlinear data. The RBF kernel formula is as follows [11]:

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2.\sigma^2}\right) \qquad (4)$$

D. Evaluation

Evaluation is carried out to choose which classification method can produce the highest level of accuracy by paying attention to the Confusion Matrix. The Confusion Matrix is a tool to find out the extent to which a classifier can recognize or predict data classes [12]. The Confusion Matrix is the size of an *mx* where *m* is the number of classes. The column section is populated by the actual label for each class, while the row section is populated by the predicted result class label. As in table 1.

Table 1. Confusion Matrix

Prediction	Actual		Total
	+	-	
+	True Positive (TP)	False Negative (FN)	P'
-	False Positive (FP)	True Negative (TN)	N'
Total	P	N	P + N = P' + N'

Accuracy is a percentage of the number of tuples in the testing data that are correctly classified by the classifier.

$$Accuracy = \frac{TP + TN}{P + N} \quad (5)$$

$$Error\ Rate = 1 - Accuracy\ = \frac{FP + FN}{P + N} \quad (6)$$

Other measures that can be used in evaluating classifications are precision and recall. Precision is the percentage of tuples labeled as positive is true in fact or also referred to as a measure of certainty. Recall is a percentage of positive tuples that are labeled as positive or also called a measure of completeness.

$$Precision = \frac{TP}{TP + FN} = \frac{TP}{P'} \quad (7)$$

$$Recall = \frac{TP}{TP + FP} = \frac{TP}{P} \quad (8)$$

3. Results and Discussion

The structure of the data used in this study can be seen in Table 2.

Table 2. Data Structure

Attribute	Definition	Category	Scale
HDI_Status	Human Development Index	1 = Medium 2 = High 3 = Very High	Polinominal
AHH	Life Expectancy	-	Ratio
EYS	Old School Expectations	-	Ratio
MYS	Average Length of Schooling	-	Ratio
Expenses	Adjusted per capita expenses	-	Ratio

In this study, experiments were carried out with two kinds of classification methods, namely Artificial Neural Network (ANN) and Support Vector Machine (SVM). The classification is carried out using the RapidMiner Studio Version 9.3 software. The results of the classification of the two methods will be compared to obtain the best method based on the highest accuracy value. For modeling and modeling purposes, the data is divided into training data and testing data. In this case, the 10-folds method is used to perform cross-validation. The data is divided into 10 groups at random, 9 groups will be training data and the other 1 group will be testing data. This process is carried out 10 times until each group has been a testing data.

A. Descriptive Analysis

The average HDI on the island of Sumatra in 2020 was 70.69 with a high category. The highest HDI of 85.41 is in Banda Aceh City while the lowest HDI of 61.09 is in the Mentawai Islands Regency. The following is the number of classifications of regencies/cities on the island of Sumatra.

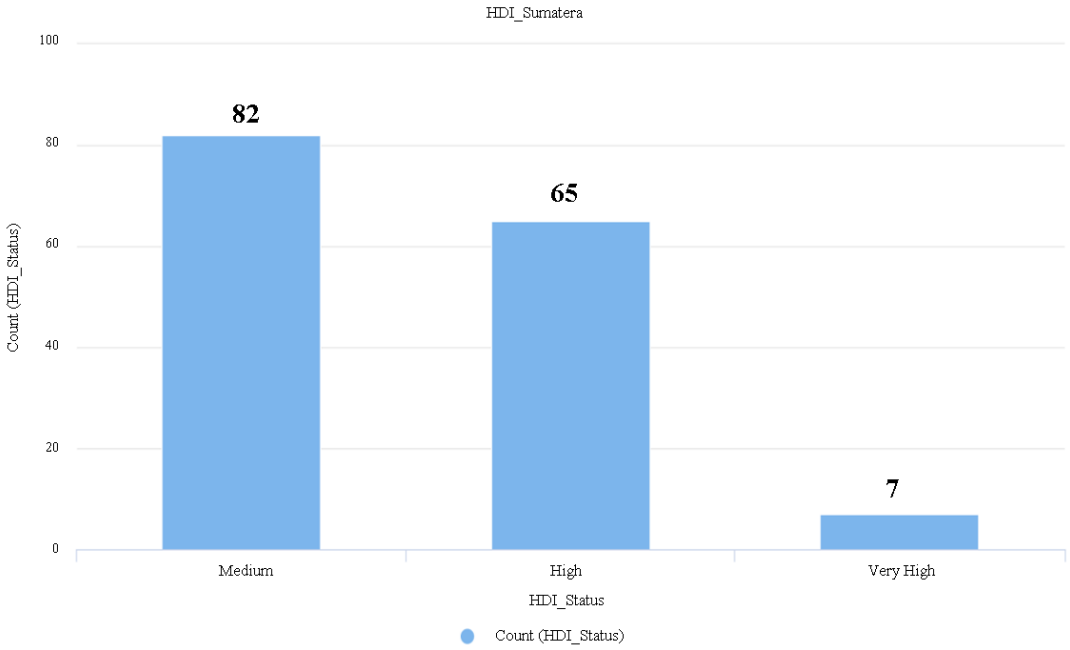

Figure 22. Classification Diagram of Regencies/Cities on the Island of Sumatra

Based on figure 2 above, regencies/cities on the island of Sumatra in 2020 which are classified as medium category there are 82 regencies/cities, 65 regencies/cities are classified as high and 7 regencies/cities are included in the very high category.

B. Artificial Neural Network (ANN)

The artificial neural network generated uses three layers consisting of an input layer, hidden layer, and output layer as shown in Figure 3.

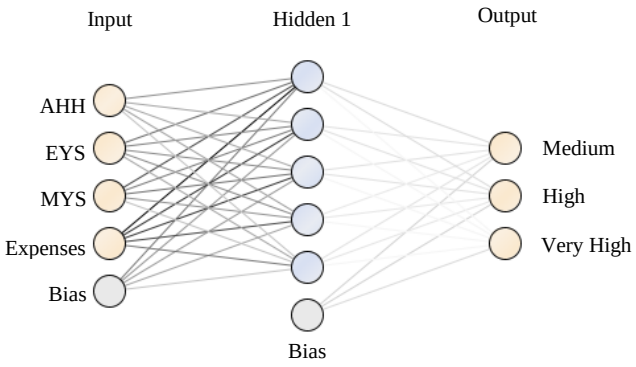

Figure 3.3 Artificial Neural Network

The first layer is the input layer consists of 5 nodes, namely 4 predictor attribute nodes (AHH, EYS, MYS, and Expenses) and 1 bias node. The second layer is the hidden layer consists of 5 nodes and 1 bias node. The third layer is an output layer that has 3 nodes that represent the classes in the Status_IPM attributes, namely medium, high, and very high.

The Artificial Neural Network (ANN) method used in this study was determined by trial using the learning rate parameter values used were 0.010 until 0.099 and momentum parameter values used were 0.1 until 0.9. The results of such experiments are shown through Table III.

Table 3. ANN By Parameters

Method	Parameter		Accuracy
	Learning Rate	Momentum	
ANN	0.010	0.1	0.826
	0.206	0.1	0.929
	0.402	0.1	0.962
	0.598	0.1	0.942
	0.794	0.1	0.916
	0.990	0.1	0.955
	0.010	0.3	0.858
	0.206	0.3	0.948
	0.402	0.3	0.948
	0.598	0.3	0.936
	0.794	0.3	0.940
	0.990	0.3	0.961
	0.010	0.4	0.884
	0.206	0.4	0.935
	0.402	0.4	0.955
	0.598	0.4	0.935
	0.794	0.4	0.955
	0.990	0.4	0.934
	0.010	0.4	0.963
	0.206	0.4	0.941
	0.010	0.6	0.890
	0.206	0.6	0.935
	0.402	0.6	0.948
	0.598	0.6	0.935
	0.794	0.6	0.928
	0.990	0.6	0.960
	0.010	0.7	0.903
	0.206	0.7	0.948
	0.402	0.7	0.909
	0.598	0.7	0.929
	0.794	0.7	0.948
	0.990	0.7	0.973
	0.010	0.9	0.942
	0.206	0.9	0.967
	0.402	0.9	0.968
	0.598	0.9	0.936
	0.794	0.9	0.974
	0.990	0.9	0.960

The results of testing training data using the Artificial Neural Network (ANN) method have an accuracy of 97.4%. Based on the Confusion Matrix table presented in Table 3, out of 154 regency/city data on the island of Sumatra, there are 81 regencies/cities that are predicted to be precisely included in the medium HDI category, 63 regencies/cities predicted to be appropriately included in the high HDI category, and 6 regencies/cities predicted to be appropriately included in the very high HDI category. Meanwhile, there are 2 regencies/cities predicted to have a moderate HDI but it turns out to be included in the high HDI, and 1 regency/city is predicted to be a high HDI but it turns out to be included in the medium HDI, and 1 regency/city is predicted to be a high HDI but it turns out to be included in the very high HDI.

Based on the picture above, regencies/cities on the island of Sumatra in 2020 which are classified as medium category there are 82 regencies/cities, 65 regencies/cities are classified as high and 7 regencies/cities are included in the very high category.

Table 4. Confusion Matrix Artificial Neural Network

	True Medium	High True	Very High True
Pred. Keep	81	2	0
Pred. Tinggi	1	63	1
Pred. Very High	0	0	6

C. Support Vector Machine (SVM)

The Support Vector Machine (SVM) method used in this study was determined by trial using the Radial Basis Function (RBF) kernel parameter [11] and the Gamma and C parameter values used were 0.1, 1, and 10. The results of such experiments are shown through Table 5.

Table 5. SVM By Parameters

Kernel	Parameter		Accuracy
	Gamma	C	
RBF	0.1	0.1	0.534
		0.1	0.534
		0.1	0.532
	1	1	0.533
		1	0.533
		1	0.532
	10	10	0.534
		10	0.528
		10	0.533

The results of testing training data using the Support Vector Machine (SVM) method have an accuracy of 53.25%. Based on the Confusion Matrix table presented in Table 5, out of 154 regency/city data on the island of Sumatra, there are 82 regencies/cities that are predicted to be precisely included in the medium HDI category, while none of the predicted regencies/cities are correctly included in the high and very high HDI category. Meanwhile, there were 65 regencies/cities predicted to have a medium HDI but it turned out to be included in the high HDI and 7 regencies/cities were predicted to be medium HDI but it turned out to be included in the very high HDI.

Table 6. Confusion Matrix Support Vector Machine

	True Medium	High True	Very High True
Pred. Medium	82	65	7
Pred. High	0	0	0
Pred. Very High	0	0	0

D. Method Comparison

Based on the results of the analysis of the two methods, the accuracy values of each method are compared, then the results of the comparison of the accuracy of each model can be seen in Table VII.

Table 7. Accuracy Comparison

Method	Precision	Recall	Accuracy
Artificial Neural Network	98.17%	93.80%	97.4%
Support Vector Machine	17.75%	33.33%	53.25%

From table 6 above, it can be seen that the accuracy value of the Artificial Neural Network (ANN) method is higher, which is 97.4% compared to the accuracy value of the Support Vector Machine (SVM) method of 53.25%. Therefore, the classification results of the Artificial Neural Network (ANN) method can be applied to classify the Human Development Index (HDI) on the island of Sumatra.

4. Conclusion

This study aims to compare the Artificial Neural Network and Support Vector Machine (SVM) methods to classify the Human Development Index (HDI) on the island of Sumatra. Based on the results of the analysis in this study, the accuracy value for the Artificial Neural Network (ANN) method was 97.4%, while the accuracy value for the Support Vector Machine (SVM) method was 53.25%. From the two accuracy values in each of these methods, it can be concluded that the Artificial Neural Network (ANN) method can be used to classify the Human Development Index (HDI) on the island of Sumatra because the accuracy value in the Artificial Neural Network (ANN) method is higher than the accuracy value in the Support Vector Machine (SVM) method.

5. References

United Nations Development Programme, *Human Development Report 2019: beyond income, beyond averages, beyond today*. 2019.

M. Y. Matdoan, "Penerapan Analisis Cluster Dengan Metode Hierarki Untuk Klasifikasi Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia," *Statmat J. Stat. Dan Mat.*, vol. 2, no. 2, p. 20, 2020, doi: 10.32493/sm.v2i2.4740.

Badan Pusat Statistik, "Indeks Pembangunan Manusia," *Badan Pusat Statistik*, 2020. .

I. Irmawati, Z. Zainudin, and Y. Yuyun, "Data Mining Untuk Penentuan Model Kelulusan Murid Sma Pada Perguruan Tinggi Negeri; Studi Kasus Di Iain Bone," *JIKO (Jurnal Inform. dan Komputer)*, vol. 3, no. 2, pp. 113–118, 2020, doi: 10.33387/jiko.v3i2.1800.

N. Kranjčić, D. Medak, R. Župan, and M. Rezo, "Support Vector Machine accuracy assessment for extracting green urban areas in towns," *Remote Sens.*, vol. 11, no. 6, 2019, doi: 10.3390/rs11060655.

F. Fauzi, "K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah," *J. Mipa*, vol. 40, no. 2, pp. 118–124, 2017.

R. H. Kusumodestoni and S. Sarwido, "Komparasi Model Support Vector Machines (Svm) Dan Neural Network Untuk Mengetahui Tingkat Akurasi Prediksi Tertinggi Harga Saham," *J. Inform. Upgris*, vol. 3, no. 1, 2017, doi: 10.26877/jiu.v3i1.1536.

P. R. Sihombing, "Perbandingan Metode Artificial Neural Network (ANN) dan Support Vector Machine (SVM) untuk Klasifikasi Kinerja Perusahaan Daerah Air Minum (PDAM) di Indonesia," *J. Ilmu Komput.*, vol. 13, no. 1, p. 9, 2020, doi: 10.24843/jik.2020.v13.i01.p02.

M. W. Talakua, Z. A. Leleury, and A. W. Talluta, "Analisis Cluster Dengan Menggunakan Metode Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014 Cluster Analysis By Using K-Means Method for Grouping of District / City in Maluku Province Industrial Based on Indicators of Maluku Dev," *J. Ilmu Mat. dan Terap.*, vol. 11, pp. 119–128, 2017.

N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, no. February 2019, p. 103676, 2020, doi: 10.1016/j.compedu.2019.103676.

F. Fauzi, M. Yamin, and T. Wahyu, “Klasifikasi Indeks Pembangunan Manusia Kabupaten / Kota Se-Indonesia dengan Pendekatan Smooth Support Vector Machine (SSVM) Kernel Radial Basis Function (RBF),” *Semin. Nas. Pendidikan, Sains dan Teknol. Fak. Mat. dan Ilmu Pengetah. Alam Univ. Muhammadiyah Semarang*, pp. 88–97, 2017, [Online]. Available: https://jurnal.unimus.ac.id/index.php/psn12012010/article/view/2986.

F. Rahmad, Y. Suryanto, and K. Ramli, “Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012076.