JURNAL FOURIER | April 2021, Vol. 10, No. 1, 29-37 ISSN 2252-763X
DOI: 10.14421/fourier.2021.101.29-37 E-ISSN 2541-5239

# Rough Set Theory for Dimension Reduction on Machine Learning Algorithm

Rani Nuraeni, Sugiyarto Surono

*Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia.*

*Korespondensi; Sugiyarto Surono, Email: sugiyarto@math.uad.ac.id; Rani Nuraeni, Email: rani1700015051@webmail.uad.ac.id*

## Abstract

**Dimension reduction is a method applied in machine learning sector to significantly improve the efficiency of computational process. The application of high number variables in certain dataset is expected to be able to provide more information to analyze. However, this application of high number of variables will impacted on the computational time and weight linearly. Dimension reduction method serves to transforming the high dimension data into much lower dimension without significantly reduce the initial information and characteristic provided by the initial data. Core and Reduct is a method acquired through the concept of Rough Set. Dataset functioning as the input and output on Machine Learning can be perceived as informational system. The objective of this research is to determine the impact of the dimension reduction application on machine learning algorithm on the reduction of computational time and weight. Core and Reduct will be applied in few popular machine learning method such as Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN). This research applied on 5 UCI machine learning dataset which are Iris, Seeds, Years, Sonar, and Hill-Valley. Furthermore, Machine learning metrics such as Accuracy, Recall, Precision, and F1-Score also observed and compared. This research resulted in the conclusion that Core and Reduct is able to decrease the computational time up to 80% and maintain the value of each evaluation model.**

**Keywords:** Core and Reduct; Dimension reduction; Machine Learning; Machine Learning Metrics; Rough Set Theory.

## Introduction

Nowadays, big data is already become a familiar variable for software developer with wide scale of projects. The data-based application is very important for structurally and systematically treating, keeping, and managing all kinds of information in the form of data [4]. Big data is a technology in the world of informational science which enabling the process of data computing, saving, and analyzing on many forms of data, very high numbers of data, and the addition of data in a very fast manner [11]. Compared to the previous technology, the application of Big Data resulting in the faster computing and analyzing processes for the high number of data [8]. The importance of Big Data is not only revolves around the number of the acquired data, but also on how to process the internal and external data. Because of that, to process these type of data, it requires the technology which have the ability to compute the given data in self-learning manner [10].

The technology which able to process the given data in self-learning manner is machine learning [8]. Machine learning is a study focuses on the learning process of computer to behave like human way of thinking by independently learn from time to time [10]. Nowadays, Machine learning emerges as one of the most applied sub-study which derived from Artificial Intelligence (AI) Technology. The main objective of machine learning is to improve the machine ability to learn new information from the given data and develop it to solve certain problems [12] [13].

In this paper, dimension reduction will be applied to machine learning algorithms such as SVM, Logistic regression, and KNN. These three strategies are being used to see how dimensional reduction works on the dataset in question. Combining it with Machine Learning Algorithms has the purpose of extracting key variables from the dataset before applying machine learning. Dimension reduction on datasets works by converting high-dimensional datasets into low-dimensional datasets while preserving

 

critical information. When the dataset is applied to the three machine learning algorithms, this has a major impact on the computation time.

The availability of the existing data contains many form of dimensions and features. This allows the existence of the data complexity containing many noise, outliner, missing value and irrelevant information which can be classified as high dimension data. Applying all of the features for certain machine learning algorithm is considered as the inefficient way of processing because not all of the information provided by the data are relevant to the objective of the algorithm itself. Therefore, to improve the accuracy, it requires the dimension reduction application on initial computing process.

There are many dimension reduction technique which already developed by researchers. In Tanet al (2006) and Prasetyo (2012), dimension reduction in linear way such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) emerged as two-dimension reduction methods which have wide-range of function. Another methods such as Factor Analysis, Locally Linear Embedding (LLE) (Roweis and Saul, 2002), Multidimensional Scaling (MDS) (Cox and Cox, 1994), FastMap (Faloutsos and Lin, 1995), ISOMAP (Tenenbaum, 1998), and few other method can also become an alternative method of dimension reduction for certain cases.

Another previous research on Rough Set Theory (RST) worked by Hendrik Fery Herdiyatmoko with the objective to decrease the parameter or fire attributes resulting in the decrease of complexity on data analysis of fire accident. The reduction output of these attributes are applied as the basic method for determining the evacuation route during fire disasters. The output result also used as the information for early detection of fire disaster inside the certain building which alarmed through wireless sensor network.

Andika Prajana, [1] also studied a research on the Rough Set Theory for forecasting. This research applied RST to analyze the data of students score for determine certain pattern. This pattern then functioned for predicting the rate of student graduation in the next test.

Based on the previous research which applied RST as the dimension reduction method, it can be concluded that Core and Reduct is one of the method of dimension reduction acquired from the concept of Rough Set Theory [2] [6]. Dataset functioning as the input and output on Machine Learning can be perceived as informational system. Based on Rough Set Theory, the core of dataset can be acquired through the reduction of these informational system. This research is studied to determine the result of the application of dimension reduction in Machine Learning Algorithm in decreasing the computational time and weight. Core and Reduct will be applied in a few popular method of Machine Learning such as Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) [13] [14]. All combinations of indiscernibility relations will be verified against each other with total indiscernibility relation obscurity to see which variables are preserved in the reduct process. Then there's the core procedure, which involves cutting the variables from the reduction results into parts. The variables that will be utilized as input to the machine learning algorithm are the outcome of the core. This method only developed for the data which considered as numerical data as is how the previous dimension reduction through RST were developed.

## Methodology

### Rough Set Theory (RST)

Rough Set Theory (RST) is a mathematical method developed by Zdzislaw Pawlak in 1982. This method already developed to process the uncertainty of the information which contain inaccuracy and incompleteness [12]. To handle the indiscernibility problem between variables in certain set, this method emerge as one of the most beneficial method for clustering, especially when the applied concept is uncertain, and the involved data are still relevantly unclear in decision-making process. The concept such as the relation of indiscernibility can be applied for the parameter of classification process.

Reduction process in the Rough Set theory can be worked by applying Core and Reduct method.

#### *Definition 2.1 (Information system)*

In rough set, certain data is represented as a table where its rows illustrate cases, events, patients, or certain objects. On the other hand, the columns inside the table are illustrated as variable attributes,

observation, properties, etc. This table also can be described as Information Systems (IS) with the definition as follows:

$$IS = (U, P) \tag{1}$$

Where U is described as Universe, an infinite set which are not unfulfilled from objects, and A is an infinite set which not unfulfilled from attributes, where:

$$a: U \rightarrow V_a \tag{2}$$

for each of $a \in A$, set of $V_a$ defined as set of values of a.

*Definition 2.2 (Indiscernibility)*
For $IS = (U, P)$ is an information system and $B \subseteq P$, the objects indiscernibility based on attributes of B symbolized as $IND_{IS}(B)$, can be defined as:

$$IND_{IS}(B) = \{(x, x') \in U^2 | \forall_a \in B\ a(x) = a(x')\} \tag{3}$$

$IND_{IS}(B)$ is described as B-indiscernibility relation. $IND_{IS}(B)$ also considered as an equivalence relation. If $(x, x') \in IND_{IS}(B)$, then object x and x are the indiscernible objects towards each other in attribute B. The classifications which equivalent with B-indiscernibility relation are notated as $[x]_B$ and defined as equivalent class.

*Definition 2.3 (Reduct)*
For $IS = (U, P)$ is information system, $B \subseteq P$ and for $a \in B$, then a is dispensable in attribute B when $IND_{IS}(B) = IND_{IS}(B) - \{a\}$. On the contrary, if a is indispensable, then a is very important in attribute B.

Set of B can be considered independent if all of its attributes are very required. Each of subset B from B defined as reduct from B if B is independent and $IND_{IS}(B') = IND_{IS}(B)$. In conclusion, reduct is a set from certain attributes which able to provide the same forecasting accuracy as when all of the attributes are applied to forecast. On the other hand, the attributes which considered as non-reduct are all of the attributes which able to erased.

*Definition 2.4 (Core)*
For $B \subseteq P$ and core of B is the set of all indispensable attributes from B, then Core can be defined as:

$$Core(B) = \cap\, Red(B) \tag{4}$$

Where $Red(B)$ is the set of all reduct output on B.

Because Core can be described as the intersection of all the reduct output, then all of Core attributes are induced in every reduction process. In other words, Core can be considered as the most important parts from attributes, because there are no indispensable attributes in it.

**Logistic Regression**
Logistic regression is one of the statistical method which illustrate the relation between response variables (y) with one or more predictor variables (x), where all of response variables in logistic regression are binary or dichotomy. The result for each of observation can be classified as success or failed. This classification is represented by y = 1 for the success output and y = 0 for the failed output. Logistic regression is the strong statistical method from the result of binomial modelling with one or more explanatory variable [7].

Mathematics model of logistic regression is the model which applied when the properties of response variables is qualitative. The function of logistic regression model is described as follows:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{5}$$

Where $\pi(x)$ is the cumulative probability of dependent variables, $\beta_0, \beta_1$ is the regression parameter, and $x$ is the dependent variable.

The properties of this equation is nonlinear in parameter, which impacted in the needs of transformation process to alter its properties into linear. This transformation process is known as logit transformation and described as follows:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{6}$$

Logistic regression creates dependent variable which is the linear combination from the independent variable. The value of this dependent variable then transformed into probability through logit transformation. Logistic regression resulting in the ratio of probability known as odds rations, which related with the each of independent variable values.

### K-Nearest Neighbor (KNN)

K-Nearest neighbor is one of the method used for decision making by applying the supervised learning where the result of the new input data is classified based on its neighboring position towards the value of the data [8]. The algorithm of K-Nearest Neighbor (KNN) is a method to classify the objects based from the closest distance of learning data to these objects. KNN is the supervised learning algorithm where the result from the new query instance is classified based on the majority of the category on KNN algorithm [15]. The most emerged cluster then will be considered as the cluster result from the classification process [10]. The distance closeness is defined in metric distances, such as Euclidean distance. The Euclidean distance [12] can be determined through this following equation:

$$D_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

Where $D$ is the closeness distance, $x$ is the data training, $y$ is the data testing, $n$ is the number of individual attributes ranged from $1, \ldots, n$, $f$ is the similarity function of attribute i between case $X$ and case $Y$, and $i$ is the individual attributes between 1 to n.

The steps to determine the K-Nearest Neighbor method are:

1. Determining the $K$ parameter (the number of nearest neighbors).
2. Measuring the square of each objects Euclid distance towards the data sample.
3. Sorting these objects into the classification which have the lowest Euclid distance.
4. Clustering the Y category (Nearest neighbor classification)
5. By applying the majority of the nearest neighbor category, then the value of the measured query instance can be predicted.

### Support Vector Machines (SVM)

Support vector machines is the classification method which works by finding the hyperplane with the highest margins. Hyperplane is the boundary line which separates data between clusters. Margin is the distance between hyperplane and the nearest data for each cluster. The nearest data with the hyperplane for each cluster can be described as support vector. The support vector machines method functioned to separate two clusters on input data which are cluster data 1 and cluster data 0. There are only two data classification because the algorithm applied in this research is binary classification algorithm. The classification problem can be solved by SVM classification method [13] [5]. This method works by determining the line which are able to separates two classification of data. The best result of hyperplane can be determined by measuring the margin of the hyperplane and its maximum point.

This method can be formulated on SVM optimization problem for linear classification as follows:

$$\min \frac{1}{2}\|\omega\|^2 \tag{8}$$

$$y_i(wx_i + b) \geq 1, i = 1, \dots, \lambda \tag{9}$$

Where $x_i$ is the data input, $y_i$ is the output of the data, data $x_i, w, b$ is the parameters where its values will be determined. From the equation, to minimalize the objective function $\frac{1}{2}\|\omega\|^2$ or maximize the quantity $\|\omega\|^2$or $w^T w$, the constraint of $y_1(wx_i + b) \geq 1$ needs to be paid attention. If the output data $y_1 = +1$, then, the constraint will be transformed into $(wx_i + b) \geq 1$. On the other hand, if $y_1 = -1$, the constraint will be transformed into $(wx_i + b) \leq -1$. In certain infeasible cases where certain data may not be able to classified, the mathematical formulation is transformed into:$\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{\lambda} t_i$

$$y_i(wx_i + b) + t_i \geq 1, \tag{10}$$

$$t_i \geq 0, i = 1, \dots, \lambda \tag{11}$$

Where $t_i$ is the slack variable. This formulation can maximize the margin between two clusters by minimalizing $\|\omega\|^2$ . To minimalize the misclassification error which presented by the appearance of slack variable of $t_i$, while in the same time we maximize the margin $\|\omega\|^2$. The application of slack variable $t_1$is to solve the infeasibility case from the constraints of $y_i(wx_i + b) \geq 1$ through the method of giving penalty to the data which unable to meet the constraints. To minimalize the value of $t_i$, the penalty will be given by applying C constant. Vector w is perpendicular with the separation function of $wx + b = 0$. B constant is applied to determine the location of the separation function relative to the origin point.

## Machine Learning Metrics

In machine learning, the metrics term is referred into certain value which can be applied to represent the performance of the resulted model. In this research, the model performance evaluation testing is done by applying the confusion matrix table. This table is functioned to observe the value of accuracy, precision, recall, and f1-score which illustrated from the following table 1 below:

### *Confusion Matrix*

Confusion matrix is the performance measurement for the machine learning classification problem where the output can be resulting in two or more clusters. Confusion matrix is a table with four different combination from the prediction value and actual value. There are 4 terms which represents the output of the classification process on confusion matrix, such as True positive, True negative, False positive, and False negative.

**Table 1**. Confusion Matrix.

| Classification | |
|---|---|
| **class 1** | **class 0** |
| True-Positive (TP) | False-Negative (FN) |
| False-Positive (FP) | True-Negative(TN) |

Where

TP: The number of data labeled true-positive and classified by model as positive label
TN: The number of data labeled true-negative and classified by model as negative label
TP: The number of data labeled true-negative and classified by model as positive label
TP: The number of data labeled true-positive and classified by model as negative label

### *Accuracy*

Referring on confusion matrix, accuracy is the ratio from the number of diagonal elements towards all of the matrix elements or:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{12}$$

### *Precision*

Precision can be described as the degree of model reliability when the given prediction is positive. Precision is the proportion of prediction labeled as positive which is true based on the all of the positive prediction, or:

$$precision = \frac{TP}{TP+FP} \tag{13}$$

### *Recall or Sensitivity*

Sensitivity can be interpreted as the degree of the model ability to detect the data labeled as true positive . Sensitivity can be defined as the proportion of the number of data which predicted as a model labeled positive from all of the data which are truly labeled positive or:

$$recall = \frac{TP}{TP+FN} \tag{14}$$

### *F1-Score*

F-1 score illustrate the average comparison of the weighted precision and recall. Accuracy is applied as the reference of algorithm performance when the dataset have the insignificant difference between the number of false negative data and false positive (symmetric). However, when the number is not insignificant, then F-1 score should be applied as the reference. The equation of F-1 score is described as follows:

$$F1 = 2 \times \frac{presisi \times recall}{presisi+recall} \tag{15}$$

### Dataset

Machine learning algorithm with core and reduct is applied on five different dataset. The first is data Iris. This dataset contains 4 variables. The second one is data Seeds which contains 7 variables. The third one is Yeas which contains 8 variables. The fourth dataset is data Sonar consisting of 60 variables. The last dataset is dataset Hill-Valley which consists of 100 variables.

## Results and Discussions

### Dataset

The application of Core and Reduct is proven for providing good result to apply dimension reduction of the high dimension data. Table 2 illustrate the result of Core and Reduct reduction for each of the dataset.

**Table 2.** Core and Reduct results

| Dataset | Before reduction | After reduction |
|---|---|---|
| Iris | 4 | 4 |
| Seeds | 7 | 2 |
| Yeast | 8 | 6 |
| Sonar | 60 | 2 |
| Hill Valley | 100 | 2 |

Based on this result, core and reduct works really well in the data with high dimension. As for the low dimension data, the application of core and reduct for dimension reduction is tends to be difficult to determine the core of the analyzed dataset.

## Computational time

The main objective of the dimension reduction is to compare the machine learning model with and without core and reduct dimension reduction. The lower the acquired result, then the better the application of core and reduct works in dimension reduction. This result clearly will affect the computational time. Figure 1 illustrate the comparison of the computational time between the machine learning model with reduction and without reduction process. The result shows that the core and reduct is relatively able to improve the computational weight and time from the machine learning model, which impacted in the decrease of the computational time. However, the result still can be considered not significant because core and reduct only decreasing the number of attributes.

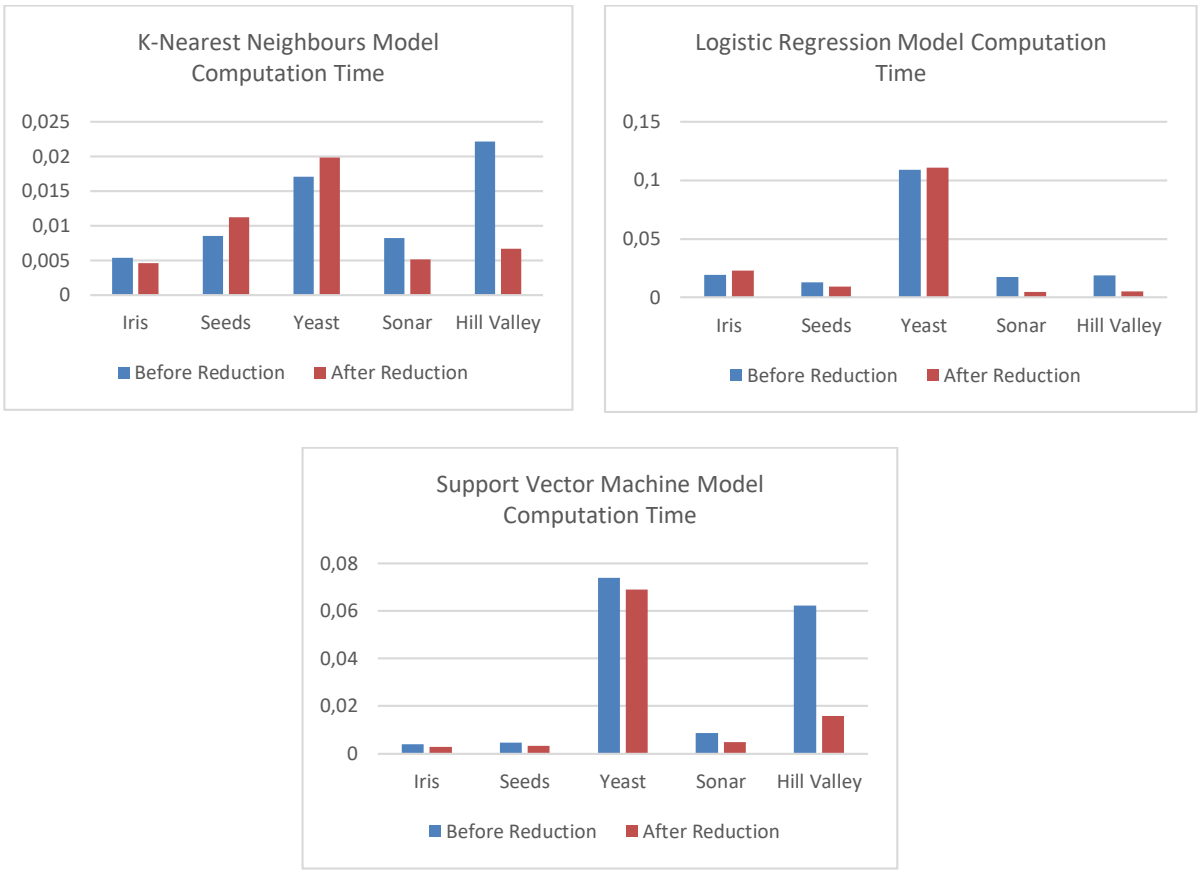


**Figure 1**. Computational time.

In the figure 1, we can observe that on dataset Hill-Valley, the computational time is improved significantly on the three method of machine learning, especially in SVM which alter the computational time from 0.06221 to 0.01583. The improvement on computational time in the application of SVM is occur in all of the given dataset. This concludes that the core and reduct dimension reduction is working well. Furthermore, it can also be observed from the other method that the improvement on computational time always occur in at least one given dataset.

## Machine Learning Evaluation

The machine learning evaluation applied to observe the performance of the applied model. The result of the machine learning model with and without core and reduct dimension reduction is clearly will affect the model performances. Figure 2 illustrate the comparison of the evaluation between the machine learning models with and without dimension reduction application.

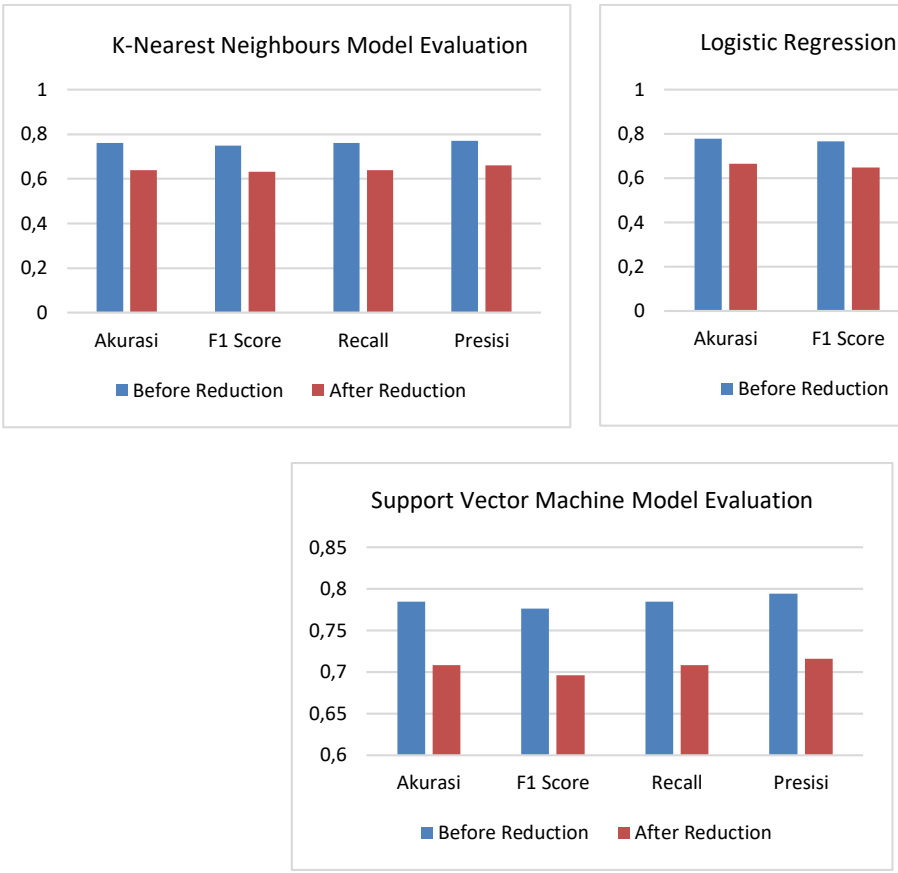


**Figure 2**. Model Evaluation.

From the figure 2, we can observe that the average model evaluation from the three applied method resulting in the insignificant output. Nevertheless, the accuracy, recall, precision, and F-1 score fro each method still can be maintained.

Few of the result illustrated above are linear with the result from the previous research [5] [7] [15]. Lei, and co. [9] also did a research in the same sector with the application of RST to decrease the variables which excessively impact on the building energy consumption. In addition to that, the critical variables from the building consumption energy also able to determined. The result shows that RST able to provide a practical and accurate solution to forecast the building energy consumption. The addition of Core process in this research guaranteeing that the dimension reduction process only occur from the core of the dataset only.

## Conclusions

Core and reduct worked better in the dataset with higher dimension. Core and Reduct is proven to be able to improve the computational time and weight on Machine learning method. This dimension reduction process only presenting the core of the dataset as the result of its reduction. The application of core and reduct focuses on the improvement of the computational time and weight in the three applied methods without significantly altering the model evaluation on each model.

## References

[1] **A. Prajana**, F. Sains, T. Universitas, I. Negeri, A. Raniry, and B. Aceh, Penerapan Teory Rough Set Untuk Memprediksi Tingkat Kelulusan Siswa Dalam Ujian Nasional Pada Sma Negeri 5 Kota Banda Aceh, *J. Islam. Sci. Technol.*, vol. 2, no. 1, pp. 7588, 2016, [Online]. Available: www.jurnal.ar-raniry.com/index.php/elkawnie.

[2] **B. Walczak** and **D. L. Massart**, Rough sets theory, *Chemom. Intell. Lab. Syst.*, 1999, doi: 10.1016/S0169-7439(98)00200-7.

[3] **C. Wu**, **Y. Yue**, **M. Li**, and **O. Adjei**, The rough set theory and applications, *Engineering Computations (Swansea, Wales)*. 2004, doi: 10.1108/02644400410545092.
[4] **G. Qi**, **Z. Zhu**, **K. Erqinhu**, **Y. Chen**, **Y. Chai**, and **J. Sun**, Fault-diagnosis for reciprocating compressors using big data and machine learning, *Simul. Model. Pract. Theory*, vol. 80, pp. 104127, 2018, doi: 10.1016/j.simpat.2017.10.005.
[5] **G. Suwardika**, Pengelompokan Dan Klasifikasi Pada Data Hepatitis Dengan Menggunakan Support Vector Machine (SVM), Classification And Regression Tree (Cart) Dan Regresi Logistik Biner, *J. Educ. Res. Eval.*, vol. 1, no. 3, p. 183, 2017, doi: 10.23887/jere.v1i3.12016.
[6] **I. De Feis**, Dimensionality reduction, in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2018.
[7] **I. Ketut**, **P. Suniantara**, and **M. Rusli**, Klasifikasi Waktu Kelulusan Mahasiswa Stikom Bali Menggunakan Chaid Regression Trees dan Regresi Logistik Biner, *Statistika*, vol. 5, no. 1, pp. 2732, 2017.
[8] **J. Qiu**, **Q. Wu**, **G. Ding**, **Y. Xu**, and **S. Feng**, A survey of machine learning for big data processing, *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, 2016, doi: 10.1186/s13634-016-0355-x.
[9] **L. Lei**, **W. Chen**, **B. Wu**, **C. Chen**, and **W. Liu**, A building energy consumption prediction model based on rough set theory and deep learning algorithms, *Energy Build.*, vol. 240, p. 110886, 2021, doi: 10.1016/j.enbuild.2021.110886.
[10] **L. Zhou**, **S. Pan**, **J. Wang**, and **A. V. Vasilakos**, Machine learning on big data: Opportunities and challenges, *Neurocomputing*, vol. 237, pp. 350361, 2017, doi: 10.1016/j.neucom.2017.01.026.
[11] **M. Mohammadi** and **A. Al-Fuqaha**, Enabling cognitive smart cities using big data and machine learning: Approaches and challenges, *arXiv*, no. February, pp. 94101, 2018.
[12] **M. S. Raza** and **U. Qamar**, An incremental dependency calculation technique for feature selection using rough sets, *Inf. Sci. (Ny).*, 2016, doi: 10.1016/j.ins.2016.01.044.
[13] **O. F.Y**, **A. J.E.T**, **A. O, H. J. O, O. O**, and **A. J**, Supervised Machine Learning Algorithms: Classification and Comparison, *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128138, 2017, doi: 10.14445/22312803/ijctt-v48p126.
[14] **S. A. Naufal**, **A. Adiwijaya**, and **W. Astuti**, Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray, *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, p. 162, 2020, doi: 10.30865/jurikom.v7i1.2014.
[15] **S. Zhang** *et al.*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 1 Efficient kNN Classification With Different Numbers of Nearest Neighbors, *Ieee Trans. Neural Networks Learn. Syst.*, pp. 112, 2017, [Online]. Available: http://ieeexplore.ieee.org.
[16] **W. K. Ching**, **M. K. Ng**, and **E. S. Fung**, Higher-order multivariate Markov chains and their applications, *Linear Algebra Appl.*, 2008, doi: 10.1016/j.laa.2007.05.021.

THIS PAGE INTENTIONALLY LEFT BLANK