Journal on Mathematics Education
Volume 15, No. 3, 2024, pp. 751-770

Radial Basis Function Neural Network with ensemble clustering for modeling mathematics achievement in Indonesia based on cognitive and non-cognitive factors

Dhoriva Urwatul Wutsqa[1], Pusparani Puan Prihastuti[2], Muhammad Fauzan[2], Endang Listyani[3]

[1]Study Program of Statistics, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2]Study Program of Mathematics, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3]Study Program of Mathematics Education, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
*Correspondence: dhoriva_uw@uny.ac.id

Received: 29 January 2024 | Revised: 14 May 2024 | Accepted: 19 June 2024 | Published Online: 30 July 2024

Abstract

Mathematics achievement could be influenced by cognitive and non-cognitive factors. The potential variable of cognitive factor is metacognition, whereas non-cognitive factors include Economic, Social, and Cultural Status (ESCS), resilience, life satisfaction, happiness, pride, fear, sadness, and gender. Those variables involve numerical and categorical data. For this reason, this study aims to apply the Radial Basis Function Neural Network (RBFNN) model with ensemble clustering to model the relation between cognitive and non-cognitive aspects and mathematics achievement. The RBFNN is a soft computing approach based on the neural network model and has been shown as an effective model and free of assumption. The ensemble clustering is a process in RBFNN modeling to capture the independent variables involving the numerical and categorical data. It employs K-means clustering for the numerical data and K-modes for categorical data and combines the results of those two methods. The data used in this study are published by PISA (Program for International Student Assessment) 2018. The results show that the RBFNN with ensemble clustering deliver good performance in modeling the students' mathematics achievement based on the cognitive and non-cognitive factors in terms of prediction accuracy. Other than RBFNN model, the use of cognitive and non-cognitive factors involving in this study also contributes to the high accuracy prediction. This further emphasizes that these factors are good predictors of mathematic achievement. Additionally, we suggest the silhouette cluster validation in the clustering process, since it leads to the number of hidden neurons of the best RBFNN model.

Keywords: Ensemble Clustering, Mathematics Achievement, PISA 2018, RBFNN

How to Cite: Wutsqa, D. U., Prihastuti, P. P., Fauzan, M., & Listyani, E. (2024). Radial Basis Function Neural Network with ensemble clustering for modeling mathematics achievement in Indonesia based on cognitive and non-cognitive factors. *Journal on Mathematics Education, 15*(3), 751-770. http://doi.org/10.22342/jme.v15i3.pp751-770

Indonesia is one of the countries participating in the Program for International Student Assessment (PISA). PISA measures the ability of 15-year-olds to use their knowledge, reading, mathematics and science skills to face real life challenges. Student achievement in PISA is an indicator of the education level in a country. Indonesian students' mathematics achievement, as represented by the PISA results, are still far from expectations. Only 28% of students in Indonesia who took mathematics tests administered by PISA achieved level 2 or above and only 1% achieved level 5 or higher (OECD, 2018). The latest PISA 2022 data were released at the end of 2023. Although the PISA 2022 showed that

http://doi.org/10.22342/jme.v15i3.pp751-770

learning outcomes especially in mathematics achievement declined globally owing to the pandemic, Indonesia rose 5-6 positions compared to 2018. This indicates the resilience of the education system in Indonesia against pandemic-associated learning loss.

In addition to conducting tests to measure students' mathematics achievements, PISA also conducts surveys of students who take the exam. According to survey results from PISA, there are many variables of student conditions that are supposed to influence or relate to students' achievement, including non-cognitive and cognitive factors. The non-cognitive factors are economic, social, and cultural status (ESCS), resilience, life satisfaction, happiness, pride, fear, sadness, and gender and the cognitive factor is metacognition. Numerous studies have demonstrated the relation of those variables to mathematics achievement. Anggraheni and Kismiantini (2022) showed that ESCS and metacognition have a positive relationship with mathematics achievement, while gender has a negative relationship, whereas Krisnamurti and Kismiantini (2022) added that the students' happiness and anxiety have a positive relationship with their mathematics achievement. Rahmawati and Kismiantini (2022) also reported that resilience has a relationship with students' mathematics achievement. Furthermore, Areepattamannil (2014) demonstrated that non-cognitive factors such as gender, students' positive attitudes towards school, and students' positive perceptions of classroom climate have a significant effect on students' mathematics achievement. Other studies on different source data are in accordance with the results of PISA data. Patmaniar et al. (2021) found the differences between male and female in the level of students' understanding in solving problems. Other factor such as motivation also showed to have relationship with mathematics achievement (Tran & Nguyen, 2021).

The PISA data has attracted many researchers to study the relationship between the variables and achievement. Multilevel analysis is applied to explain the influence of gender, ESCS, metacognition, and study time on students' mathematics achievement (Anggraheni & Kismiantini, 2022), and to analyze the relationship between socio-economic status and school resources on mathematics achievement (Efendi & Kismiantini, 2022). In addition, still using multilevel analysis, a study has explained the importance of school size and teacher-student ratio (Samnufida & Kismiantini, 2022) as well as growth mindset, gender, ESCS (Kismiantini et al., 2021) in influencing students' mathematics achievement in Indonesia. Pitsia et al. (2017) also utilized multilevel analysis to examine the contribution of non-cognitive factors, such as students' mathematics self-efficacy, motivation to learn mathematics, and attitudes toward school to predict the students' mathematics achievement in Greece. Lee and Stankov (2013) tested several models, such as Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and Structural Equation Modeling (SEM) to investigate the high-level factor structure of fifteen main variables selected from four non-cognitive factors, namely academic self-confidence, motivation, learning strategies, and attitudes towards school.

The model used in the previous studies use a parametric approach, in which many assumptions must be met, such as normality, homogeneity of variance and linearity. On the other hand, the non-parametric approach, which is more flexible and does not require strict assumptions, has not received much attention from researchers, especially in modeling relationships between variables in the PISA data. Therefore, in this study, a non-parametric approach is proposed, namely a neural network (NN) model. This model is chosen because of its flexibility and ability to theoretically model any function explaining the relationship among variables.

The PISA 2018 data can be considered as big data since they provide information of large number of students and collected in different formats from various sources. As stated by Vaitsis et al. (2016), the definition of big data might vary depending on the application scope, but the size and the source are important determinants. The machine learning approach such as NN become more desirable than parametric approach for big data analysis. It is more promising to obtain more accurate prediction for big data. The NN model has been widely applied in various fields and yields high accuracy prediction. However, there have not been many published studies on PISA data using NN model. Thus, the use of NN to model student achievement using PISA data is still open to be explored. Several studies promote NN model to analyze International PISA data for many countries. Aksu et al. (2022) used the M5P algorithm and artificial neural network to predict students' mathematical literacy on PISA data for 6 different countries, namely Singapore, Japan, Norway, the United States, Turkey and the Dominican Republic through separate analysis by adjusting for different ability levels. Demir and Karaboğa (2021) compared several NN methods such as Multi-Layer Perceptron (MLP), Elman Neural Network (ENN), and Jordan Neural Networks (JNN), to model and predict student mathematics achievement based on the PISA 2018 data. Meanwhile, Koyuncu (2020) tested the importance of mathematics-specific trend variables in PISA 2003 and 2012 to predict inter-year mathematics achievement using a two-step analysis method involving data analysis methods, multilayer perceptron and Radial Basis Function Neural Network (RBFNN), and multiple linear regression.

Some studies reveal the use of various NN models to predict students' achievement. Haviluddin et al. (2014) compared simple linear regression and radial basis function neural network (RBFNN) methods to predict the achievement using motivation factor of 108 students from the mathematics department at an Islamic university in Bengkulu. The result shows that the RBFNN model has better performance in terms of the smaller sum-square error (SSE) value. Aybek and Okur (2018) compared multilayer perceptron (MLP) and radial basis function neural network (RBFNN) to predict final exam scores and students' pass/fail rates. The results show that the network of MLP yields more precise predictions than RBFNN. Meanwhile, Huang and Fang (2013) compared four methods, namely, Multilayer Perceptron (MLP), RBFNN, and Support Vector Machine (SVM) to predict students' academic performance and found that SVM was the best model to predict individual academic performance. However, in other contexts, RBFNN has been proven to have good performance in modeling non-linear relationships, overfitting, and has better generalization ability (Huang et al., 2005; Huang & Fang, 2013). The RBFNN model is a very popular type of NN. This model has also been widely applied in various fields, including for disease classification. Abadi et al. (2017; 2019; 2021) use RBFNN for lung, brain, and prostate cancer detection, while Cheruku et al. (2017) and Kamble and Kokate (2020) for diabetes detection, Sateesh and Suresh (2013) for parkinson, Wutsqa and Farhan (2020) for lung cancer, and Wutsqa and Fauzan (2022) for breast and brain cancer detection. The ability of RBFNN to predict has been proven in several studies in many areas, such as agriculture (Ditakristy et al., 2016), socioeconomics (Shen et al., 2011), and health (Dhamodharavadhani et al., 2020).

RBFNN is a neural network model which consists of three layers, namely the input layer, hidden layer and output layer. The learning in the RBFNN model aims to determine the parameters of the activation function in the hidden layer, the number of hidden layers, and the weight between the hidden layer and the output layer. The clustering process is an important process in RBFNN because the prediction accuracy of the RBF network is influenced by the number of the hidden neurons whose values are the same as the number of clusters; therefore, the clustering process must be carried out using the suitable method (Huang & Fang, 2013). K-means is the most popular and simple clustering method in RBFNN. This method places objects in clusters based on the closest Euclidean distance to the centers (cluster average). The K-means method in the RBFNN model learning process was carried out by Sing et al. (2003), which compared the use of the K-means method and its improved version to select hidden neurons from RBFNN. Dubey (2015) also conducted research on the use of K-means clustering in RBFNN modeling to predict rainfall in Yokohama, Japan. According to Chrisinta et al. (2020), the K-means method is more appropriate for clustering numerical data. However, for mixed type data consisting of numerical and categorical data, it is more appropriate to use the ensemble clustering. Ensemble clustering is a method that combines the best outputs from several groupings to achieve more accurate and stable final results (Sowjanya & Mrudula, 2015). The ensemble clustering method for mixed type data (numerical and categorical) can combine the K-means method for clustering numerical data and K-modes for categorical data. That combination has been done by Ali et al. (2017), which is proven to be efficient and effective in handling mixed data problems.

The cognitive and non-cognitive factors that potentially influence mathematics achievement involve numerical and categorical data. For this reason, we propose RBFNN with ensemble clustering as a combination of the K-means and K-modes methods to model mathematics achievement on PISA data. An ensemble clustering is a new approach and has not been applied in previous studies on RBFNN modeling. The clustering process in the RBFNN model in previous studies did not pay attention to the type of data, whether numerical or categorical, or a combination of both. Therefore, this study aims to model the mathematics achievement on PISA 2018 data using RBFNN with ensemble clustering based on cognitive and non-cognitive factors. From the obtained model, we can predict mathematics achievement using cognitive and noncognitive factors. Model accuracy can also indicate the suitability of prediction factors.

METHODS

Data and Variable

This study uses the PISA 2018 data, which focuses on reading, mathematics, science and global competencies as areas of assessment (OECD, 2019). The PISA 2018 survey in Indonesia involves 12,098 students from 397 schools. However, this study only uses 10,628 data from Indonesian students who give complete responses to each variable required in the research.

Table 1. Dataset Structure

Symbol	Variable	Code	Data Type
x_1	Economic, Social, and Cultural Status	ESCS	Numerical
x_2	Metacognition	UNDREM	Numerical
x_3	Resilience	RESILIENCE	Numerical
x_4	Life satisfaction	ST016Q01NA	Categorical
x_5	Happiness	ST186Q05HA	Categorical
x_6	Pride	ST186Q09HA	Categorical
x_7	Fear	ST186Q02HA	Categorical
x_8	Sadness	ST186Q08HA	Categorical
x_9	Gender	ST004D01T	Categorical
y	Plausible Values in Mathematics	PVMATH	Numerical

The predictor variables of mathematics achievement are economic, social, and cultural Status (ESCS), metacognition, resilience, life satisfaction, happiness, pride, fear, sadness, and gender. An explanation of the dataset structure including symbol, code, and data type is presented in Table 1.

The predictor variables in this study are categorized as cognitive and non-cognitive factors that influence mathematics achievement. Cognitive factors are represented by metacognition. Metacognition is a person's awareness of cognitive processes and the ability to control them (Ovan et al., 2018). Metacognition in PISA 2018 data is divided into 3 aspects, namely understanding and remembering (UNDREM), summarizing (METASUM), and assessing credibility (METASPAM) (Firat & Koyuncu, 2023). However, this study only uses the UNDREM aspect with scores in the range of -1.64 to 1.50. The non-cognitive factors include Economic, Social, and Cultural Status (ESCS), resilience, life satisfaction, happiness, pride, fear, sadness, and gender. ESCS refers to the economic, social, and cultural status of a student's family, which has a range of -5.78 to 2.97. Resilience is a positive adjustment in overcoming difficulties, especially academic resilience from disadvantaged socio-economic backgrounds (Agasisti et al., 2018). In PISA 2018, this variable reflects students' perceptions of themselves around achievement, overcoming difficult situations, multitasking, and self-confidence (Govorova et al., 2020). The resilience variable in this study has a range of -3.17 to 2.37. Furthermore, the life satisfaction variable evaluates students' overall satisfaction with life on a scale of 0-10, indicating “not at all satisfied” to “very satisfied” (Govorova, et al., 2020). Positive feelings are represented by the happiness and pride variables. These variables relate to the frequency with which students usually feel happy and proud on a scale of 1-4, namely “never”, “rarely”, “sometimes”, and “always” (Jerrim, 2022). The variables of fear and sadness represent negative feelings in students' daily lives. This index reflects how often students feel afraid and sad. Responses to this question are divided into a four-point scale, namely “never”, “rarely”, “sometimes”, and “always”. Furthermore, the gender variable is categorized into two, namely category 1 for female students and category 2 for male students (Marcq & Braeken, 2023). The output variable used in this research is the mathematics achievement of Indonesian students, which is calculated from the average of the PV1MATH-PV10MATH variables with a value range of 129.60 to 721.00.

Data Analysis

In this section, we explain theories used to build the procedure of RBFNN modeling starting from the K-means clustering, K-modes clustering, ensemble clustering, and cluster validation. Then, we describe the RBFNN model and its estimation method, followed by the accuracy of the model. Based on the theory, we describe the procedure analysis of RBFNN with ensemble clustering in the last subsection.

K-means Clustering

Clustering method is a process of grouping datasets into a number of groups or clusters in a way that objects inside a cluster have high similarity with one another, and high dissimilarity with objects in other clusters. One of the methods that can be used to group numerical type data is the K-means method. In the K-means method, the object is put in a cluster with closest distance to cluster mean (Johnson & Wichern, 2007).

The K-means method algorithm is arranged as follows:

1. Define vector $\boldsymbol{x}$ as input data
2. Determine the number of m cluster
3. Partition object into m initial cluster
4. Determine the m cluster center point using cluster mean
5. Determine the distance of each object $\boldsymbol{x}$ to cluster center $\boldsymbol{c_j}$ using Euclid distance

$$d(\boldsymbol{x}, \boldsymbol{c_j}) = \sqrt{\sum_{i=1}^{p} (x_i - c_{ji})^2} \tag{1}$$

where $\boldsymbol{x} = [x_1, x_2, \dots, x_p]$ is an input data vector, $\boldsymbol{c_j} = [c_{j1}, c_{j2}, \dots, c_{jp}]$ is *j*th cluster center vector, x_i is *i*th input variable, c_{ji} is *j*th cluster center from *i*th input variable

6. Place each object to clusters that have the closest distance to center
7. Repeat step 4 and 5 until the old center value is equal to the new center value

K-modes Clustering

K-modes clustering was first introduced by Hwang in 1997. K-modes clustering is a modification of K-means clustering that is used for categorical data. K-modes method makes a modification to K-means by changing the means with modes. The steps of K-modes algorithm are almost the same as those in the K-means method. In addition, change also occurs in the distance function (Huang, 2009). The distance (1) becomes

$$d(\boldsymbol{x}, \boldsymbol{c_j}) = \sum_{i=1}^{p} \epsilon\ (x_i, c_{ji}) \tag{2}$$

where $\epsilon\ (x_i, c_{ji})$ is

$$\epsilon\ (x_i, c_{ij}) = \begin{cases} 0\ (x_i = c_{ji}) \\ 1\ (x_i \neq c_{ji}) \end{cases} \tag{3}$$

Ensemble Clustering

Ensemble clustering is a method to combine a number of different clustering methods to obtain general grouping results from the original dataset. The clustering results from each clustering method are set as input for ensemble clustering process (He et al., 2005). Ensemble clustering algorithm is also known as the CEBMDC (Cluster Ensemble Based Mixed Data Clustering) algorithm.

The procedure in ensemble clustering is as follows:

1. Mixed data are separated into categorical and numerical data.
2. K-means method is used for numerical data while K-modes is used for categorical data.
3. Combining the cluster results from both methods. Both clusters yield results with a categorical type, so ensemble clustering is performed using clustering method for categorical data. In this study, we implement again K-modes clustering. The attributes of the second K-modes clustering step include two variables, whose values are the categories of each clustering result.

Cluster Validation

Cluster validation is carried out to obtain the best clusters representing the grouping of the data. Silhouette is one method to see the quality and strength of clusters. This method uses a coefficient known as Silhouette width which is a measure of the difference of the distance between objects in a cluster and the distance of a separate cluster from another cluster (Dubey, 2015). For a particular object *k*, the Silhouette width can be formulated as follows:

$$s(k) = \frac{b(k) - a(k)}{\max(a(k),\ b(k))} \tag{4}$$

where $a(k)$ is the average distance between object k and all other objects in the cluster and $b(k)$ is the minimum distance between the k-*th* object in the cluster and all members of the other clusters. The Silhouette width ranges from -1 to 1. If the value is close to -1, it means that the object has been misclassified. If the value is 0, it is considered an intermediate case because object k is located equally far from both clusters. If the value is close to 1, it means that the objects are grouped well (Kaufman & Rousseeuw, 1990). The average value of $s(k)$ across all the data in a cluster shows how closely the objects in the cluster are grouped. The average value of $s(k)$ across the dataset shows how precisely the data has been grouped. The k value that produces the highest average Silhouette width in the dataset is considered to be the optimum number of clusters (Kodinariya & Makwana, 2013).

Radial Basis Function Neural Network (RBFNN)

RBFNN is special type of neural network that uses radial function as its activation function. The characteristic feature of radial functions is that their response decreases or increases monotonically along with the distance from a central point. The RBFNN model is composed of three layers, namely the input layer, hidden layer, and output layer (see Figure 1).

On Figure 1, $\boldsymbol{x} = [x_1, x_2, \dots, x_p]$ is an input vector, while $\boldsymbol{\varphi} = [\varphi_1, \varphi_2, \dots, \varphi_m]$ is an activation function vector on hidden layer, and y is an output neuron. Weight vector between hidden layer and output layer is symbolized as $\boldsymbol{w} = [w_0, w_1, \dots, w_m]$, with w_0 as a bias. The activation function in RBFNN usually uses the Gaussian function with center and radius as the parameters. In this study, those parameters are computed from results of ensemble clustering. Each function includes the center and maximum distance of cluster.

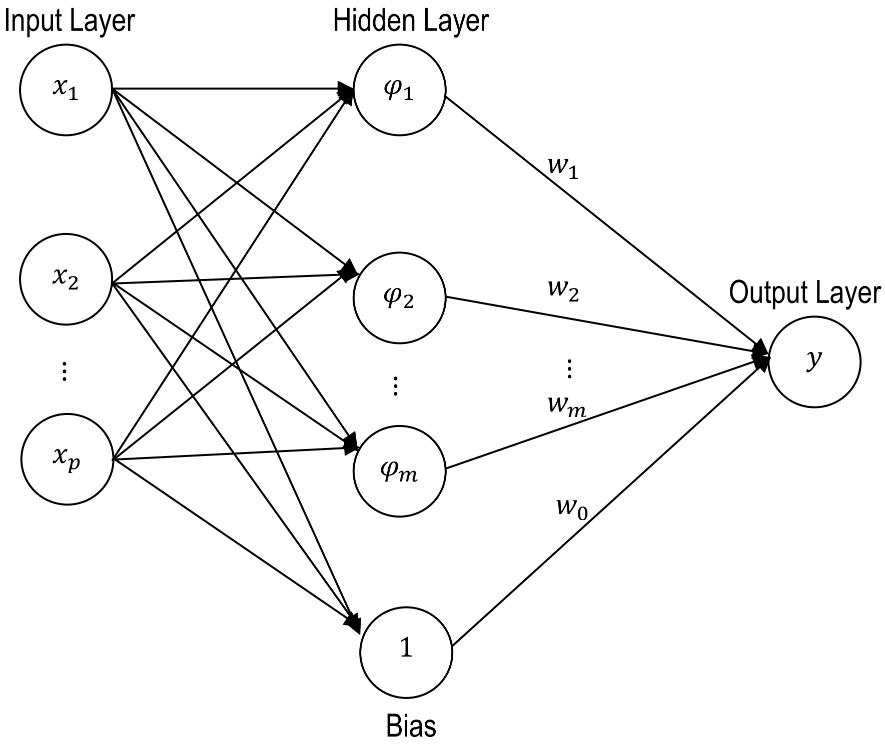


Figure 1. Architecture of RBFNN

The output y is a linear combination of weight w_j with activation function $\varphi_j(\boldsymbol{x})$,bias weight w_0, and error model

$$\mathrm{y} = \sum_{j=1}^{m} w_j \varphi_j(\boldsymbol{x}) + w_0 + \varepsilon \tag{5}$$

with Gaussian function

$$\varphi_j(\boldsymbol{x}) = \exp\left(-\sum_{i=1}^{p}\frac{(x_i - c_{ji})^2}{r_j^2}\right) \tag{6}$$

where c_{ji} is the *j*th cluster center of *i*th input variable, and r_j is the maximum distance of object $\boldsymbol{x}$ from cluster center $\boldsymbol{c_j}$ on *j*th cluster. Global ridge regression method is used to estimate weight by adding regulation parameter λ on the sum of squared error (Orr, 1996). The cost function to be minimized is

$$C = \sum_{k=1}^{n}\left(y_k - \sum_{j=1}^{m} w_j\varphi_j(\boldsymbol{x}) + \ w_0\right)^2 + \lambda\sum_{j=1}^{m} {w_j}^2 + {w_0}^2$$

Then, the vector of optimum weights after addition of regulation parameter is

$$\boldsymbol{\hat{w}} \ = \ (\boldsymbol{\varphi}^T\boldsymbol{\varphi} + \ \lambda\boldsymbol{I}\,)^{-1}\boldsymbol{\varphi}^T\boldsymbol{y} \tag{7}$$

where,

$$\boldsymbol{\varphi} = \begin{bmatrix} \varphi_1\,(\boldsymbol{x}_1) & \varphi_2\,(\boldsymbol{x}_1) & \dots & \varphi_m\,(\boldsymbol{x}_1) & 1 \\ \varphi_1\,(\boldsymbol{x}_2) & \varphi_2\,(\boldsymbol{x}_2) & \dots & \varphi_m\,(\boldsymbol{x}_2) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \varphi_1\,(\boldsymbol{x}_\mathrm{k}) & \varphi_2\,(\boldsymbol{x}_\mathrm{k}) & \dots & \varphi_m\,(\boldsymbol{x}_\mathrm{k}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \varphi_1\,(\boldsymbol{x}_\mathrm{n}) & \varphi_2\,(\boldsymbol{x}_\mathrm{n}) & \dots & \varphi_m\,(\boldsymbol{x}_\mathrm{n}) & 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\mathrm{k} \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\hat{w}} = \begin{bmatrix} \widehat{w}_1 \\ \widehat{w}_2 \\ \vdots \\ \widehat{w}_\mathrm{j} \\ \vdots \\ \widehat{w}_m \\ \widehat{w}_0 \end{bmatrix}$$

Error prediction measurement is one of selecting criteria of a model which aims to know how well a model will work on a testing data with unknown input. Choosing a good value for regulation parameter λ is one issue associated with selecting the model with lowest error prediction. Golub et al. (1979) suggested generalized cross validation GCV to select λ as the most convenient and the simplest optimization formula. The GCV is expressed as

$$\hat{\sigma}^2{}_{GCV} = \frac{n\boldsymbol{\hat{y}}^T\boldsymbol{P}^2\boldsymbol{\hat{y}}}{(trace\ (\boldsymbol{P}))^2} \tag{8}$$

where,

$$\boldsymbol{P} = \boldsymbol{I} - \ \boldsymbol{\varphi}\boldsymbol{A}^{-1}\boldsymbol{\varphi}^T$$
$$\boldsymbol{A} = \ \boldsymbol{\varphi}^T\boldsymbol{\varphi} + \lambda\boldsymbol{I}$$

Accuracy of the Model

The accuracy of the model or method can be perceived from the prediction error. It is frequently measured using the difference between actual value and prediction value. Prediction error can be measured by several criteria, but the most used are Mean Absolute Percent Error (MAPE) and Mean Squared Error (MSE) (Hanke & Wichern, 2005). MAPE is the average value of the overall error percentage (difference) between actual value and predicted data. The MSE value is used to measure the accuracy of the estimated model value which is expressed in the average square of the error. The best model can be determined by looking at the smallest MAPE and MSE results. The MAPE and MSE values are determined as

$$MAPE = \frac{1}{n}\sum_{k=1}^{n}\frac{|Y_k - \hat{Y}_k|}{|Y_k|} \times 100\% \tag{9}$$

$$MSE = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \hat{Y}_k\right)^2 \tag{10}$$

where Y_k is the k-*th* observation value, $\hat{Y}_k$ is the k-*th* forecasting value, and n observations number.

The Procedure of Modeling the RBFNN with Ensemble Clustering

In this study, the RBFNN model with ensemble clustering is used to predict Indonesian students' mathematics learning achievement. The procedure includes the steps described as follows:

Step 1 − Determining the input and output variables of the network

The input variables in this study consist of cognitive and non-cognitive factors of mathematics achievement. They are attributed as ESCS, UNDREM, RESILIENCE, ST016Q01NA, ST186Q05HA, ST186Q09HA, ST186Q02HA, ST186Q08HA, and ST004D01T as explained in data description. The output variable used in this study is the average students' mathematics score of the ten PVMATH variables.

Step 2 - Dividing the training and testing data

The data are divided into two parts, namely training data and testing data. Training data is used to find the best model, while testing data is used to evaluate the accuracy of the model to predict the out-sample data. We attempt three data division sets, i.e. 80% training data and 20% testing data, 70% training data and 30% testing data, and 60% training data and 40% testing data from a total of 10,628 data.

Step 3 - Clustering using ensemble method

The clustering process aims to estimate the parameters of Gaussian function and to determine the number of hidden neurons. It starts by separating the data according to the data type. The K-means method is used for numerical data, while the K-modes method is used for categorical data. The number of clusters in both methods are determined using Silhouette width. The results of these two methods are then used as input data for the ensemble clustering process. Since the input data is categorical, the method used in the ensemble clustering process is the K-modes method.

Step 4 - Modeling the Radial Basis Function Neural Network (RBFNN)

RBFNN modeling is divided into three parts. The first part involves determining the value of the cluster center and the maximum distance of the object to the cluster center using the ensemble clustering method. The cluster center and the maximum distance constitute the values of Gaussian function (6). The second part is determining the number of neurons in the hidden layer. The number of neurons in the hidden layer is determined according to the number of clusters obtained from the ensemble clustering method. The third part is estimating the weight from the hidden layer to the output layer using the global ridge regression method with GCV (8) criteria. The best RBFNN is obtained by attempting the learning process with several numbers of hidden neurons. The trial-and-error process will stop if the network is optimal or in other words, when the error rate has reached a fairly small level or does not lead a significant change. The error rate is evaluated using MAPE and MSE criteria.

RESULTS AND DISCUSSION

The data used in this research is the PISA 2018 data with a sample of 12,098 Indonesian students. There are 10 variables used, including variables with the attributes ESCS, UNDREM, RESILIENCE, ST016Q01NA, ST186Q05HA, ST186Q09HA, ST186Q02HA, ST186Q08HA, ST004D01T, and PVMATH. Data preprocessing is needed since some missing values and no response data are found on those variables. The ESCS, UNDREM, and ST016Q01NA variables have 90, 746, 73 missing values, respectively. The rest variables ST186Q05HA, ST186Q09HA, ST186Q02HA, and ST186Q08HA, each has 72 missing values. Meanwhile, the RESILIENCE, ST016Q01NA, ST186Q05HA, ST186Q09HA, ST186Q02HA, and ST186Q08HA variables have 317, 665, 173, 272, 280, and 298 no response data. The handling of missing values is carried out by deleting uncompleted data. This way is chosen because it produces perfect data without making any assumptions (Dixon, 1979). The amount of data after data preprocessing is 10,628.

The data type are mixed data containing numerical data and categorical data. An overview of the data is explained in descriptive statistic. It is presented in the form of a statistical summary table, which includes the average value, standard deviation, minimum value, and maximum value for numerical data. Meanwhile, categorical data is presented in the form of a statistical summary table, which contains the mode, minimum value, and maximum value. Descriptive statistics for numerical and categorical data are presented in Table 2 and Table 3.

Table 2. Descriptive statistics of numerical data

Symbol	Variable	Code	Mean	sd	Min	Max
x_1	Economic, Social, and Cultural Status	ESCS	-1,34	1,12	-5,78	2,97
x_2	Metacognition	UNDREM	-0,37	0,96	-1,64	1,50
x_3	Resilience	RESILIENCE	-0,01	0,83	-3,17	2,37
y	Plausible Values in Mathematics	PV1MATH-PV10MATH	408,10	78,71	129,60	721,00

Based on Table 2, the ESCS (x_1), UNDREM (x_2), and RESILIENCE (x_3) range from negative to positive values since those values are the normalized version of the original data. PISA publishes the normalized data instead of the original data. Meanwhile, the PVMATH (y), or plausible values in mathematics, has an average of 408.10, a standard deviation of 78.71, a minimum value of 129.60, and a maximum value of 721.00, which is below the OECD average of 490 (Ministry of Education and Culture of Indonesia, 2019).

Table 3. Descriptive statistics of categorical data

Symbol	Variable	Code	Mode	Min	Max
x_4	Life satisfaction	ST016Q01NA	10	0	10
x_5	Happiness	ST186Q05HA	4	1	4
x_6	Pride	ST186Q09HA	3	1	4
x_7	Fear	ST186Q02HA	3	1	4
x_8	Sadness	ST186Q08HA	3	1	4
x_9	Gender	ST004D01T	1	1	2

Table 3 shows varied statistics results of the variables. The variable ST016Q01NA (x_4) and variable ST186Q05HA (x_5) achieve the maximum modes, which means that mostly Indonesian students feel very satisfied with their lives and always feel happy. The variable ST186Q09HA (x_6) has a mode of 3, with a value range between 1 and 4, indicating that most Indonesian students feel proud on certain occasions. Meanwhile, negative feelings represented by the variables ST186Q02HA (x_7) and

ST186Q08HA (x_8) have a mode of 3, with the same value range, namely between 1 and 4. This shows that most Indonesian students sometimes feel afraid and sad. Furthermore, the variable ST004D01T (x_9) has a mode of 1, with a value between 1 and 2, meaning that there are more female students than male students.

The first step in RBFNN modeling is to determine the input and output variables of the network. Referring to the explanation in the method, the number of neurons in the input layer is nine, which are attributed as ESCS (x_1), UNDREM (x_2), RESILIENCE (x_3), ST016Q01NA (x_4), ST186Q05HA (x_5), ST186Q09HA (x_6), ST186Q02HA (x_7), ST186Q08HA (x_8), ST004D01T (x_9). The output variable is PVMATH (y), which is the student's mathematics score obtained by calculating the average of ten PVMATH (plausible values in mathematics).

To obtain the best RBFNN model, we go to the second step, which is dividing the data into three compositions of training and testing data, namely 80%-20%, 70%-30%, and 60-40%. The data distribution of each composition is given in Table 4.

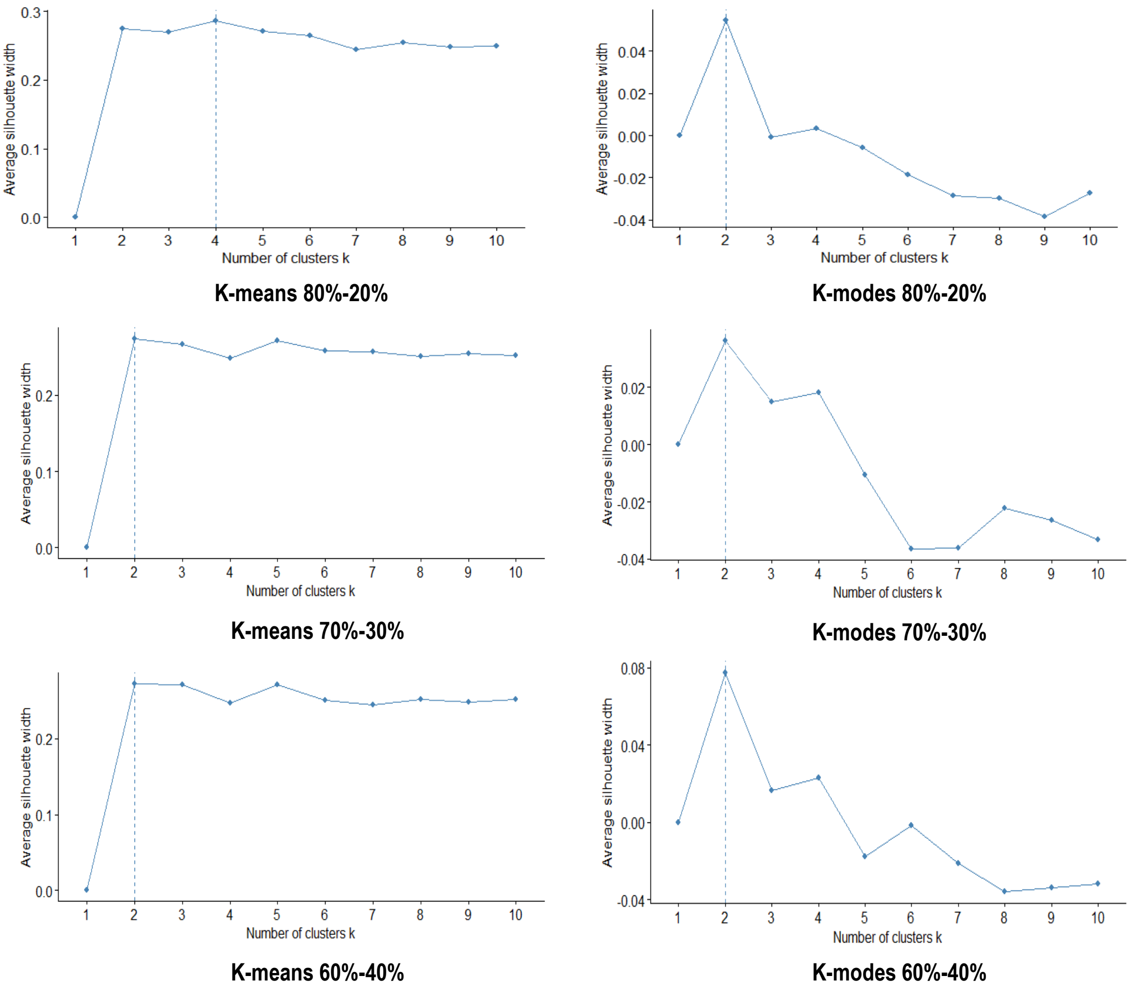
Table 4. Description of training and testing dataset compositions

Proportion of training-testing data	Number of observations	
	Training	Testing
80%-20%	8,502	2,126
70%-30%	7,440	3,188
60%-40%	6,377	4,251

The input training data variables are then used in the clustering process for RBFNN learning using the ensemble clustering method. The data is a mixture of numerical and categorical types as listed in Table 1, so it is necessary to separate the data first before carrying out the ensemble clustering process. Numerical data on input variables is processed using the K-means method, while categorical data is processed using the K-modes method. The number of clusters formed using the K-means and K-modes methods is determined using the silhouette coefficient method. The results of the silhouette coefficient method are displayed in Figure 2.

The optimum number of clusters is achieved when the average silhouette coefficient reaches its maximum. As shown in Figure 2., the best number of clusters are 4 in K-means and 2 in K-modes for the data proportion of 80%-20. Meanwhile the optimal number of clusters for data proportions of 70%-30% and 60%-40% are both 2 in K-means and 2 in K-modes. The K-modes are then applied to the two categorical variables, where for a proportion of 80%-20%, variable 1 consists of 4 categories and variable 2 consists of 2 categories, and for proportions of 70%-30% and 60%-40%, variables 1 and 2 consist of two categories.

Following the steps 4 in the procedure analysis, we attempt several numbers of clusters to obtain the best RBFNN. This number is determined using trial and error method, which result in change of cluster center, maximum distance value, and automatically also change the Gaussian function value. Then global ridge regression method is used to estimate weight. Optimum cluster number then chosen based on RBFNN model that have the highest accuracy, both on training and testing data. The accuracy results of RBFNN in terms of MAPE and MSE values using the ensemble clustering method for all training and testing data in proportions of 80%-20%, 70%-30%, and 60%-40% are presented in Table 5.



K-means 80%-20%

K-modes 80%-20%

K-means 70%-30%

K-modes 70%-30%

K-means 60%-40%

K-modes 60%-40%

Figure 2. Optimum number of clusters based on silhouette coefficient

The results in Table 5 show that the highest accuracy in training and testing data for the proportion of 80%-20% was achieved in the formation of 7 clusters. The highest accuracy in training and testing data for the proportion of 70%-30% was achieved when forming 3 clusters, while that for the proportion of 60%-40% was achieved when forming 4 clusters. This is considered by looking at the smallest MAPE and MSE values for training and testing data. Thus, the best RBFNN model for predicting students' mathematics learning achievement in Indonesia has an architecture of 9 neurons in the input layer, and the numbers of hidden neurons are 7, 3, and 4 for training and testing data proportions of 80%-20%, 70%-30% and 60%-40%, respectively. Table 5 also shows the performance of the best RBFNN model with ensemble clustering from each proportion of training and testing data to predict students' mathematics achievement in Indonesia. Based on Table 5, the MAPE values obtained from training and testing data are 13.89, 14.33, and 14.24 for training and 13.07, 13.72, and 13.76 for testing data on the proportions of 80%-20%, 70%-30%, and 60%-40%, respectively. Therefore, both the training and testing data are around an accuracy level of 86 %. The accuracies on testing data are slightly higher than on those on testing data. This shows that the RBFNN model with ensemble clustering can determine the characteristics of training data quite well and is able to predict students' mathematics achievement in Indonesia.

Table 5. Accuracy of RBFNN with ensemble clustering for each data proportion

Proportion of training-testing data	Number of clusters	Training data		Testing data	
		MAPE (%)	MSE	MAPE (%)	MSE
80%-20%	2	15.12	5,466.99	14.22	4,937.11
	3	14.70	5,160.96	13.70	4,593.72
	4	14.19	4,760.79	13.17	4,176.80
	5	14.11	4,694.78	13.29	4,189.07
	6	13.94	4,618.50	13.08	4,094.07
	7*	**13.89**	**4,598.34**	**13.07**	**4,092.06**
	8	13.30	4,619.32	13.09	4,077.09
70%-30%	2	15.01	5,391.27	14.54	5,162.25
	3*	**14.33**	**4,878.91**	**13.72**	**4,561.78**
	4	14.37	4,875.02	13.80	4,592.86
60%-40%	2	15.57	5,714.18	15.14	5,553.89
	3	14.51	4,963.14	14.05	4,732.17
	4*	**14.24**	**4,783.10**	**13.76**	**4,532.57**

Note: *) Best Model

This research focuses more on modeling learning outcomes with cognitive and non-cognitive factors, which have the potential to be good predictors. The results of the analysis on model performance are shown by the low prediction error rate of MAPE values, namely less than 20%. The characteristics of modeling with neural networks are more about investigating the predictive ability of the model, not the influencing factors. Therefore, although this research does not examine which variables have a significant influence or which variables have the greatest influence, it shows the accuracy level of the independent variables to predict mathematics achievement using the PISA 2018 data. This is different from previous studies, which focus more on analyzing which variables have a significant effect on students' mathematics achievement. Pitsia et al. (2017) tests several non-cognitive factors, such as students' mathematics self-confidence, motivation to learn mathematics, and students' attitudes towards school, which contribute to the prediction of students' mathematics achievement in Greece using the PISA 2012 data using a multilevel modeling approach. Other similar studies also use PISA data to predict or investigate factors that influence achievement using the multilevel model (Anggraheni & Kismiantini, 2022; Efendi & Kismiantini, 2022).

Research to predict achievement using a machine learning approach has been carried out but is still very limited. Table 6 shows some results from studies conducted by Aksu et al. (2022), Demir & Karaboğa (2021), and Koyuncu (2020) using a machine learning approach to predict students' mathematics achievement based on PISA data.

Table 6 shows that the level of accuracy of the model varies depending on the learning algorithm and the data used. The results mostly yield low to medium accuracy. As shown by Koyuncu (2020), the PISA 2003 and 2012 data reveal a low level of accuracy, ranging from 23% to 36%, whereas according to Aksu et al. (2022), results on the PISA 2015 data indicate a low level of accuracy ranging from 46% to

60%, with the exception for Singapore, in which the result has a very high level of accuracy (above 90%). Meanwhile, the study of Demir & Karaboğa (2021) on PISA 2018 data performs better than the studies of Koyuncu (2020) and Aksu et al. (2022) with the accuracy rate of around 67%-82%. The current study has an accuracy rate of around 85%-86, which outperforms the studies listed in Table 6. The high accuracy rate indicates that the factors involving in this study lead to be good predictors for mathematics achievement. This finding can be referred by other researchers or next studies to consider those factors as predictors of mathematics achievement with different method or goal. However, this study has not investigated which factor has more effect to mathematics achievement. It is the challenging for next studies to develop the algorithm of RBFNN modeling to address that problem. And of course, the results of the new algorithm will support the reference in mathematics education studies regarding the factors that have high effect to mathematics achievement.

Table 6. Comparison of research results using machine learning methods on PISA data

Reference	Method	Accuracy (%)
Aksu, et al. (2022)	M5P Algorithm and Artificial Neural Network	Singapore: 93.18% (Train); 91.96% (Test) Japan: 57.56% (Train); 55.33% (Test) Norway: 57.05% (Train); 51.73% (Test) USA: 64.83% (Train); 57.05% (Test) Turkey: 53.80% (Train); 46.31% (Test) Dominic: 56.01% (Train); 51.06% (Test)
Demir & Karaboğa (2021)	Multi-Layer Perceptron (MLP)	69.80% (Train); 70.50% (Test)
	Elman Neural Network (ENN)	70.60% (Train); 71.10% (Test)
	Jordan Neural Networks (JNN)	71.60% (Train); 82.60% (Test)
	Logistic Regression	68.40% (Train); 67.10% (Test)
Koyuncu (2020)	Multi-Layer Perceptron (MLP)	PISA 2003: 35.00% (Train); 35.00% (Test) PISA 2012: 31.00% (Train); 31.00% (Test)
	Radial Basis Function Neural Network (RBFNN)	PISA 2003: 31.00% (Train); 29.00% (Test) PISA 2012: 26.00% (Train); 23.00% (Test)
	Multiple Linear Regression	PISA 2003: 36.00% PISA 2012: 27.00%

In addition to giving excellent performance, this study contributes to the method development by offering the use of ensemble clustering method to deal with numerical and categorical data on factors that influence students' achievement. The RBFNN model, which has been applied in many cases, usually only uses one clustering method, regardless of whether the data is categorical, numerical, or mixed. The studied employed by Sing et al. (2003), Dubey (2015), and Wutsqa and Fauzan (2022) applied the RBFNN only using one clustering method without considering the type of data.

In the clustering process, to determine the number of clusters in K-means and K-modes, the silhouette coefficient is used. An ensemble process is carried out with K-modes using the results of the K-means and K-modes clustering. It is found that the combination of the number clusters in K-means and K-modes at the beginning will determine the number of clusters needed to obtain the optimal number of clusters. Likewise, at the proportion of 80%-20%, the formation of 4 clusters in K-means and 2 clusters in K-modes produces an optimal network of 7 clusters. The formation of 2 clusters in K-means and 2 clusters in K-modes in proportions of 70%-30% and 60%-40% at the beginning produces an optimal network with 3 clusters and 4 clusters, respectively. These results indicate that the number of clusters yielding RBFNN with the smallest MAPE value tends not to exceed the multiplication of the optimal number of clusters resulting from K-means and K-modes. Results from previous studies on RBFNN modeling without using a cluster validation such as silhouette coefficient require many trials ranging from 10-30 to obtain the optimal model (Abadi et al., 2017; 2019; 2021; Wutsqa & Farhan, 2020; Wutsqa & Fauzan, 2022). This finding suggests that the use of the silhouette coefficient leads to more efficient computation than the usual computation.

CONCLUSION

In this study we propose a new approach in modeling mathematics achievement on PISA 2018 data using soft computing. The RBFNN with ensemble clustering model is applied to numerical and categorical data. We investigated the model's performance based on its MAPE value which represents the accuracy of the model in predicting mathematics achievement. The small MAPE value indicates the high accuracy of the model for prediction. We attempt three training-testing datasets, namely 80%-20%, 70%-30%, and 60%-40%. The RBFNN with ensemble clustering delivers good performance for all datasets. The accuracies of the model on training and testing data are almost the same, and even tend to increase as the MAPE values tend to decrease. This means that the RBFNN with ensemble clustering yield the generalized ability to out-sample data or predict new data well. The good performance of the proposed model indicates that the cognitive and non-cognitive factors involving in this study can be regarding as good predictors for students' mathematics achievement. This study shows the potential of neural network model, specifically the RBFNN with ensemble clustering model, to be employed to other PISA data, since PISA data are usually analyzed using parametrical statistics which needs several assumptions. This study also finds that the use of silhouette coefficient can give insight on the number of hidden neurons producing the best RBFNN model. The limitation of this study is the factor selections of the PISA 2018 data. We mainly focused on the internal factors, such as cognitive and noncognitive factors, and only considered economy as the external factor. Thus, additional factors such as parents, school environments, and teachers' quality should be considered for better prediction in future research.

Acknowledgments

The authors would like to thank the OECD for providing access to the 2018 PISA data, via https://www.oecd.org/pisa/data/, which is the main data in this research. The author also thanks the reviewers for the suggestions provided for this research.

Declarations

Author Contribution : DUW: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Writing-Review and Editing; Funding acquisition, Data curation, Project administration, and Supervision.

PPP: Investigation, Software, Resources, Formal analysis, Data curation Visualization, and Writing – Original Draft.

MF: Writing - review & editing, Validation, Funding acquisition and Supervision.

EL: Writing - review & editing, Validation, and Supervision.

Funding Statement : This research was supported by Yogyakarta State University [grant number B/122/UN34.13/PT.01.03/2023].

Conflict of Interest : The authors declare no conflict of interest.

Additional Information : Additional information is available for this paper.

REFERENCES

Abadi, A. M., Wutsqa, D. U., & Pamungkas, L. R. (2017). Detection of lung cancer using radiograph images enhancement and radial basis function classifier. *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1-6.* https://doi.org/10.1109/CISP-BMEI.2017.8302052.

Abadi, A. M., Wustqa, D. U., & Nurhayadi. (2019). Diagnosis of brain cancer using radial basis function neural network with singular value decomposition method. *International Journal of Machine Learning and Computing*, *9*(4), 527–532. https://doi.org/10.18178/ijmlc.2019.9.4.836

Abadi, A. M., Wutsqa, D. U., & Ningsih, N. (2021). Construction of fuzzy radial basis function neural network model for diagnosing prostate cancer. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(4), 1273–1283. https://doi.org/10.12928/TELKOMNIKA.v19i4.20398

Agasisti, T., Avvisati, F., Borgonovi, F., & Longobardi, S. (2018). *Academic resilience: What schools and countries do to help disadvantaged students succeed in PISA*. https://doi.org/10.1787/e22490ac-en

Aksu, N., Aksu, G., & Saracaloglu, S. (2022). Prediction of the factors affecting PISA mathematics literacy of students from different countries by using data mining methods. *International Electronic Journal of Elementary Education*, *14*(5), 613-629. https://iejee.com/index.php/IEJEE/article/view/1757

Ali, D. S., Ghoneim, A., & Saleh, M. (2017). Data clustering method based on mixed similarity measures. *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems (ICORES 2017)*, 192-199*.* https://doi.org/10.5220/0006245601920199

Anggraheni, F. Y., & Kismiantini. (2022). Relationships of metacognition and learning time to mathematics achievement-PISA 2018 findings in Indonesia. *AIP Conference Proceedings*, *2575.* https://doi.org/10.1063/5.0108028

Areepattamannil, S. (2014). International note: What factors are associated with reading, mathematics, and science literacy of Indian adolescents? A multilevel examination. *Journal of adolescence*, *37(4), 367-372.* https://doi.org/10.1016/j.adolescence.2014.02.007

Aybek, H. S. Y., & Okur, M. R. (2018). Predicting achievement with artificial neural networks: The case of Anadolu University open education system. *International Journal of Assessment Tools in Education, 5(3), 474-490.* https://doi.org/10.21449/ijate.435507

Cheruku, R., Edla, D. R., & Kuppili, V. (2017). Diabetes classification using radial basis function network by combining cluster validity index and BAT optimization with novel fitness function. *International Journal of Computational Intelligence Systems*, *10*(1), 247. https://doi.org/10.2991/ijcis.2017.10.1.17

Chrisinta, D., Sumertajaya, I. M., & Indahwati, I. (2020). Evaluasi kinerja metode cluster ensemble dan latent class *clustering* pada peubah campuran. *Indonesian Journal of Statistics and Its Applications, 4*(3)*, 448-461.* https://doi.org/10.29244/ijsa.v4i3.630

Demir, I., & Karaboğa, H. A. (2021). Modeling mathematics achievement with deep learning methods. *Sigma Journal of Engineering and Natural Sciences*, *39*, 33–40. https://doi.org/10.14744/sigma.2021.00039

Dhamodharavadhani, S., Rathipriya, R., & Chatterjee, J. M. (2020). COVID-19 mortality rate prediction for India using statistical neural network models. *Frontiers in Public Health*, *8*(441), 1-12. https://doi.org/10.3389/fpubh.2020.00441

Ditakristy, M. L., Saepudin, D., & Nhita, F. (2016). Analisis dan implementasi radial basis function neural network dalam prediksi harga komoditas pertanian. *eProceedings of Engineering* (Vol. 3(1), pp. 1130-1139). https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/3658

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(10)*, 617-621.* https://doi.org/10.1109/TSMC.1979.4310090

Dubey, A. D. (2015). K-Means based radial basis function neural networks for rainfall prediction. *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)* (pp. 1-6). Bangalore, India. https://doi.org/10.1109/ITACT.2015.7492664

Efendi, R., & Kismiantini. (2022). Analysis of PISA 2018 results in Indonesia: Perspective of socioeconomic status and school resources. *AIP Conference Proceedings* (Vol. 2575). https://doi.org/10.1063/5.0108065

Fırat, T., & Koyuncu, İ. (2023). Examining metacognitive strategy preferences of students at different reading proficiency levels. *International Journal of Psychology and Educational Studies, 10*, 224-240. https://doi.org/10.52380/ijpes.2023.10.1.997

Golub, G. H., Heath, M., & Wahba, G. (1979) Generalised cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215-223. https://doi.org/10.1080/00401706.1979.10489751

Govorova, E., Benítez, I., & Muñiz, J. (2020). Predicting student well-being: Network analysis based on PISA 2018. *International Journal of Environmental Research and Public Health*, *17*(11), 1–18. https://doi.org/10.3390/ijerph17114014

Hanke, J. E., & Wichern, D. W. (2005). *Business forecasting (9th edition)*. London: Pearson Education.

Haviluddin, Sunarto, A., & Yuniarti, S. (2014). A comparison between simple linear regression and Radial Basis Function Neural Network (RBFNN) models for predicting students' achievement. *International Conference on Education* (pp. 299-308). Sabah, Malaysia. https://doi.org/10.13140/2.1.3878.5600

He, Z., Xu, X., & Deng, S. (2005). *Clustering* mixed numeric and categorical data: A cluster ensemble approach. *ArXiv Computer Science e-prints*, 1-14. https://doi.org/10.48550/arXiv.cs/0509011

Huang, J. Z. (2009). Clustering categorical data with k-Modes. In *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 246-250). IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch040

Huang, G., Saratchandran, P., & Sundararajan, N. (2005). A generalized growing and pruning rbf neural network for function approximation. *IEEE Transactions On Neural Networks, 16*(1), 57–67*.* https://doi.org/10.1109/TNN.2004.836241.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133-145*.* https://doi.org/10.1016/j.compedu.2012.08.015

Jerrim, J. (2022). The power of positive emotions? The link between young people's positive and negative affect and performance in high-stakes examinations. *Assessment in Education: Principles, Policy and Practice*, *29*(3), 310–331. https://doi.org/10.1080/0969594X.2022.2054941

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Upper Saddle River, New Jersey: Prentice Hall.

Kamble, V. V., & Kokate, R. D. (2020). Automated diabetic retinopathy detection using radial basis function. *Procedia Computer Science*, *167*, 799–808. https://doi.org/10.1016/j.procs.2020.03.429

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New Jersey: John Wiley & Sons. https://doi.org/10.1002/9780470316801

Kismiantini, Setiawan, E. P., Pierewan, A. C., & Montesinos-López, O. A. (2021). Growth mindset, school context, and mathematics achievement in Indonesia: A multilevel model. *Journal on Mathematics Education*, *12*(2), 279–294. https://doi.org/10.22342/jme.12.2.13690.279-294

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6). www.ijarcsms.com

Koyuncu, İ. (2020). Investigation of mathematics-specific trend variables in pisa studies with neural networks and linear regression. *Journal of Curriculum and Teaching*, *9*(4), 40. https://doi.org/10.5430/jct.v9n4p40

Krisnamurti, A. W., & Kismiantini. (2022). PISA 2018: Non-cognitive factors and school characteristics towards mathematics achievement in Indonesia. *AIP Conference Proceedings*, (Vol. 2575). https://doi.org/10.1063/5.0107787

Lee, J., & Stankov, L. (2013). Higher-Order structure of noncognitive constructs and prediction of PISA 2003 mathematics achievement. *Learning and Individual Differences*, *26*, 119-130. https://doi.org/10.1016/j.lindif.2013.05.004

Marcq, K., & Braeken, J. (2023). *Gender Differences in Item Nonresponse in the PISA 2018 Student Questionnaire*. https://www.oecd.org/pisa/data/2018database/.

Ministry of Education and Culture of Indonesia. (2019). Pendidikan di Indonesia: Belajar dari hasil PISA 2018. Jakarta: Badan Penelitian dan Pendidikan, Kemdikbud. http://repositori.kemdikbud.go.id/id/eprint/16742

OECD. (2018). *What 15-year-old students in Indonesia know and can do*. Paris: OECD Publishing. https://www.oecd.org/pisa/publications/PISA2018_CN_IDN.pdf

OECD. (2019). PISA 2018 assessment and analytical framework. Paris: OECD Publishing. https://doi.org/10.1787/b25efab8-en

Orr, M. J. L. (1996). *Introduction to radial basis function networks*. Edinburgh: *Edinburgh University*.

Ovan, Waluya, S. B., & Nugroho, S. E. (2018). Analysis mathematical literacy skills in terms of the students' metacognition on PISA-CPS model. *Journal of Physics: Conference Series*, *983*(1). https://doi.org/10.1088/1742-6596/983/1/012151

Patmaniar, P., Amin, S. M., & Sulaiman, R. (2021). Students' growing understanding in solving mathematics problems based on gender: Elaborating folding back. *Journal on Mathematics Education, 12*(3), 507-530. https://doi.org/10.22342/jme.12.3.14267.507-530

Pitsia, V., Biggart, A., & Karakolidis, A. (2017). The role of students' self-beliefs, motivation, and attitudes in predicting mathematics achievement: a multilevel analysis of the programme for international student assessment data. *Learning and Individual Differences*, *55*, 163-173. https://doi.org/10.1016/j.lindif.2017.03.014

Rahmawati, D., & Kismiantini. (2022). Gender differences in mathematics achievement, competitiveness, fear of failure, and resilience: Analysis of PISA 2018 in Indonesia. *AIP Conference Proceedings*, (Vol. 2575). https://doi.org/10.1063/5.0107819

Samnufida, R., & Kismiantini. (2022). How school size and student teacher ratio affecting the students' mathematics achievement. *AIP Conference Proceedings*, (Vol. 2575). https://doi.org/10.1063/5.0130178

Sateesh, B. G., & Suresh, S. (2013). Parkinson's disease prediction using gene expression- A projection-based learning meta-cognitive neural classifier approach. *Expert Systems with Applications*, *40*(5), 1519–1529. https://doi.org/10.1016/j.eswa.2012.08.070

Shen, W., Guo, X., Wu, C., & Wu, D. (2011). Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. *Knowledge-Based Systems*, *24*(3), 378–385. https://doi.org/10.1016/j.knosys.2010.11.001

Sing, J. K., Basu, D. K., Nasipuri, M., & Kundu, M. (2003). Improved k-means algorithm in the design of rbf neural networks. *Conference on Convergent Technologies for Asia-Pacific Region* (Vol. 2, pp. 841-845). Bangalore, India. https://doi.org/10.1109/TENCON.2003.1273297

Sowjanya, A. H. M., & Mrudula, M. O. (2015). Cluster ensemble approach for clustering mixed data. *International Journal of Computer Techniques*, 2(5), 43–51. http://www.ijctjournal.org/Volume2/Issue5/IJCT-V2I5P9.pdf

Tran, L. T., & Nguyen, T. S. (2021). Motivation and mathematics achievement: a Vietnamese case study. Journal on Mathematics Education, 12(3), 449-468. http://doi.org/10.22342/jme.12.3.14274.449-468

Vaitsis, C., Hervatis, V., & Zary, N. (2016). Introduction to big data in education and its contribution to the quality improvement processes. *Big Data on Real-World Applications, 113*, 58. http://dx.doi.org/10.5772/63896

Wutsqa, D. U., & Farhan, A. (2020). Lung cancer detection using the SOM-GRR based radial basis function neural network. *Journal of Physics: Conference Series*, *1581*(1). https://doi.org/10.1088/1742-6596/1581/1/012007

Wutsqa, D. U., & Fauzan, M. (2022). The hybrid model of radial basis function neural network and principal component analysis for classification problems. *Industrial Engineering and Management*

Systems, *21*(3), 409–418. https://doi.org/10.7232/iems.2022.21.3.409