Feature Selection for Analyzing Data Errors Toward Development of Household Big Data at the Sub-District Level Using Multi-Layer Perceptron Neural Network

https://doi.org/10.3991/ijim.v16i05.22523

Sumitra Nuanmeesri(✉)
Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand
sumitra.nu@ssru.ac.th

Abstract—This research aims to analyze the patterns of data errors in order to fulfill the data required for household big data development at the sub-district level in Thailand. Feature Selection and Multi-Layer Perceptron Neural Network were applied, while the data imbalance was solved by the SMOTE method and the comparison between the CFS feature selection method and Information Gain (IG) feature selection method. Afterward, the datasets were classified the data errors by the Multi-Layer Perceptron Neural Network. Each model's effectiveness was measured by the 10-fold cross-validation method. The research results revealed that the suitable data size after being adjusted data imbalanced was 400%. Once the data had been processed for developing the model, it was found that after being adjusted data size towards the application of the SMOTE, CFS feature selection technique, and classified data errors by the Multi-Layer Perceptron Neural Network, the model provided the highest level of effectiveness in data errors classification with an accuracy of 98.29%. Moreover, the application could effectively classify data errors and display the household big data at the highest level. The application evaluation results given by the experts and the users had an average mean of 4.69 and higher, a standard deviation of 0.47 and lower, which has the level of effectiveness of 93.78% and higher, while interquartile range values not over 1, a quartile deviation of no more than 0.5.

Keywords—big data, data errors, feature selection, MLP, sub-district, SMOTE

1 Introduction

The development of big data in the field of health, economics, environment, activities, developments, and household demographics is crucial for community development. This is because comprehensive and accurate data can demonstrate the community's genuine problems and demands in which the governmental agencies or responsible figures such as village leaders, subdistrict administrators, local people themselves, researchers, and the business sector can take advantage to solve the problems. Community demographics are considered a big data prototype linked with the national big data system, facilitating the data processing cycle and reflecting the genuine problems embedded in the data. Information is a highly valuable asset for any state or any agency because inequality issues between the rich and the poor could be solved by rapidly developing grassroots economies in a systematic and clear direction. Quite often, governmental development projects are not consistent or correspond with the local people's demand or cannot effectively solve problems. Community demographic data could contribute to the project proposals of the subdistrict development agencies since reliable data could be used for supporting their reports or budget proposals. Researchers, local people, and the public sector can altogether benefit from the data in terms of community and national developments.

In order to obtain the community data, it is necessary to get in touch or coordinate with community leaders. Primary data collection is required to develop big data for community development as well as the national development in each dimension. One of the most common problems while collecting community data is that the local people are hesitant to provide information. Even though both public and private sectors have tried to collect data from the local communities, local people rarely understand the overall picture because the analyzed data has not been accessible for the local people. Hence, they are reluctant to provide further information. As a result, the local people may not be able to see the whole picture or the real situations in their communities, leading to the inability to address the quick-response-needed issues. It is important to identify problems before resolving problems in the community.

Information technology at present offers many free services to establish online platforms and store data in the cloud system. For example, Google Forms can be instantly processed as long as there are internet connections. Although many areas lack internet access, it is still possible to record the data manually and then input it on Google Forms later. Collecting household data requires intervention in each household. However, some household leaders may lack literacy skills or have lousy eyesight, leading to an inability to read and fill in the questionnaire. Furthermore, this kind of project may also disturb the daily activities of the local people. Therefore, collecting household data for creating big data is considerably challenging. Nonetheless, when there is communication about the significance of data for local development, the local people are more likely to cooperate with the researchers. During the data collection process, it might have some problems. For instance, many data collectors were visiting each household, as shown in Figure 1. The data was recorded manually on the papers by data collectors before being input into the system, such as Google Forms by staff or officers. Therefore, the data was vulnerable to incorrect or repeated recordings. The fact that there are many respondents who may not have sufficient knowledge or literacy skills to understand the questions also leads to the same problems when inputting or processing data. Consequently, the charts in the application developed based on the deficient data may represent incorrect comparison results.

The goal of this study, hence, is to select features of data errors in order to support the system to fix data, provide accurate data for the users, and accurately predict features. As the collected dataset was small in size, the Synthetic Minority Over-sampling technique (SMOTE) was applied to adjust the data imbalance. Next, the dataset was used for selecting features towards two feature selection techniques, including Correlation-based Feature Selection and Information Gain. Once the features had been selected, the dataset was used for developing a model by the Multi-Layer Perceptron (MLP) Neural Network method. The model's effectiveness was measured by the 10-fold cross-validation method in order to use the model for developing a mini big data system of community information in Samut Songkhram Province, Thailand. The system is expected to be accessible anywhere and anytime by any user who has a smartphone. Local communities, governmental agencies, researchers, and the business sector can make use of this information for supporting and developing the communities in the future.

Fig. 1. Household data collection for establishing big data at the district level

2 Related work

The related works for applying the feature selection for analyzing data errors toward developing household big data at the sub-district level using Multi-Layer Perceptron Neural Network were described as follows.

2.1 Synthetic minority over-sampling technique

Classifying data which includes more than one class can lead to data imbalance problems and, eventually, an inaccurate classification that inclines towards majority classes. In this research, the Synthetic Minority Over-sampling technique (SMOTE) was applied to resolve the data imbalance issues. It is an empirical over-sampling algorithm, which extracts artificial samples from the minority class by inserting nearby existent samples. It is a method that resynthesizes data by increasing the class's data size as much as the biggest class. A value was randomized to find the distance between it and every other value, and then the closest value was selected. The resynthesized data is represented in (1) [1][2].

$$x_{nb} = x_o + R*(\hat{x}_{nb} - x_o) \tag{1}$$

where:
x_{nb} is the newly synthesized data,
x_o is the original random data,
$\hat{x}_{nb}$ is the nearest value or neighbor data,
R is the random value range from 0 to 1.

2.2 Feature selection

Feature selection is a technique that selects features of datasets before classifying them. There are many feature selection methods, but they all aim to select the significant data solely. The data whose features had been selected could be used for rapid model synthesization and effective data classification. In this study, two feature selection techniques were applied as follows.

1) Information Gain (IG) [3] is a feature selection technique which measures the gain value of each node. The node with the highest gain value will be selected as the root node. Then, the rest of the data will be measured again to find the next node. The information gain was calculated and represented in (2).

$$Gain(Y; X) = H(Y) - H(Y \mid X) \tag{2}$$

where:

Y is the feature value, which is a data class ranging between $\{Y_1, Y_2, \ldots, Y_n\}$ where n is the number of features,
X is the value of other features that are not classes ranging between $\{X_1, X_2, \ldots, X_n\}$ where n is the number of features,
$Gain(Y; X)$ is the score value gained from sample randomization ranging between 0 and 1,
$H(Y)$ is the probability value gained from the randomization of Y samples,
$H(Y \mid X)$ is the probability value gained from the randomization of Y samples when compared to X.
$H(Y)$ and $H(Y \mid X)$ are calculated in (3) and (4), respectively.

$$H(Y) = -\sum_{i=1}^{i=k} P(Y = y_i) \log_2 P(Y = y_i) \tag{3}$$

$$H(Y \mid X) = -\sum_{i=1}^{i=k} P(X = x_i) H(Y \mid X = x_i) \tag{4}$$

where:
$P(Y = y_i)$ is the probability value from y_1 to y_k,
$P(X = y_i)$ is the probability value from x_1 to x_k,
k is the number of features.

2) Correlation-based Feature Selection (CFS) is a feature selection technique that depends on the link between the collections of features acquired from the assessment of feature prediction capacity utilized for data classification and inconsequential data management. CFS can rank the information subsets dependent on the information measurements and select the information subsets dependent on the information measurements concerning high and low connections between classes. For any immaterial data or any information with a low degree of relationship, they will be rejected, the equivalent with complex data dimensions which will be barred from the data dimensions with a high level of relationship. The equation for assessing the subsets of CFS data dimensions has appeared in (5) [4].

$$M_{zc} = \frac{k\overline{r}_{zf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}} \tag{5}$$

where:
M_{zc} is the value of data dimension subset which composes of k data dimension,
k is the data dimension or features,
$\overline{r}_{ff}$ is the average value of the relationship of data dimension,
$\overline{r}_{zf}$ is the average value of the relationship between the variable and classes ($f \in S$).

2.3 Multi-layer perceptron neural network

Multi-Layer Perceptron Neural Network consists of the input layer, hidden layer, and output layer. In each layer, there are nodes or processing units. Moreover, there could be more than one layer hidden in each layer. The network applies the supervised learning pattern, which involves both Feed Forward and Backpropagation techniques, allowing the network to learn how to classify complex data. This network is widely used in the field of medicine. The operation of the Multi-Layer Perceptron Neural Network starts from inputting data into the input layer, followed by delivering the processed data to the output layer, as illustrated in Figure 2.

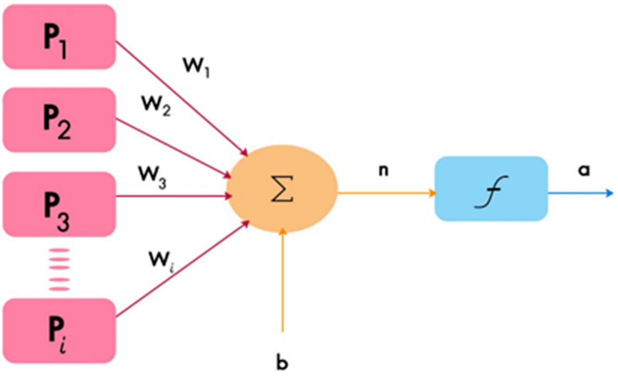


Fig. 2. Multi-layer perceptron neural network

Data processing requires the sum of the multiplication of inputs and weights, as shown in Equation 6. Next, the outputs shall be calculated by the Sigmoid function (See Equation 7. The outputs in the hidden layer will be transferred to the output layer where the processed outputs and the target outputs are compared. If the difference is not acceptable, the outputs will be sent back by the backpropagation process to the hidden layer and the input layer, respectively. Meanwhile, they also enter the weight adjustment process before being tested with the data. Finally, the outputs are then calculated with the Sigmoid function once again [5].

$$n = \sum_{i=1}^{k} P_i W_i \tag{6}$$

$$f_x = \frac{1}{1+e^{-x}} \tag{7}$$

where:
n is the sum total of input P_i multiplied by weight W_i value,
i is the number of inputs or weight value,
x is to the input value.

2.4 Literature review

According to the previous studies on the solutions to data imbalance of big data, the most commonly used resolution techniques included Random Over Sampling (ROS), Random Under Sampling (RUS), and Synthetic Minority Over-sampling Technique (SMOTE), especially in case of the "double layer" scenarios. Nevertheless, with regards to big data cases, multiple layered imbalances have not been comprehensively explored; there have been only a few examinations so far. In this study, the model's effectiveness was analyzed under the circumstances of multiple layered imbalances of big data by the SMOTE technique. The analysis results showed that it is necessary to slightly increase the overall effectiveness of the classifiers to the non-random datasets [6].

The Synthetic Minority Oversampling Technique has been applied to resolve the data imbalance by the random sampling method, together with the data cleaning techniques such as neighborhood adjustment or Tomek's link in terms of big data. The results of competency and probability analysis of the heuristic sampling method based on the deep-learning Multi-Layer Perceptron Neural Network in the big data domain showed that the most effective classification could emerge when there was the application of the data cleaning process with the ANN output instead of using only the input attribute space. It can be seen that the adjustment of imbalanced classes could be applied to deep learning and big data scenarios [7].

Apparently, big data class imbalance resolutions have been adapted from conventional methods, especially sampling methods [8][9]. Nonetheless, recent research demonstrates that some conclusions drawn by the machine learning process were not applicable to the context of big data. To illustrate, it is normal for machine learning that SMOTE can provide better results than ROS [10]. However, in some cases, the results did not represent similar trends in big data contexts [11][12]. Additionally, not so many previous studies have focused on big data class imbalance resolution towards the application of "intelligent" or heuristic sampling methods [13][14].

Therefore, this research points out there should be more studies on how to resolve the data imbalance problems and select suitable features for predicting data errors. Effective machine learning performances can solve data imbalance problems by using heuristic sampling algorithms with regards to the scale of big data, including subdistrict household demographics in Thailand. The research results could be further applied to the research and development of grassroots economies, which can reduce poverty and inequality in society. The findings can also be used for preparing accurate, in-depth information at the household level of the country, as they can formulate a big data system displayed in the form of graphs that clearly illustrate the comparisons of the data.

Both public and private sectors, including the local communities, can access the data, which is provided via an online application, in order to develop the country together.

3 Methodology

The model development consists of 6 stages, including 1) preprocessing for data transformation, 2) data imbalance adjustment by SMOTE, 3) feature selection by CFS and Information Gain (IG), 4) model creation by Multi-Layer Perceptron Neural Network, 5) model's effectiveness measurement by 10-fold cross-validation, and 6) development and deployment of the application, as illustrated in Figure 3.

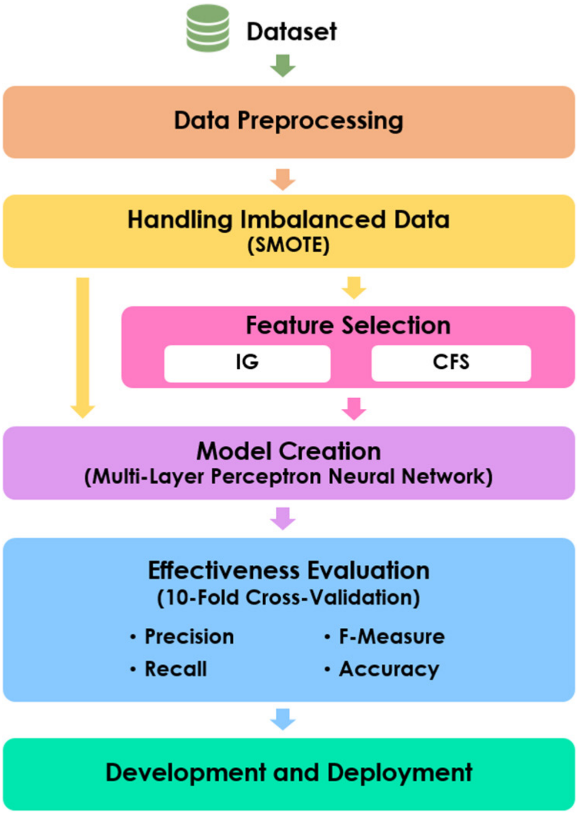


Fig. 3. The overall research framework

3.1 Data preprocessing

This study collected household demographic data from 22 villages in 3 subdistricts, including Kradangnga and Bang Khonthi in Khonthi District and Bang Kaeo in Mueang District of Samut Songkhram Province. The household demographic data involved general information of family members, health data, economic data, environment and surroundings data, local activities data, and local development data. These data contain 124 attributes or features in a total of 2,845 records, with approximately 14% of incomplete or inaccurate data, including duplicates, missing data, incorrect input, typo error, and inconsistent data or violated attribute dependency. The data was transformed into a comma-separated value (.CSV) file to be processed by Weka version 3.9 later.

3.2 Handling imbalanced data by SMOTE

The preprocessed data was found to be imbalanced in the solution classes. Thus, the research applied the Synthetic Minority Over-sampling technique (SMOTE), which is a data re-synthesization method, to increase the number of randomized datasets with small classes by increasing the K value from 1–5 as an experiment. The results showed that the most effective K value was 5, and the random seed was 1. Then, the data size was increased starting from 100% until it reached the highest effective value measured by the 10-fold cross-validation. The experiment results showed that the most suitable data size was 400%. Thus, the data size increased from 2,845 records to 2,990 records, as illustrated in Figure 4.

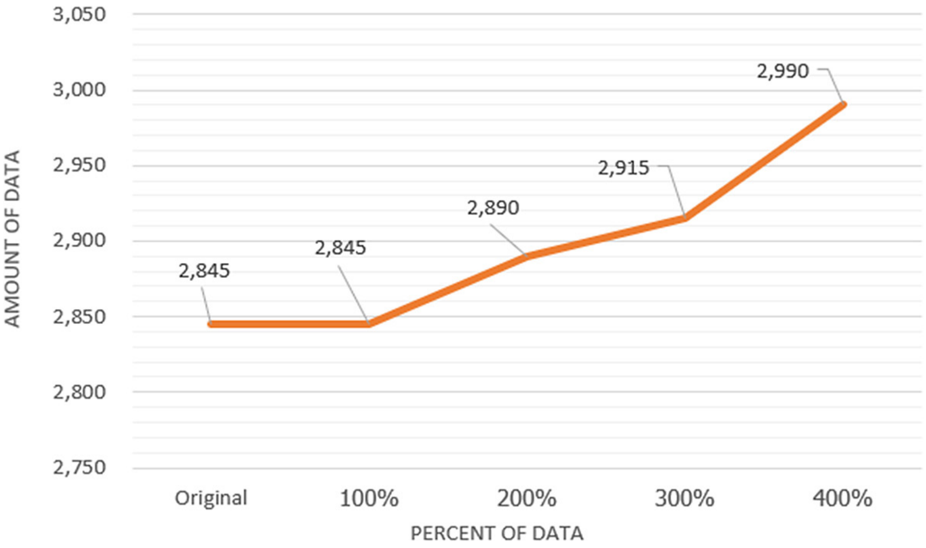


Fig. 4. The adjustment of the data using the SMOTE method

3.3 Feature selection by CFS and IG

This research applied the preprocessed and adjusted-size data towards SMOTE and then inputted the data to the feature selection processes using the CFS and the IG in Weka. In this section, there are ten groups of data applied in this process, consisting of 1) the original dataset + CFS, 2) the original dataset + IG, 3) 100% of SMOTE + CFS,

4) 200% of SMOTE + CFS, 5) 300% of SMOTE + CFS, 6) 400% of SMOTE + CFS, 7) 100% of SMOTE + IG, 8) 200% of SMOTE + IG, 9) 300% of SMOTE + IG, and 10) 400% of SMOTE + IG. These resulting ten datasets were processed by feature selection technique will be used in modeling at the next process.

3.4 Model creation by multi-layer perceptron neural network

The adjusted-size data was delivered to the learning process in order to create a model by which the research team employed two feature selection techniques, CFS and IG, together with the Multi-Layer Perceptron Neural Network. The research team specified the parameters for the Multi-Layer Perceptron Neural Network model creation as follows: Hidden Layer = 4, Training Time = 500, Learning Rate = 0.3, and Momentum = 0.2. These parameters provided the highest effective results as measured by 10-fold cross-validation.

3.5 Effectiveness evaluation of the model

The developed model's effectiveness was measured by the four effectiveness evaluation methods: the precision, the recall, the F-measure, and the accuracy [14][15][16][17]. Here are the equations that represented the tests of the model's effectiveness.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

where:
TP refers to when the targeted class is "Yes" and the model predicts it "Yes".
TN refers to when the targeted class is "No" and the model predicts it "No".
FP refers to when the targeted class is "No" but the model predicts it "Yes".
FN refers to when the targeted class is "Yes" but the model predicts it "No".

3.6 Development and deployment of the application

This study developed an application that can be operated on both computers and smartphones based on the web application. It was scripted in PHP, HTML5, jQuery JavaScript-based. In addition, the Bootstrap framework was based on a cascade style sheet (CSS) and the jQuery function was also applied to the component arrangement on the screen design and user interface for displaying the output on both computers and mobile devices responsively. The XAMPP was set up and run to manage the MySQL database and Apache web server. The main workflow of the system is connected to the Weka software's instruction set through Java for creating or loading the best-performing MLP model for use in data error handling. The PHP script controlled the overall processes between client and server and communicated via extensible markup language (XML) formatted. The developed application infrastructure is illustrated in Figure 5.

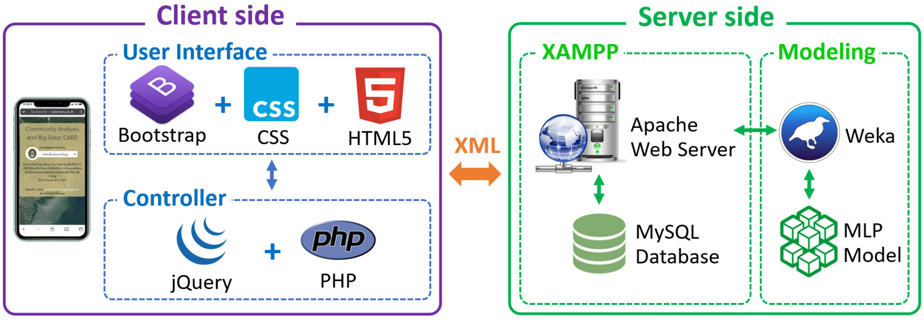
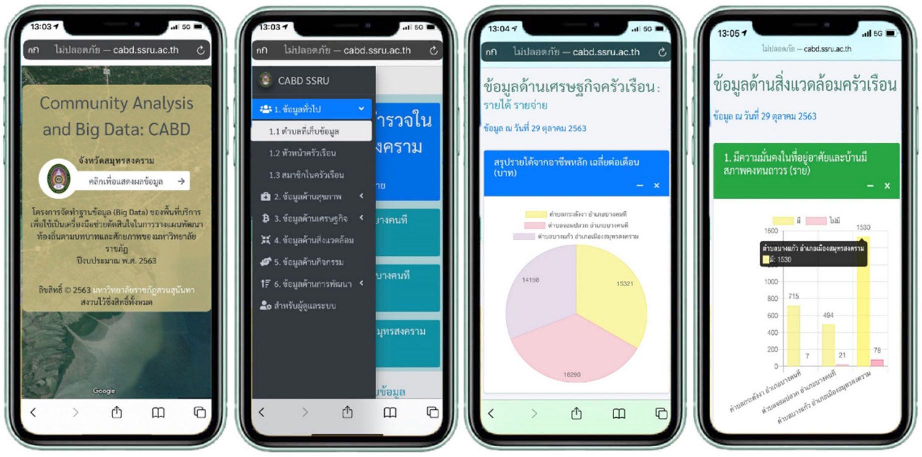


Fig. 5. The developed application infrastructure

The application development was divided into two major dimensions of users: the power users and the general users. The power users have rights for massive data processing, including exporting, importing, detecting, and managing data errors in the database. In case of detecting incorrect data, the system will run semi-automatically by displaying a message advising users to edit the correct information or automatically let the system take care of it. For general users, they are allowed to enter data individually into the database. Therefore, the system will immediately notify the users if any data errors are found in the data input process, such as data duplicates, missing values. Figure 6 illustrates the display of the application system on a smartphone.


Community Analysis and Big Data: CABD

จังหวัดสมุทรสงคราม

คลิกเพื่อแสดงผลข้อมูล

โครงการจัดทำฐานข้อมูล (Big Data) ของพื้นที่บริการเพื่อใช้เป็นเครื่องมือช่วยตัดสินใจในการวางแผนพัฒนาท้องถิ่นตามบทบาทและศักยภาพของมหาวิทยาลัยราชภัฏ
ปีงบประมาณ พ.ศ. 2563

ลิขสิทธิ์ © 2563 มหาวิทยาลัยราชภัฏสวนสุนันทา สงวนไว้ซึ่งสิทธิ์ทั้งหมด

ข้อมูลด้านเศรษฐกิจครัวเรือน: รายได้ รายจ่าย

ข้อมูล ณ วันที่ 29 ตุลาคม 2563

สรุปรายได้จากอาชีพหลัก เฉลี่ยต่อเดือน (บาท)

ข้อมูลด้านสิ่งแวดล้อมครัวเรือน

ข้อมูล ณ วันที่ 29 ตุลาคม 2563

1. มีความมั่นคงในที่อยู่อาศัยและบ้านมีสภาพคงทนถาวร (ราย)

Fig. 6. The example of the user interface of the application on mobile devices

After developing the application to correct the data errors and display the subdistrict household big data of 3 subdistricts in Samut Songkhram Province, Thailand, a 10-minute video was filmed alongside creating a user guide to the developed application. They were published on Facebook, a popular online platform in Thailand with more than 45 million accounts [18]. This video aimed to help the users actively and continuously learn how to use the application. They could play, pause, repeat, fast forward or backward anytime they wanted to learn how to use the application. There were different icons and the 'Help' section that showed the advice and how to use the application for the users to learn by themselves. If the users have any questions, they could send an inquiry or videocall the operators via Facebook services.

Before the mobile application was evaluated, the training had been provided to 46 samples who voluntarily participated in the mobile application testing advertised on social media. In this research, the documents describing the protocols and research ethics were sent to the participants, asking them to give their consent to participate in the study.

The Black-box testing evaluated the developed system with five criteria indicators, including functional testing, compatibility testing, usability testing, performance testing, and security testing. All forty-six participants evaluated the system based on a 5-point Likert scale [19], as illustrated in Table 1.

Table 1. The scoring based on the Likert scale

Scoring	Weighted Mean	Level of Effective
5	4.51–5.00	The highest
4	3.51–4.50	The high
3	2.51–3.50	The medium
2	1.51–2.50	The little
1	1.00–1.50	The least

In addition, the developed system was assessed in quartiles (Q), including the first quartile (Q1), the third quartile (Q3), the interquartile range (IQR), and the quartile deviation (QD), based on mean, standard deviation (SD), median (MED), and percentage.

4 Research results

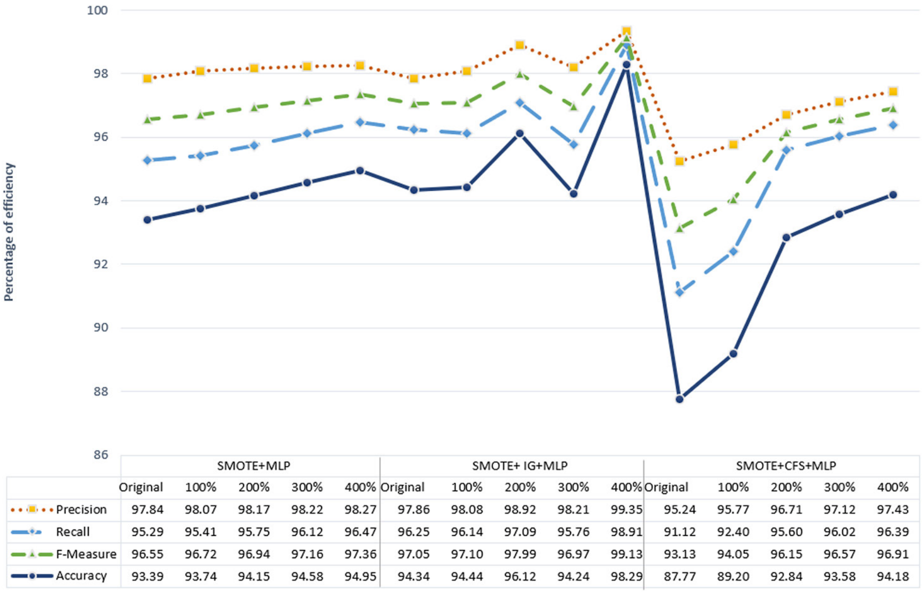
In this experiment, the results of the model's effectiveness and application evaluation and can be summarized as follows:

4.1 The results of data analysis by feature selection and multi-layer perceptron neural network

After adjusting the data imbalance towards the SMOTE method, the most suitable data size was 400%. The balanced data was then used to select features by comparing the effectiveness with two feature selection techniques, including CFS and IG. Multi-Layer Perceptron Neural Network was applied to create the model, and the 10-fold cross-validation method was employed to measure the model. The results of the model evaluation are illustrated in Table 2 and Figure 7.

Table 2. The results of model evaluation

Techniques	% of SMOTE	Precision	Recall	F-Measure	Accuracy
SMOTE+MLP	Original	97.84	95.29	96.55	93.39
	100%	98.07	95.41	96.72	93.74
	200%	98.17	95.75	96.94	94.15
	300%	98.22	96.12	97.16	94.58
	400%	98.27	96.47	97.36	94.95
SMOTE+CFS+MLP	Original	97.86	96.25	97.05	94.34
	100%	98.08	96.14	97.10	94.44
	200%	98.92	97.09	97.99	96.12
	300%	98.21	95.76	96.97	94.24
	400%	**99.35**	**98.91**	**99.13**	**98.29**
SMOTE+ IG+MLP	Original	95.24	91.12	93.13	87.77
	100%	95.77	92.40	94.05	89.20
	200%	96.71	95.60	96.15	92.84
	300%	97.12	96.02	96.57	93.58
	400%	97.43	96.39	96.91	94.18



	SMOTE+MLP					SMOTE+ IG+MLP					SMOTE+CFS+MLP				
	Original	100%	200%	300%	400%	Original	100%	200%	300%	400%	Original	100%	200%	300%	400%
Precision	97.84	98.07	98.17	98.22	98.27	97.86	98.08	98.92	98.21	99.35	95.24	95.77	96.71	97.12	97.43
Recall	95.29	95.41	95.75	96.12	96.47	96.25	96.14	97.09	95.76	98.91	91.12	92.40	95.60	96.02	96.39
F-Measure	96.55	96.72	96.94	97.16	97.36	97.05	97.10	97.99	96.97	99.13	93.13	94.05	96.15	96.57	96.91
Accuracy	93.39	93.74	94.15	94.58	94.95	94.34	94.44	96.12	94.24	98.29	87.77	89.20	92.84	93.58	94.18

Fig. 7. The comparison of the model's effectiveness evaluation

Moreover, the author compares the performance of the data error handling method for big data with other studies, as shown in Table 3. The scope of data error handling includes data duplicates (DUP), incorrect input (IC), missing value (MV), typo error (TPE), and inconsistent data or violated attribute dependency (VAD).

Table 3. The comparison of the data error handling method for big data

Method	Dataset	No. of Features	Handling	Performance
MLClean [20]	Adult Census Income [21]	15	DUP	Accuracy of 78.00%
Corleone [22]	Restaurants [23]	6	DUP	Recall of 96.30%
CrowdER [24]	Restaurants [23]	6	DUP	Recall of 100.00%
HoloClean [25]	Flights [26]	6	TPE, VAD	Recall of 67.00%
Baran [27]	Flights [26]	6	TPE, VAD	Recall of 100.00%
HoloClean [25]	Hospital [28]	19	TPE, VAD	Recall of 71.00%
NADEEF [29]	Hospital [28]	19	TPE, VAD	Recall of 92.80%
BigDansing [30]	Hospital [28]	19	TPE, VAD	Recall of 92.90%
Baran [27]	Hospital [28]	19	TPE, VAD	Recall of 88.00%
Mean-Median imputation with Decision Tree [31]	Apollo Hospitals [31]	25	MV	Accuracy of 100.00%
SMOTE+CFS+MLP (This study)	Household	124	DUP, IC, MV, TPE, VAD	Accuracy of 98.29% (Recall of 98.91%)

4.2 Effectiveness evaluation results of the application

The application was evaluated towards the black-box testing by 9 experts in the field of information technology and community development and 37 users composed of community leaders, local people, researchers, and community developers.

The results of application evaluation by the nine experts towards the black-box testing revealed that the functional testing and compatibility testing had the highest mean value of 4.78 with a standard deviation of 0.44, while the usability testing and performance testing had a mean value of 4.67 and a standard deviation of 0.50. The security testing had a mean value of 4.56 with a standard deviation of 0.33. Overall, it had a mean value of 4.69 and a standard deviation of 0.47. In other words, the effectiveness evaluation results provided by the experts showed that the application was effective at the highest level.

Meanwhile, the results of application evaluation by the 37 users towards the black-box testing showed that the functional testing had the highest mean value of 4.84 with a standard deviation of 0.37. The compatibility testing, usability testing, and performance testing had a mean value of 4.81 with a standard deviation of 0.40. The Security testing had a mean value of 4.73 and a standard deviation of 0.49. Overall, evaluated by the users, it had a mean value of 4.80 and a standard deviation of 0.40. That is to say, the effectiveness evaluation results provided the users showed that the application was effective at the highest level.

Additionally, the developed system was also analyzed in terms of conformity by both the experts and the users in quartiles. The results suggested that the interquartile range values were not over 1, and the quartile deviation had no more than 0.5, indicating that all participants had the same opinion and evaluated the application in the same manner, as demonstrated in Table 4.

Table 4. The results of the application evaluation

Indicators	Mean	SD	Quartiles			IQR	QD	Level of Effective
			Q1	MED	Q3			
Experts								
Functional testing	4.78	0.44	5.00	5.00	5.00	0.00	0.00	The highest (95.56%)
Compatibility testing	4.78	0.44	5.00	5.00	5.00	0.00	0.00	The highest (95.56%)
Usability testing	4.67	0.50	4.00	5.00	5.00	1.00	0.50	The highest (93.33%)
Performance testing	4.67	0.50	4.00	5.00	5.00	1.00	0.50	The highest (93.33%)
Security testing	4.56	0.53	4.00	5.00	5.00	1.00	0.50	The highest (91.11%)
Total	**4.69**	**0.47**	**4.00**	**5.00**	**5.00**	**1.00**	**0.50**	**The highest (93.78%)**
Users								
Functional testing	4.84	0.37	5.00	5.00	5.00	0.00	0.00	The highest (96.76%)
Compatibility testing	4.81	0.40	5.00	5.00	5.00	0.00	0.00	The highest (96.22%)
Usability testing	4.81	0.40	5.00	5.00	5.00	0.00	0.00	The highest (96.22%)
Performance testing	4.81	0.40	5.00	5.00	5.00	0.00	0.00	The highest (96.22%)
Security testing	4.73	0.45	4.00	5.00	5.00	1.00	0.50	The highest (94.59%)
Total	**4.80**	**0.40**	**5.00**	**5.00**	**5.00**	**0.00**	**0.00**	**The highest (96.00%)**

5 Conclusion

Dealing with data errors is challenging for big data, including missing data, incorrect input, typo error, inconsistent data, or violated attribute dependency. Therefore, this research proposed the technique for analyzing and classifying the data errors using feature selection and Multi-Layer Perceptron Neural Network. Additionally, the original dataset with imbalanced data was synthesized the minority class based on SMOTE up to 400%. Thus, the dataset was increased from 2,845 records to 2,990 records. Further, these datasets were processed by applying the feature selection technique between CFS and IG methods to compare the results. All datasets were classified the data errors based on Multi-Layer Perceptron Neural Network. Finally, each model's effectiveness was evaluated by the 10-fold cross-validation technique.

It can be concluded from the research findings that the most suitable MLP model was the dataset that adjusted the data imbalanced was 400% of SMOTE and applied the CFS method. This model provided the highest effectiveness in data errors classification with an accuracy of 98.29%. The results showed that the application of CFS improved the accuracy of the model better than IG. Therefore, the most suitable model could be used to develop the application for data error handling and displaying household big data of the developing subdistrict household in Thailand. Moreover, the effectiveness of the developed application was evaluated by experts and users. It was shown that the application had the highest level of effectiveness. All mean of 4.56 and over on each indicator which has a standard deviation of not more than 0.53. Besides, the interquartile range values were not over 1, the quartile deviation was no more than 0.5, and the percentage was higher than 93%. All of the above showed that the development of sub-district household big data in Thailand successfully analyzed data errors by CFS feature selection technique and Multi-Layer Perceptron Neural Network, with SMOTE data imbalance adjustment. Therefore, the developed application which based on MLP and CFS can help users process large amounts of data or enter data to be more accurate and help perform data cleansing in big data for the household at the sub-district level. However, this research does not cover all of them in detail when analyzing character-level errors in in-depth and unstructured data. Still, this study can only classify which data records have errors in attributes and how to correct them. Additionally, this research supports many attributes or features which have errors in the dataset.

Future studies should explore the development of data error classifying techniques to fulfill the data collected on social media and unstructured data, which can be further used to record household information to develop big data of diverse agricultural occupations and productions. Then compare the runtime with published studies.

6 Acknowledgment

I would like to express our gratitude to the Institute for Research and Development, Suan Sunandha Rajabhat University, who offered us the opportunities to conduct this research.

7 References

[1] Tsipouras, M. G. (2018). Uterine EMG signals spectral analysis for pre-term birth prediction. Engineering, Technology & Applied Science Research. 8(5): 3310–3315. https://doi.org/10.48084/etasr.2146

[2] Paranya, P. (2016). Improving decision tree technique in imbalanced data sets using SMOTE for internet addiction disorder data. Information Technology Journal. 12(1): 54–62.

[3] Puripat, T. (2016). Ensemble algorithm for feature selection. M.Sc. Thesis in Computer Science. Thammasat University, Bangkok: Thailand.

[4] Mark, A. H. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis. The University of Waikato. New Zealand.

[5] Boubaker, S., Kamel, S., & Kchaou, M. (2020). Prediction of daily global solar radiation using resilient-propagation artificial neural network and historical data: A case study of Hail, Saudi Arabia. Engineering, Technology & Applied Science Research. 10(1): 5228–5232. https://doi.org/10.48084/etasr.3278

[6] González-Barcenas, V. M., Rendón, E., Alejo, R., Granda-Gutiérrez, E. E., & Valdovinos, R. M. (2019). Addressing the big data multi-class imbalance problem with oversampling and deep learning neural networks. Springer International Publishing. https://doi.org/10.1007/978-3-030-31332-6_19

[7] Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega F. J., & Granda-Gutiérrez, E. E. (2020). Data sampling methods to deal with the big data multi-class imbalance problem. Applied Sciences. 10(1276): 1–15. https://doi.org/10.3390/app10041276

[8] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer International Publishing: Cham, Switzerland. https://doi.org/10.1007/978-3-319-98074-4

[9] García-Gil, D., Holmberg, J., García, S., Xiong, N., & Herrera, F. (2020). Smart data based ensemble for imbalanced big data classification. 2001(5759): 1–25.

[10] Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., & Herrera, F. (2019). An analysis of local and global solutions to address big data imbal-anced classification: A case study with smote preprocessing. In Cloud Computing and Big Data. 75–85. https://doi.org/10.1007/978-3-030-27713-0_7

[11] Reyes-Nava, A., Cruz-Reyes, H., Alejo, R., Rendón-Lara, E., Flores-Fuentes, A. A., & Granda-Gutiérrez, E. E. (2019). Using deep learning to classify class imbalanced gene-expression microarrays datasets. In: Vera-Rodriguez R., Fierrez J., Morales A. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP 2018). Lecture Notes in Computer Science. 11401. Springer, Cham. https://doi.org/10.1007/978-3-030-13469-3_6

[12] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal Big Data. 5: 42. https://doi.org/10.1186/s40537-018-0151-6

[13] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A., Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access. 4: 7940–7957. https://doi.org/10.1109/ACCESS.2016.2619719

[14] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-measure to ROC. Informedness & Correlation Journal of Machine Learning Technologies. 2(1): 37–63.

[15] Nuanmeesri, S. (2020). Mobile application for the purpose of marketing, product distribution and location-based logistics for elderly farmers. Applied Computing and Informatics. https://doi.org/10.1016/j.aci.2019.11.001

[16] Gadebe, M. L., & Kogeda, O. P. (2020). Top-K human activity recognition dataset. International Journal of Interactive Mobile Technologies. 14(18): 68–86. https://doi.org/10.3991/ijim.v14i18.16965

[17] Nurhaida, I., Noprisson, H., Ayumi, V., Wei, H., Putra, E. D., Utami, M., & Setiawan, H. (2020). Implementation of deep learning predictor (LSTM) algorithm for human mobility prediction. International Journal of Interactive Mobile Technologies. 14(18): 132–144. https://doi.org/10.3991/ijim.v14i18.16867

[18] Thai Social Media Users in 2020 World's Ranking. (2020). Retrieved February 12, 2021, from https://marketeeronline.co/

[19] Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, New York University.

[20] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019). Data cleaning for accurate, fair, and robust models: A big data—AI integration approach. arXiv:1904.10761. https://doi.org/10.1145/3329486.3329493

[21] Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.

[22] Gokhale, C, Das, S., Doan, A. H., Naughton, J. F., Rampalli, N., Shavlik, J. & Zhu, X. (2014). Corleone: Hands-off crowdsourcing for entity matching. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. 2014, Association for Computing Machinery: Snowbird, Utah, USA. 601–612. https://doi.org/10.1145/2588555.2588576

[23] Bilenko, M. (2003). RIDDLE: Repository of information on duplicate detection, record linkage, and identity uncertainty. Retrieved February 12, 2021, from https://www.cs.utexas.edu/users/ml/riddle/data.html

[24] Wang, J., Kraska, T., Franklin, M. J., & Feng, J. (2012). CrowdER: Crowdsourcing entity resolution. arXiv:1208.1927. https://doi.org/10.14778/2350229.2350263

[25] Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. arXiv:1702.00820. https://doi.org/10.14778/3137628.3137631

[26] Li, X., Dong, X. L., Lyons, K., Meng, W. & Srivastava, D. (2013). Truth finding on the deep web: Is the problem solved? arXiv:1503.00303. https://doi.org/10.14778/2535568.2448943

[27] Mahdavi, M. & Abedjan, Z. (2020). Baran: effective error correction via a unified context representation and transfer learning. In Proceedings of VLDB Endow. 13(12): 1948–1961. https://doi.org/10.14778/3407790.3407801

[28] Chu, X., Ilyas, I. F., & Papotti, P. (2013). Holistic data cleaning: Putting violations into context. In 2013 IEEE 29th International Conference on Data Engineering, 458–469. https://doi.org/10.1109/ICDE.2013.6544847

[29] Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., & Tang, N. (2013). NADEEF: A commodity data cleaning system. In SIGMOD. 541–552. https://doi.org/10.1145/2463676.2465327

[30] Khayyat, Z., Ilyas, I F., Jindal, A., Madden, S., Ouzzani, M., Papotti, P., Quiané-Ruiz, J.-A., Tang, N., & Yin, S. (2015). BigDansing: A System for Big Data Cleansing, in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Association for Computing Machinery: Melbourne, Victoria, Australia. 1215–1230. https://doi.org/10.1145/2723372.2747646

[31] Sessa, J., & Syed., D. (2016). Techniques to deal with missing data. In 2016 5th International Conference on Electronic Devices, Systems and Applications. https://doi.org/10.1109/ICEDSA.2016.7818486

8 Author

Sumitra Nuanmeesri is lecturer at Suan Sunandha Rajabhat University (SSRU), Bangkok 10130 Thailand. She received the Ph.D. in information technology from King Mongkut's University of Technology North Bangkok. Her research interests include speech recognition, data mining, deep learning, machine learning, image processing, web and mobile application, supply chain management system, augmented reality (AR) and virtual reality (VR) development, robotics, and the internet of things (IoT).

Article submitted 2021-03-08. Resubmitted 2021-10-15. Final acceptance 2021-10-30. Final version published as submitted by the authors.