International Journal of Research and Review
Vol.7; Issue: 4; April 2020
Website: www.ijrrjournal.com
Research Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Genetic Algorithm Optimization Through Handling of Incomplete Data and Reduction

Sri Novida Sari[1], Pahala Sirait[2], Arwin Halim[2]

[1]Postgraduate Student at STMIK Mikroskil, Medan, Indonesia, 20212
[2]Lecturer Student at STMIK Mikroskil, Medan, Indonesia, 20212

Corresponding Author: Sri Novida Sari

ABSTRACT

Cervical cancer is a leading gynecologic malignancy worldwide. There are still many weaknesses in the data processing system in handling risk factors for cervical cancer. This research presents optimization techniques and shows the selection of features for the best combination of attributes in the risk of cervical cancer. There are thirty two attributes with eight hundred fifty eight samples. In addition, this data also has incomplete values due to respondents' privacy issues and imbalance data. Therefore, preprocessing techniques are used with regression imputation and data normalization methods. Furthermore, attribute reduction techniques with the optimum index factor (OIF) method are needed to improve the accuracy of the results. The results of research using the Genetic Algorithm method show that the comparison of optimization with reduction and without reduction of attributes has a difference of 1%. Optimization with no attribute reduction results in 95% while results with attribute reduction are 94%.

Keywords: Preprocessing, Incomplete data, regression imputation, optimum index factor, Genetic Algorithm, Optimization.

INTRODUCTION

The development of information technology has entered the healthcare world, ranging from service system to diagnosis of disease-based specialist systems. However, early detection of existing types of diseases is still limited so that more people or organizations use algorithms to analyse large collections of data in diagnosing or predicting something, and this kind of technology is also Widely used, such as the medical field usually predicts some severe diseases in the early stages, for example, cervical cancer. Cervical cancer is one of the more health-threatening diseases of women (Deng et al., 2019) Cervical cancer is the cause of significant deaths in low-income countries, more than half a million cases per year, and kills Over a quarter million in the same period (Fernandes et al., 2017).

From the study of the literature shows that the quality of disease diagnosis is still found many weaknesses such as the limitation of disease data to the accuracy of the diagnosis results, this is so by various factors such as lack of knowledge data is owned by the system, the incompleteness of data (incomplete data), and the method of optimization is less precise. As in previous research, Choudhury et al compares optimization 4 methods namely Gaussian Naive Bayes (GNB), Decision tree (DT), Support Vector Mechines (SVM) and Logistic Regregession (LR) in which the research Decision tree (DT) is much more Superior to other optimization methods. With the similar datasets of Wu, W., and Zhou, H. Also conducted a comparative optimization of the method to support vector mechine-recursive feature elimination (SVM-RFE) and Support Vector Mechine Principal Component Analysis (SVM-PCA), with the results of SVM-PCA having Better ability though with the same number of features.

Another problem that arises is that there are many attributes that have an impact on the length of time of decision-making, in recent years, many detection methods are proposed and applied in the field to provide timely diagnosis, including Data-driven approaches (Wu, W., & Zhou, H., 2017). In this study the Optimum Index Factor (OIF) method was used to reduce the 32 attribute by measuring the statistical value of the existing optimal combination selection. But the data-driven approach should also be supported with complete data, in order to avoid missing data value. An optimization issue for incomplete data is a challenge for data research development because the most widely used technique assumes complete data but does not conform to Incomplete data problems (Yildirim., 2018). In the study Yildirim conducted a comparative analysis of subsitution of average value with an ensemble algorithm to address incomplete data as well as evaluation of results based on accuracy size and execution time. In this study presents the process of incomplete data with the model regression Imputation, which is expected to achieve a relatively better accuracy and precision value.

Most of the research is also trying to get the right method to increase the accuracy value, one of which proposes an optimization method for some of the available datasets. The optimization of medical data with the accuracy of the results-through the optimization approach Genetic Algorithm has been conducted in the research of Gorzałczany and Rudzinski (2017), in the research of a kind Wissanu Thungrut and Naruemon (2019) also do the same thing that Optimize dataset Diabetes with Genetic Algorithm method, the results can prove that the method of optimization Genetic Algorithm is superior to problem solving similar, therefore in this research, the author proposes Genetic method Algorithm on process optimization.

LITERATURE REVIEW

Cervical cancer

Cervical cancer is one of the leading causes of cancer deaths among women. Around the world, cervical cancer is the fourth most common violence in women, and produces about 530,000 new cases each year with 270,000 deaths. Approximately 85% of deaths worldwide due to cervical cancer occur in low-income or developing countries, and mortality rates are 18 times higher in low-and middle-income countries compared to wealthy countries.

Machine Learning

According to Shwartz and David (2014) Machine Learning is a study of the algorithm to learn something in doing certain things that are done by humans automatically. Learning in this regard relates to how to complete various tasks, or to create a prediction of accurate new conclusions from the various patterns that have been learned before.

a. Incomplete data

Incomplete data is an incompleteness of data that affects the statsistic value of the data. This ' missing data ' case will result in an incomplete data (incomplete data) in a model, thereby inhibiting the statistical analysis to be performed. Missing data will be an important issue on the case of dataset optimization. In general the method of imputation and optimization techniques are two different things but the important point of optimization is how to get good training data. Because in addition to the proper selection of methods, the accuracy of the results of the optimization is influenced by the characteristics and completeness of an instance of a data (Acuna, 2004).

a) Approach Incomplete

An optimization issue for incomplete data is a challenge for data research development because the most widely used technique assumes complete data but does not conform to Incomplete data problems. In the study Yildirim conducted a comparative analysis of substitution of average value with an ensemble algorithm

to address incomplete data as well as evaluation of results based on accuracy size and execution time (Yildirim., 2018).

b) Regression Imputation method

In the regression imputation method, the regression model is estimated to predict the values observed and mounted for each variable with the missing values (Yuan 2001).

Next suspect missing data by calculating values from parameters $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ and $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\boldsymbol{2}}$ first.

Suspect parameters $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ used the smallest quadratic method of equation, or

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\boldsymbol{x'x})^{-\mathbf{1}}(\boldsymbol{x'y}) \qquad (2.1)$$

Next look for the value of variances

$$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\boldsymbol{2}} = \frac{\boldsymbol{y'y}-\widehat{\boldsymbol{\beta}}'\boldsymbol{x'y}}{n-p} \qquad (2.2)$$

Stages of use of imputation on variables X_2. Calculating a new Parameter variance value β_* and σ_{*m}^2 gained from

$$\sigma_{*m}^2 = \frac{\widehat{\sigma}_i^2(n-p)}{g_m} \qquad (2.3)$$

g is a random value of the chi-square distribution and is the value of unlost observation.

Calculating regression coefficient

$$\beta_* = \beta + \sigma_{*m}{L_{ij}}^T Z \qquad (2.4)$$

Which σ_{*m} is the root of σ_{*m}^2 obtained from (1), Z is a randomly-sized random variable vector with normal distribution $N(0,1)$ and ${L_{ij}}^T$ is a matrix of upper triangle from decomposition Cholesky. According Rencher (1934), Cholesky decomposition is used if the matrix V is a symmetrical matrix that is $V = V^T$, $V = (x'x)^{-1}$.

So the data value is lost in variables X_2 Be suspected by using the equation.

$$Y_{mis} = \beta_{*0} + \beta_{*1}X_1 + sE \qquad (2.5)$$

Reduction

Reduction is a process by which data obtained from the field is carried out reduction, compiled the essential things and focused on the important things and arranged systematically with the aim that the data become more Easy to understand and control. The Data that has been reduced will give a clearer picture of the research results in the field. In this reduction process researchers are not the origin of reducing the data but do the selection or choose what data is relevant and meaningful. Focusing on solving the problem of discovery of the use or answering of research questions, the data reduction process takes place continuously during ongoing research (Moleong, 2006:288).

a) Attribute reduction approach

A common problem is the many attributes that exist, impacting the length of decision-making time. In recent years, many detection methods are proposed and applied in the field to provide timely diagnosis, including a data-driven approach (Wu, W., & Zhou, H., 2017).

b) The Optimum Index Factor (OIF) reduction

Optimum Index factor (OIF) is one of the statistical methods to choose the optimal combination based on the total standard deviation and koefesien correlation between bands. The OIF formula is as follows (Sirait, P., & Arymurthy, 2010):

$$OIF = \frac{\sum_{h=1}^{3} S_h}{\sum_{h=1}^{3} abs(f,g)} \qquad (2.6)$$

Description:

Sh : Standard deviation for channels h

abs (f.g) : Absolute value for any channel

OIF : Optimum Index Factor

Genetic Algorithm

The algorithm was discovered at the University of Michigan, USA by John Holland (1975) through a study and popularized by one of his disciples, David Goldberg (1989). Where to present this Genetic Algorithm as a method of search algorithms based on natural selection mechanism and genetic nature. Genetic Algorithm is an algorithm that seeks to implement an understanding of the natural evolution of problem-solving tasks (problem solving). The approach taken by this algorithm is to randomly combine a wide selection of the best solutions in a group to get the next generation of best solutions in a condition that maximizes the match or Often called fitness (Desiani et al., 2006).

This generation will represent improvements to its initial population. By

doing this process repeatedly, the algorithm is expected to simulate the evolutionary process. In the end, you will get the most appropriate solutions for the problems faced. To use Genetic Algorithm, the problem solution is represented as a khromosom. Three important aspects for the use of Genetic Algorithm (Desiani et al, 2006):

a) The Defenisi fitness function
b) The defendant and the implementation of genetic representation
c) The defendant and the implementation of genetic surgery

MATERIALS & METHODS

Methodology

An overview of the steps done in this study will begin from inputting a whole dataset of cervical cancer risk factors. The next step will be to process the incomplete data in which to input the missing value by regression imputation method. Next the second stage will be continued for the normalization process followed by the third stage of the attribute reduction. Then the fourth stage will be done process optimization with Genetic Algorithm method. Once all processing stages are completed, the next stage will be analysis of the results obtained by comparing the accuracy, precision, and Recall values of those results.

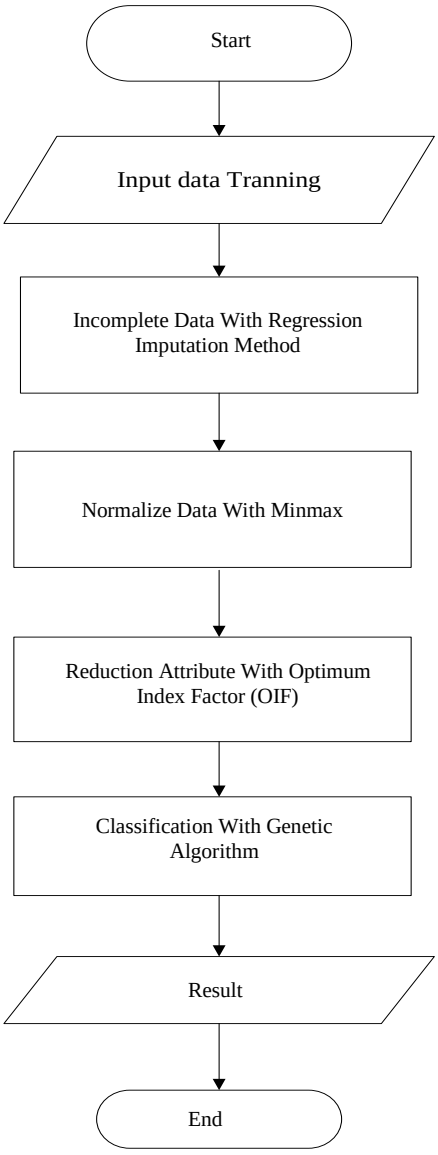

Figure 1 Methodology

RESULT

Imputation Of Incomplete Data With Regression Imputation

At this stage, there will be data input that is not complete with rapid miner. The results can be seen in the following image:

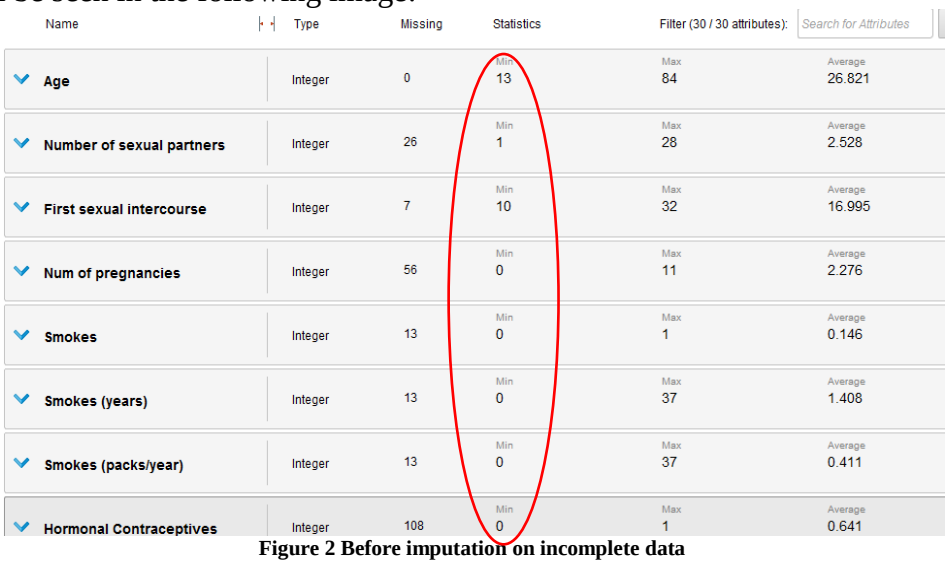

Figure 2 Before imputation on incomplete data

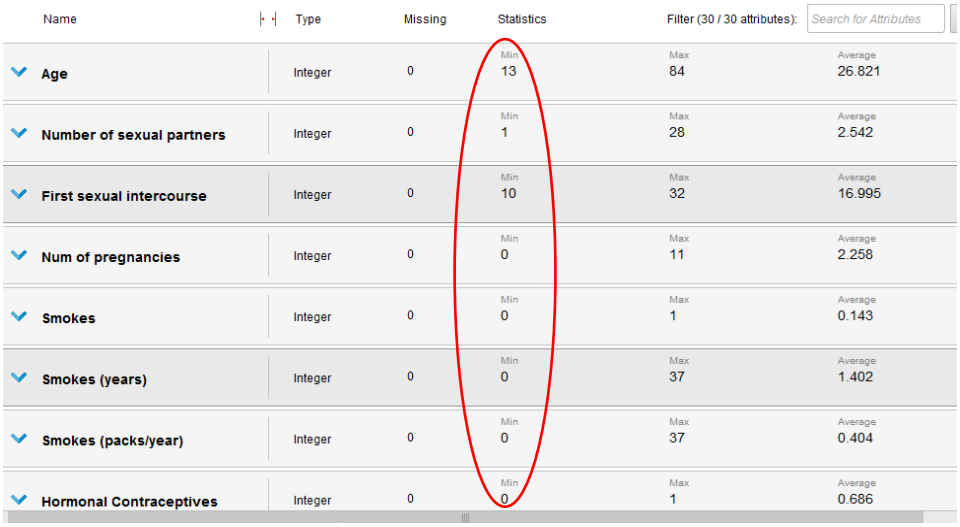

Figure 3 After imputation with Regresion Imputation

Normalization with Min-max

At this stage, the dataset that has been imputated will be normalized with Min-Max, due to the data value that has the difference in interval (value range) is quite far, in the neighboring data bandwidth (related data). The results can be seen in the following image:

Table 2 Data before normalization

Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)
18.00	4.00	15.00	1.00	0.00	3.00	2.00
15.00	1.00	17.50	1.00	1.00	2.00	3.00
34.00	0.00	15.00	1.00	2.00	1.50	1.00
52.00	5.00	16.00	4.00	2.00	3.00	0.00
46.00	0.00	17.40	4.00	0.00	2.00	3.00
42.00	3.00	23.00	2.00	1.00	2.00	2.00
51.00	3.00	17.00	6.00	1.00	2.00	4.00
26.00	1.00	26.00	0.00	1.00	8.70	3.00
45.00	1.00	20.00	5.00	0.00	0.00	0.00
44.00	3.00	15.00	2.36	1.00	12.00	2.80
44.00	3.00	26.00	4.00	0.00	0.00	0.00

Table 3 Data after normalization

Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)
0.07	0.14	0.44	0.09	0.00	0.09	0.08
0.03	0.04	0.52	0.09	0.50	0.06	0.12
0.30	0.00	0.44	0.09	1.00	0.05	0.04
0.55	0.18	0.47	0.36	1.00	0.09	0.00
0.46	0.00	0.52	0.36	0.00	0.06	0.12
0.41	0.11	0.70	0.18	0.50	0.06	0.08
0.54	0.11	0.51	0.55	0.50	0.06	0.16
0.18	0.04	0.80	0.00	0.50	0.27	0.12
0.45	0.04	0.60	0.45	0.00	0.00	0.00
0.44	0.11	0.44	0.21	0.50	0.38	0.11
0.44	0.11	0.80	0.36	0.00	0.00	0.00

Attribute Reduction Testing Results

At this stage the data that has been normalized will process through the reduction phase of the attribute where the entire attribute will be process using the optimum index factor (OIF) method. So in the get the best attribute combinations of the 32 attributes are as follows:

1. Reduction of 1 attribute with 31 combinations

From 32 The attribute is done 31 combinations so that in can result as follows:

Table 4 Reduction 1 attribute

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	12	0.289131	22	0
3	0.0935	13	0.132327	23	0.143314
4	0.127361	14	0.225253	24	0.03412
5	0.182497	15	0	25	0.048224
6	0.124159	16	0.076115	26	0.100789
7	0.08186	17	0.220572	27	0.080287
8	0.463809	18	0.143314	28	0.078329
9	0.125237	19	0.03412	29	0.150886
10	0.295599	20	0.03412	30	0.10188
11	0.097211	21	0.03412	31	0.150886
				32	0.164893

From table you get 31 combinations that produce OIF = 8190936314439.64 with 1 attribute reduction of Number of sexual partners.

2. Reduction of 2 attributes with 30 combinations

From 32 The attribute is done 30 combinations so that in can result as follows:

Table 5 Reduction Of 2 Attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	23	0.143314
3	0.0935	14	0.225253	24	0.03412
4	0.127361	15	0	25	0.048224
5	0.182497	16	0.076115	26	0.100789
6	0.124159	17	0.220572	27	0.080287
8	0.463809	18	0.143314	28	0.078329
9	0.125237	19	0.03412	29	0.150886
10	0.295599	20	0.03412	30	0.10188
11	0.097211	21	0.03412	31	0.150886
12	0.289131	22	0	32	0.164893

From table you get the result of 30 combinations that generate OIF = 20275004097744.5 value by reducing the 2 attributes of Smokes (Packs/year).

3. Reduction of 3 attributes with 29 combinations

From 32 The attribute is done 29 combinations so that in can result as follows:

Table 6 Reduction Of 3 Attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	23	0.143314
3	0.0935	14	0.225253	24	0.03412
4	0.127361	15	0	26	0.100789
5	0.182497	16	0.076115	27	0.080287
6	0.124159	17	0.220572	28	0.078329
8	0.463809	18	0.143314	29	0.150886
9	0.125237	19	0.03412	30	0.10188
10	0.295599	20	0.03412	31	0.150886
11	0.097211	21	0.03412	32	0.164893
12	0.289131	22	0		

From table gained 29 combinations resulting in OIF = 103994219938707.84 value with 3 attribute reduction i.e. STDs: Number of diagnosis

4. Reduction of 4 attributes with 28 combinations

From 32 The attribute is done 28 combinations so that in can result as follows:

Table 7 Reduction of 4 Attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	24	0.03412
3	0.0935	14	0.225253	26	0.100789
4	0.127361	15	0	27	0.080287
5	0.182497	16	0.076115	28	0.078329
6	0.124159	17	0.220572	29	0.150886
8	0.463809	18	0.143314	30	0.10188
9	0.125237	19	0.03412	31	0.150886
10	0.295599	20	0.03412	32	0.164893
11	0.097211	21	0.03412		
12	0.289131	22	0		

From table, there are 28 combinations that result in OIF = 412544795700876.75 value with 4 attribute reduction of STDs: HIV.

5. Reduction 5 attributes with 27 combinations

From 32 The attribute is done 27 combinations so that the results can be as follows:

Table 8 Reduction of 5 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	22	0
3	0.0935	14	0.225253	24	0.03412
4	0.127361	15	0	26	0.100789
6	0.124159	16	0.076115	27	0.080287
8	0.463809	17	0.220572	28	0.078329
9	0.125237	18	0.143314	29	0.150886
10	0.295599	19	0.03412	30	0.10188
11	0.097211	20	0.03412	31	0.150886
12	0.289131	21	0.03412	32	0.164893

From table, you get 27 combinations that produce OIF = 1108915686361230.5 with 5 attributes reduction of Smokes.

6. Reduction of 6 attributes with 26 combinations

From 32 The attributes are performed 26 combinations so that in can result as follows:

Table 9 Reduction of 6 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	24	0.03412
3	0.0935	14	0.225253	26	0.100789
4	0.127361	15	0	27	0.080287
6	0.124159	16	0.076115	28	0.078329
8	0.463809	17	0.220572	29	0.150886
9	0.125237	18	0.143314	30	0.10188
10	0.295599	19	0.03412	31	0.150886
11	0.097211	20	0.03412	32	0.164893
12	0.289131	21	0.03412		

From table, there are 26 combinations that result in OIF = 1108915686361230.5 value with 6 attribute reduction of STDs: AIDS.

7. Reduction 7 attributes with 25 combinations

From 32 The attributes are performed 25 combinations so that in can result as follows:

Table 10 Reduction 7 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	26	0.100789
3	0.0935	14	0.225253	27	0.080287
4	0.127361	16	0.076115	28	0.078329
6	0.124159	17	0.220572	29	0.150886
8	0.463809	18	0.143314	30	0.10188
9	0.125237	19	0.03412	31	0.150886
10	0.295599	20	0.03412	32	0.164893
11	0.097211	21	0.03412		
12	0.289131	24	0.03412		

From table, Obtained 25 combinations that produce OIF = 1108915686361230.5 value with a reduction of 7 attributes, namely STDs: Cervical condylomatosis.

8. Reduction of 8 attributes with 24 combinations

From 32 The attribute is done 24 combinations so that in can result as follows:

Table 11 Reduction of 8 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	12	0.289131	24	0.03412
3	0.0935	13	0.132327	26	0.100789
4	0.127361	14	0.225253	27	0.080287
6	0.124159	16	0.076115	28	0.078329
8	0.463809	18	0.143314	29	0.150886
9	0.125237	19	0.03412	30	0.10188
10	0.295599	20	0.03412	31	0.150886
11	0.097211	21	0.03412	32	0.164893

From table, obtained results 24 combinations that produce OIF = 982274108598563.5 value with a reduction of 8 attributes namely Vulvo-perineal condylomatosis.

9. Reduction 9 attributes with 23 combinations

From 32 attributes done 23 combinations so that in can result as follows:

Table 12 Reduction of 9 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	12	0.289131	24	0.03412
3	0.0935	13	0.132327	27	0.080287
4	0.127361	14	0.225253	28	0.078329
6	0.124159	16	0.076115	29	0.150886
8	0.463809	18	0.143314	30	0.10188
9	0.125237	19	0.03412	31	0.150886
10	0.295599	20	0.03412	32	0.164893
11	0.097211	21	0.03412		

From table, obtained 23 combinations resulting in OIF = 214681812350125.12 value with a reduction of 9 attributes, namely Vulvo-perineal condylomatosis.

10. Reduction of 10 attributes with 22 combinations

From 32 The attribute is done 22 combinations so that the results can be as follows:

Table 13 Reduction of 10 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	13	0.132327	27	0.080287
3	0.0935	14	0.225253	28	0.078329
4	0.127361	16	0.076115	29	0.150886
6	0.124159	18	0.143314	30	0.10188
8	0.463809	19	0.03412	31	0.150886
10	0.295599	20	0.03412	32	0.164893
11	0.097211	21	0.03412		
12	0.289131	24	0.03412		

From table, there are 22 combinations that result in OIF = 103492254294081.81 value with 10-attribute reduction of Hormonal Contraceptives (years).

11. Reduction of 11 attributes with 21 combinations

From 32 The attributes are performed 21 combinations so that in can result as follows:

Table 14 Reduction of 11 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	12	0.289131	24	0.03412
3	0.0935	13	0.132327	27	0.080287
4	0.127361	14	0.225253	28	0.078329
6	0.124159	16	0.076115	29	0.150886
8	0.463809	19	0.03412	30	0.10188
10	0.295599	20	0.03412	31	0.150886
11	0.097211	21	0.03412	32	0.164893

From table 4.14, there are 21 combinations that result in OIF = 299628431180378.7 with the reduction of 11 attributes, namely STDs: Syphilis.

12. Reduction of 12 attributes with 20 combinations
From 32 The attribute is done 20 combinations so that in can result as follows:

Table 15 Reduction of 12 attributes

Attribute	Standard Deviation	Attribute	Standard Deviation	Attribute	Standard Deviation
1	0.11962	12	0.289131	24	0.03412
3	0.0935	13	0.132327	27	0.080287
4	0.127361	14	0.225253	28	0.078329
6	0.124159	16	0.076115	30	0.10188
8	0.463809	19	0.03412	31	0.150886
10	0.295599	20	0.03412	32	0.164893
11	0.097211	21	0.03412		

From table 4.15, there are 20 combinations resulting in OIF = 6,826,926,378,341,489 with a reduction of 12 attributes, Dx: Cancer.
Best combination: [1, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 24, 26, 27, 28, 29, 30, 31, 32] with the best combination of 25 combinations and the maximum value gained is 1108915686361230.5, the attribute being reduced at this stage is [1 , 7, 25, 23, 5, 22, 15].

Optimization test results with Genetic Algorithm

The tests will be conducted using a trial of the data that has been in the imputation and the reduction of the attributes with the data that has been in the imputation without the reduction of the attributes, the experiment was done with 800 data testing and 58 data tranning in the extract from the top and bottom Dataset. Optimization is done with 2 targets, Hinselmann and Schiller.
Testing will use accuracy and performance measurements as a comparison of results:

1. Testing with complete data without attribute reduction

Testing done with complete data without attribute reduction, so the number of attributes is still 32, with the number of feature selection cost [2.2897624507670153], the determination of cost value in can of standard deviation from each attribute and with the result Accuracy of 0.95. For more details can be seen in the table below:

Table 16 The GA Optimization Test Without Reduction Of Attributes

Attribute	Cost	Attribute	Standard Deviation	Attribute	Cost
1	0.11961964	13	0.132327	23	0.143314
3	0.093500264	14	0.225253	24	0.03412
4	0.127361328	15	0	25	0.048224
5	0.182496791	16	0.076115	26	0.100789
6	0.124158994	17	0.220572	27	0.080287
8	0.463809311	18	0.143314	28	0.078329
9	0.12523705	19	0.03412	29	0.150886
10	0.295598761	20	0.03412	30	0.10188
11	0.09721098	21	0.03412	31	0.150886
12	0.289131219	22	0	32	0.164893

2. Testing optimization with reduction

Testing conducted with complete data with attribute reduction, so that the number of attributes 25 of the best combination of the 7 attributes is reduced, with the number of feature selection cost [1.3951494739999997], determination of the value of the cost also can be from the standard deviation from each attribute. For more details can be seen in the table below.

Table 17 of GA-Reduction Testing With Attribute Reduction

combination	un-classified sample	accuracy	recall
31	31	0,94	0,91
30	31	0,93	0,91
29	31	0,90	0,91
28	31	0,93	0,91
27	31	0,94	0,91
26	33	0,95	0,97
25	31	0,94	0,91
24	31	0,93	0,91
23	31	0,95	0,91
22	33	0,95	0,97
21	33	0,95	0,97
20	31	0,93	0,91

In table can be seen that the best combination that can be measured from the value of the recall accuracy as well as the consideration of the sample that can not be optimized then the result can be from a combination is 25 combinations.

DISCUSSION

Incomplete Data and Normalization

In the previous part has been conducted test of imputation on incomplete data by using regression imputation, where the total amount of data is 858 with 32 attributes, after done Imputasi on incomplete data then data then in Normalization using the Minmax method. Results obtained from the regression imputation method where the data incomplete obtained from the number of correlates on each row there is the whole dataset.

Reducing attribute with Optimum Index Factor (OIF)

The attribute reduction is done with the value of data already in incomplete and in normalization where the value of data is averaged later in search of variance value and then get standard deviation. After the cost or standard deviation is obtained then look for the correlation of the whole attribute, then get the best combination results. Of the 32 attributes in the get 25 best combinations so 7 attributes in the reduction according to the above discussion. The test results can be seen as follows:

Table 18 Combination OIF

Combination	OIF
31	8.19E+12
30	2.03E+13
29	1.04E+14
28	4.13E+14
27	1.11E+15
26	1.11E+15
25	1.11E+15
24	9.82E+14
23	2.15E+14
22	1.03E+14
21	3E+14
20	6.83E+12

From table 15 can be seen that the best combination of 32 combinations are 25 combinations are the best combination = [1, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 24, 26, 27, 28, 29, 30, 31, 32] with the 1108915686361230.5 best OIF value of

Optimization Genetic Algorithm

In the Genetic Algorithm optimization use complete data and comparison with the testing of data that has been reduced and data before the reduction. These comparisons can be a different amount of cost or weight, and the comparison of the samples cannot be optimised. The final test results can be seen in the table below:

Table 19 Results of Optimization Genetic Algorithm

combination	un-classified sample	accuracy	recall
31	31	0,94	0,91
30	31	0,93	0,91
29	31	0,90	0,91
28	31	0,93	0,91
27	31	0,94	0,91
26	33	0,95	0,97
25	31	0,94	0,91
24	31	0,93	0,91
23	31	0,95	0,91
22	33	0,95	0,97
21	33	0,95	0,97
20	31	0,93	0,91

The analysis results are obtained according to the 4.19 table, namely:

1. From the test results can be seen that the verification results of the genetic algorithm optimization of the data that is complete and in normalization but without the attribute reduction has an average accuracy value of 95% while the optimization results with the reduction of attributes The best OIF combination is 94%.
2. Then from the attributes that have been reduced and without the reduction of attributes apparently the result of the optimization process does not overaffect that is with the results of a comparison of 0.01% so that with data already reduced attributes allow more efficient and Effective in processing compared to 32 attributes without reduction.

CONCLUSION

Throughout the results of the tests that have been done to imputasi incomplete data using

regression imputation and reduction of attributes using Optimum index factor (OIF) as well as genetic algorithm optimization. The following conclusions are obtained:

1. In the process of reduction attribute obtained the number of 25 best attribute combinations as well as 7 attributes are reduced with the maximum value of Optimum Index Factor (OIF).
2. From the optimization test that has been obtained the results of optimization genetic algorithm from the data that is complete and in normalization but without the attribute reduction has an average accuracy value of 95% while from the results of the optimization with reduction The best OIF combination attribute is 94%.
3. On the process obtained from the optimization with genetic algorithm, the value of accuracy with reduction and without attribute reduction has a comparison accuracy of 0.01% thus allowing more efficient and effective in the processing process compared to 32 attribute without reduction.

REFERENCES

1. Bahadure, N. B., Ray, A. K., & Thethi, H. P. (2018). Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm. Journal of digital imaging, 31(4), 477-489.
2. Ceylan, Z., & Pekel, E. (2017). Comparison of multi-label classification methods for prediagnosis of cervical cancer. International Journal of Intelligent Systems and Applications in Engineering, 5(4), 232-236.
3. Choudhury, A., Wesabi, Y. M., & Won, D. (2018). Classification of Cervical Cancer Dataset. arXiv preprint arXiv:1812.10383.
4. Deng, X., Luo, Y., & Wang, C. (2019, April). Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods. In 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS) (pp. 631-635). IEEE.
5. Desiani, Anita dan M. Arhami, 2006. Konsep Kecerdasan Buatan. Yogyakarta: ANDI.
6. Exner, M., Kühn, A., Stumpp, P., Höckel, M., Horn, L. C., Kahn, T., & Brandmaier, P. (2016). Value of diffusion-weighted MRI in diagnosis of uterine cervical cancer: a prospective study evaluating the benefits of DWI compared to conventional MR sequences in a 3T environment. Acta Radiologica, 57(7), 869-877.
7. Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017, June). Transfer learning with partial observability applied to cervical cancer screening. In Iberian conference on pattern recognition and image analysis (pp. 243-250). Springer, Cham.
8. Handayani, I. F. (2015). Perbandingan Karakteristik Dan Pengetahuan Tentang Kanker Serviks Pada Wanita Dengan Inspeksi Visual Asam Asetat (Iva) Positif Di Pesisir Dan Perkotaan. Jurnal Kebidanan, 7(01).
9. Hidayat, E., Sari, D. H., & Fitriyati, Y. (2014). Hubungan Kejadian Kanker Serviks Dengan Jumlah Paritas Di RSUD Dr. Moewardi Tahun 2013. Jurnal Kedokteran dan Kesehatan Indonesia, 6(3), 128-136.
10. Kartikawati Erni, 2013. Bahaya Kanker Payudara dan Kanker Serviks. Bandung :Buku Baru
11. Kessler, T. A. (2017, May). Cervical cancer: prevention and early detection. In Seminars in oncology nursing (Vol. 33, No. 2, pp. 172-183). WB Saunders.
12. Krisno Agus, 2011. Kajian Mikrobiologi Kesehatan. Hubungan Kandidiasis dengan Kanker Serviks, 11 Januari 2011
13. Kusumadewi, S, 2003. Artificial Intelligence (Teknik dan Aplikasinya), Yogyakarta: Graha Ilmu
14. Nurwijaya.2010. Hubungan Pemakaian Alata Kontrasepsi Dengan Resiko Terjadinya Kanker Serviks Menggunakan Metode IVA Di Puskesmas Seragen. Stikes Kusuma Husada Surakarta.
15. Ramana, B. V., & Boddu, R. S. K. (2019, January). Performance Comparison of Classification Algorithms on Medical Datasets. In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0140-0145). IEEE.
16. Rasjidi, I., 2014. Manual Prakanker Serviks, edisi 1, Sagung Seto, Jakarta.
17. Sasongko, T. B. (2016). Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Optimasi Jalur

Minat SMA). Jurnal Teknik Informatika dan Sistem Informasi, 2(2).
18. Sirait, P., & Arymurthy, A. M. (2010, August). Cluster centres determination based on KD tree in K-Means clustering for area change detection. In 2010 International Conference on Distributed Frameworks for Multimedia Applications (pp. 1-7). IEEE.
19. Small Jr, W., Bacon, M. A., Bajaj, A., Chuang, L. T., Fisher, B. J., Harkenrider, M. M., ... & Gaffney, D. K. (2017). Cervical cancer: a global health crisis. Cancer, 123(13), 2404-2412.
20. Stewart, T. S., Moodley, J., & Walter, F. M. (2018). Population risk factors for late-stage presentation of cervical cancer in sub-Saharan Africa. Cancer epidemiology, 53, 81-92.
21. Thungrut, W., & Wattanapongsakorn, N. (2018, July). Diabetes classification with fuzzy genetic algorithm. In International Conference on Computing and Information Technology (pp. 107-114). Springer, Cham.
22. Wahyuningsih, T., & Mulyani, E. Y. (2014). Faktor risiko terjadinya lesi prakanker serviks melalui deteksi dini dengan metode IVA (inspeksi visual dengan asam asetat). In Forum ilmiah (Vol. 11, No. 2, pp. 192-209).
23. Wu, W., & Zhou, H. (2017). Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access, 5, 25189-25195.
24. Yildirim, P. (2018, June). Classification with incomplete data and ensemble learners for the prediction of cervical cancer risk. In Proceedings of the International Conference on Intelligent Science and Technology (pp. 1-5). ACM/
25. Zhang, K., Gao, H., Han, X., Cai, Z., & Li, J. (2019). Modeling and Computing Probabilistic Skyline on Incomplete Data. IEEE Transactions on Knowledge and Data Engineering.

How to cite this article: Sari SN, Sirait P, Halim A. Genetic algorithm optimization through handling of incomplete data and reduction. International Journal of Research and Review. 2020; 7(4): 346-357.

******