Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 1 2024, pp. 51 – 61

Optimization Of Neural Network Method Using Chi-Square Feature Selection In Poverty Data Classification

Tresi Aprilia*[1]
Faculty of Computer Science, Dian Nuswantoro University
Semarang, Indonesia
*E-mail : p31202102413@mhs.dinus.ac.id **[1]*

Received 8 December 2015; Revised 10 February 2016; Accepted 2 March 2016

Abstract - Poverty is a fundamental problem that has become the center of attention in several aspects, for example from the government. The government needs poverty data and analyzes it to determine which poverty alleviation programs should be delivered to the right target or the poor. The aim of this study is to determine the accuracy of the classification of poverty in Batang District using the Neural Network method using the chi square feature selection. The dataset used in this study uses poverty data sourced from the Batang district BPS based on the results of the Susenas survey (National Economic Survey) for the 2022 time period. The results of this study indicate that the accuracy obtained for poverty classification using a neural network is 96.38% , with a precision value of 100%, and a recall value of 89.38%. Whereas when using a neural network with feature selection chi square, it gets an accuracy value of 93.68%, with a precision value of 91.07%, and a recall value of 90.26%. The contribution of this research is to develop a neural network method using feature selection chi square to improve the results of the accuracy of the classification is not poor or poor.

Keywords - Classification, Neural Network, Feature Selection, Chi Square

1. INTRODUCTION

In the process of measuring poverty data, the Central Statistical Agency approach carried out by BPS uses the concept of ability to meet basic needs (basic needs approach). The data used to calculate poverty in district or city level areas is data from the National Economic Survey of Consumption and Expenditure for the month of March of 2022. The method of poverty calculation used is to calculate the poverty line (GK), which consists of two components: the Food Poverty Line (GKM) and the Non-Food Povertyline (NKNM) [1].

The classification process for this study will be done using data from poor subjects collected in Tibawa district using data mining techniques. Attributes to be used to classify individuals include Age, Education, Activity, Income, Dependency, and Status (Wife/Unwife). The method to use is the Naive Bayes Classifier, which is the most important single classification technique in data mining. Based on the research carried out, the classification system of the general public in the territory of Pemda Tibawa Kab. Gorontalo can be implemented and used. The use of nave bayes classification methods against data sets taken from the research objects obtained an accuracy of 73% [2].

In order to determine to whom the schizophrenia program should be targeted, the government needs to collect and analyze the scizophrenic data. Data collection is carried out periodically using 14 survey variables. To identify a range of values in each category, we use k-means clustering at 57,522 data points to identify the range of value in each cluster or category. After defining the cluster value range, you can directly send the label output of the clustering using a single category defined by the value range of each cluster. The results of the evaluation of clustering are almost poor, poor and very poor. Accuracy was 13%, 46%, and 41% [3]. The data used was obtained from the Badan Pusat Statistik (BPS) of Banjar district, South

Kalimantan in 2014. However, this study only processed data on households in the area of Aranio district, which is 289 data. Subsequently, it tested the data of the Aranio district's poverty variables using the Neural Network method with various testing models, namely experimentation with testing of different models with a combination of Training Cycle, Learning Rate, Momentum, Number of Validation, and Sampling Type used. We will then compare the levels of Accuracy, Precision, and Recall. From various tests, the conclusion is that the Neural Network model, Training Cycle = 200, Learning Rate = 0,1, Momentum = 0,2, Number of Validation = 6, and Sampling Type = Strarified, yields a better degree of accuracy than other models. The model obtained yielded an Accuracy value of 89.97% +/- 3.46% [4].

Choosing the right algorithm for the attributes that exist in classifying between the loyal and the unloyal of a customer is essential. Feature selection is an important part of maximizing the performance of the classifier. Reducing large feature space on feature selection is the basis for creating performance optimization, for example by eliminating inaccurate or irrelevant attributes. Using proper feature selection algorithms can improve accuracy. The Neural Network Method is a classification machine that is modeled and used to mimic the biological structures of human nerves [5].

A feature selection algorithm can be distinguished into two types, namely, a filter and a wrapper. Examples of the type of a filter are information gain (IG), chi-square, and log likelihood ration. The accuracy result of the wrapper type is higher than the filter type, but this result is achieved with bright complexity. High complexity problems can also cause problems. According to Yang and Padersen in 1997, their research results showed that IG and chi-square obtained better results than the proposed Bi-Normal Separation method [6]

The research carried out in Batang district using poverty data on BPS Batang District, the data can be used to do classification about the number of poverty after the outbreak of Covid 19 in Batangan District. During this time in the settlement of poor family data in Batang District has not yet been done a study in detail, then on this penelitin, in performing the determination of a poverty family data using Neural Network method to obtain more accurate data using neural network method that is optimized with feature selection chi square [7]. From the method of classification of mining data, the data that has been prepared will be tested in order to obtain the best pattern or model in determining classification determination of poor family data of Batang district that can be grouped or tagged based on variables obtained from the National Economic Survey (SUSENAS) Period March 2022 on Batang District BPS.

2. RESEARCH METHOD

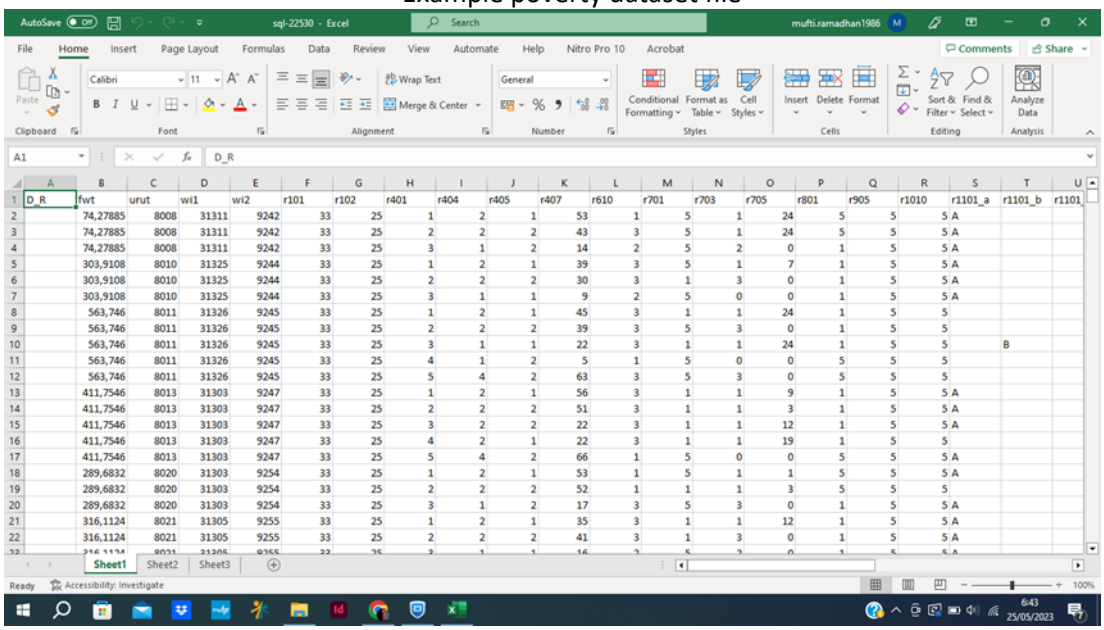
2.1 Dataset

The data set in this study is obtained from the poverty data in Batang district BPS. This data is based on the results of the survey conducted by the Central Statistical Agency every one semester per year, namely the National Economic Survey (SUSENAS) Period March 2022. This data set is private and not publicly accessible because it is micro- and confidential. However, it can only be accessed through a link through some procedures that have been specified by BPS through https://silastik.bps.go.id/. From the results of SUSENAS survey period March 2022 in Batang district, there is a record number of 83 census blocks. Where each 1 census block consists of 10 heads of families. So, the total record of the datasets used in SUSENAS is 830 heads of family data [8].

The process of obtaining the dataset begins with making a transaction to the BPS. The process from each stage is about 2 to 5 working days. Then proceed with the transaction verification. Where in the process, the customer gets a message sent via e-mail that aims to verify the accuracy of the customer's data [9]. In the process of uploading the file, it is required to fill in the data that must be completed about the data itself and the background of the

reasons why you want to obtain the data and also supplemented with the proof of the payment transaction of the invoice that has been sent by the admin silastik. Where these transaction funds go into the state's treasury. Then at the verification stage this file includes the upload process of the SPPD (Letter of Data Use Agreement) which has been signed by the parties concerned with the creation of the material. The goal is that later on, the data provided can be used properly and responsibly without misuse of data for criminal acts. Because the data is private or confidential. If the data is well verified, then the poverty data is successfully downloaded and used as research data.

Example poverty dataset file



D_R	fwt	urut	wi1	wi2	r101	r102	r401	r404	r405	r407	r610	r701	r703	r705	r801	r905	r1010	r1101_a	r1101_b	r1101
	74,27885	8008	31311	9242	33	25	1	2	1	53	1	5	1	24	5	5	5	A		
	74,27885	8008	31311	9242	33	25	2	2	2	43	3	5	1	24	5	5	5	A		
	74,27885	8008	31311	9242	33	25	3	1	2	14	2	5	2	0	1	5	5	A		
	303,9108	8010	31325	9244	33	25	1	2	1	39	3	5	1	7	1	5	5	A		
	303,9108	8010	31325	9244	33	25	2	2	2	30	3	1	3	0	1	5	5	A		
	303,9108	8010	31325	9244	33	25	3	1	1	9	2	5	0	0	1	5	5	A		
	563,746	8011	31326	9245	33	25	1	2	1	45	3	1	1	24	1	5	5			
	563,746	8011	31326	9245	33	25	2	2	2	39	3	5	3	0	1	5	5			
	563,746	8011	31326	9245	33	25	3	1	1	22	3	1	1	24	1	5	5		B	
	563,746	8011	31326	9245	33	25	4	1	2	5	1	5	0	0	5	5	5			
	563,746	8011	31326	9245	33	25	5	4	2	63	3	5	3	0	5	5	5			
	411,7546	8013	31303	9247	33	25	1	2	1	56	3	1	1	9	1	5	5	A		
	411,7546	8013	31303	9247	33	25	2	2	2	51	3	1	1	3	1	5	5	A		
	411,7546	8013	31303	9247	33	25	3	2	2	22	3	1	1	12	1	5	5	A		
	411,7546	8013	31303	9247	33	25	4	2	1	22	3	1	1	19	1	5	5			
	411,7546	8013	31303	9247	33	25	5	4	2	66	1	5	0	0	5	5	5	A		
	289,6832	8020	31303	9254	33	25	1	2	1	53	1	5	1	1	5	5	5	A		
	289,6832	8020	31303	9254	33	25	2	2	2	52	1	1	1	3	5	5	5			
	289,6832	8020	31303	9254	33	25	3	1	2	17	3	5	3	0	1	5	5	A		
	316,1124	8021	31305	9255	33	25	1	2	1	35	3	1	1	12	1	5	5	A		
	316,1124	8021	31305	9255	33	25	2	2	2	41	3	1	3	0	1	5	5	A		

Figure 1. Poverty Dataset Original File

Based on the dataset obtained at BPS Batang Regency, in the form of raw data in the form of questionnaire results consisting of many parameters with a total of 830 data.

Tabel 1. Variable SUSENAS March 2022

No	Variabel Name
1.	Activity Status?
2.	Do you accept the bansos of the central government?
3.	Have you ever been denied a health check using JKN/Jamkesda?
4	During the last year have you not eaten because of a lack of money or other resources?
5	The roof material of the house?

6	Wall building materials?
7	Floor building materials?
8	What are the facilities for defecating?
9	Where is the final disposal site for feces?
10	The main source of water used by households for drinking?
11	Where is the water source used for drinking?
12	What is the main water used by the household for bathing/washing/etc?
13	How much electricity is installed?
14	Are there any household members aged 15 years and over who have received wage subsidy assistance/BSU?
15	Has your household ever been a recipient of BPNT (Non-Cash Food Assistance)/basic food assistance program?
16	Classification of output Not Poor and Poor.

Based on the table above, the number of variables selected to be used as research parameters consists of 16 variables.

2.2 Method

In this study, the process carried out in this study started with data training, data normalization, feature selection using chi-square, then continued with the process of classification using Neural Network method [10]. A general overview can be described in the flow diagram below:

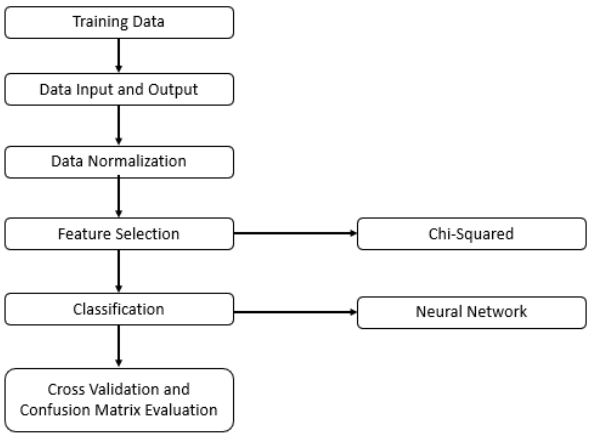


Figure 2. Proposed method

Based on the research flow carried out by the researcher, the research flow starts from data training, input and output, data normalization, feature selection using the chi-square method, classification using the Neural Network method, and finally evaluation using the confusion matrix.

2.3. Data Normalization

An input variable is a variable that will be inserted into the system to be processed and obtain the required results. In this study, the variable in question is population data that is expected to influence the results of the poor population classification [11]. Variable conversion is done so that it can be used for experiments in Neural Network optimization using feature selection chi square.

Table 2. Results of Data Normalization with Label Encoding with Ordinal Approach

No.	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**	**X11**	**X12**	**X13**	**X14**	**X15**	**Y**
1	1	5	5	5	2	3	8	2	1	3	1	3	1	5	5	2
2	1	5	5	5	2	1	6	1	1	7	1	7	2	1	5	1
3	1	5	1	5	2	1	2	1	1	7	1	7	1	5	5	2
4	1	5	5	5	2	3	2	1	1	5	1	7	1	1	5	2
5	1	5	5	5	3	1	2	1	1	5	1	3	1	5	1	2
6	1	5	5	5	2	3	2	1	1	7	1	7	2	5	5	2
7	1	5	5	5	2	1	2	1	1	7	1	7	2	5	5	1
8	1	5	5	5	2	1	2	1	1	2	1	3	1	5	5	1
9	1	5	5	5	2	1	2	1	1	7	1	7	1	5	5	1
10	1	5	5	5	2	1	2	1	1	7	1	7	1	1	5	1
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	...
830	3	5	5	5	2	3	8	1	1	7	1	7	1	5	1	2

Based on the table above, the data that has gone through a normalization process is then processed for feature selection using the chi-square feature selection method.

2.4 *Skala Likert*

In the calculation of the questionnaire in this study use calculations using the Likert Scale. Based on the results of the questionary that has been obtained then can be explained as follows. Before calculating the results of the questionnaire, it is necessary to specify intervals (distance intervals) first to know the assessment with the method of finding the values of intervals percentage score (P) [12]. Here is the calculation equation for determining intervals in the likert scale:

$$P = (\text{Total Score} / (\text{The highest Likert scale} \times \text{Number of respondents})) \times 100$$

Table 3. Answer Weight

Choice	**Information**		**Weight**
A	Very good	SB	11

B	Good	B	10
C	Pretty good	CB	9
D	In accordance	S	8
E	Suitable enough	CS	7
F	Help	M	6
G	Quite Helpful	CM	5
H	Enough	C	4
I	Quite Less	CK	3
J	Not enough	K	2
K	Very less	SK	1

Based on the table above, the answer weights are used to determine the results of the Likert scale with a total weight value from 1 to 11.

Table 4. Best Feature Results based on likert scale

No	Fitur terbaik	Nilai Fitur	Prosentase (%)
1.	**x11**	**9.071**	**99,35**
2.	**x8**	**8.861**	**97,05**
3.	**x1**	**8.840**	**96,82**
4.	**x9**	**8.821**	**96,62**
5.	x6	8.803	94,42
6.	x13	8.759	93,94
7.	x15	8.458	92,64
8.	x14	8.443	92,48
9.	x3	8.311	91,03
10.	x4	8.303	90,94
11.	x2	8.300	90,91
12.	x5	8.048	88,15
13.	x7	7.250	79,41
14.	x10	5.814	63,68
15.	x12	5.722	62,67

Based on table 2.5, it can be concluded that the four best features to be taken for neural network optimization using feature selection chi-square are starting from X11, X8, X1, and X9 with results 9.071, 8.861, 8.840, and 8.821.

2.5 Chi-Square

Chi Square used for the selection of features, will range from features with the highest rank up to features that have the higher rank to features which have the lowest rank [13]. Here's the formula for the chi square according to Liu (2014).

$$x^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

..... (1)

Table 5. Calculation of Chi Square X9

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Beton	677	81.6	81.6	81.6
	Genteng	1	.1	.1	81.7
	Seng	145	17.5	17.5	99.2
	Asbes	2	.2	.2	99.4
	Jerami	5	.6	.6	100.0
	Total	830	100.0	100.0	

Based on the table above, chi-square data is calculated for variable X9, namely What is the building material for the walls of the house?

Table 6. Parameter Neural Network

	Max. Epoch	Kinerja Tujuan	Learning Rate	Hidden Neuron
NN	1000	1e-6	0.01	[4,4,1]
NN with Chi-Square	1000	1e-6	0.01	[15,15,1]

Based on the table above, the neural network parameters are calculated with epoch 100, objective performance 1e-6, learning rate 0.01 and hidden neurons [4,4,1].

Table 7. Neural Network Optimization Results before using Chi-Square

Epoch	***Learning Rate***	***Hidden Layer***	***Akurasi***	***Presisi***	***Recall***
1000	0.01	[15, 15, 1]	93,68%	91,07%	90,26%

Based on the table above, the calculation results before using chi-square resulted in an accuracy of 93.68%, precision of 91.07% and recall of 90.26%.

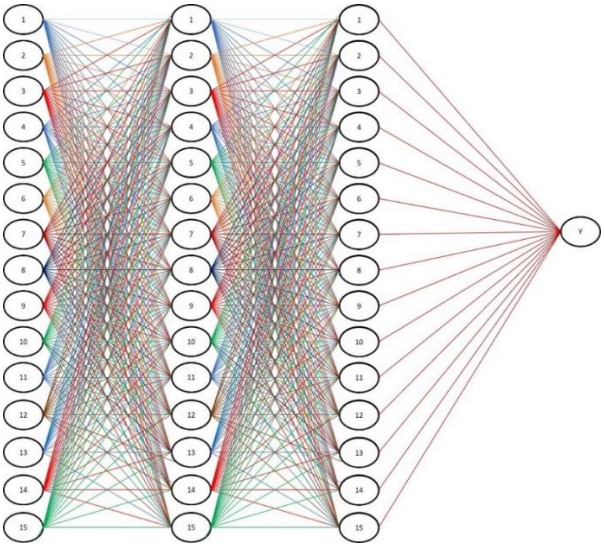


Figure 3. Neural Network Architecture 2 hidden layer [15,15,1]

Based on the image above, the results of the neural network architecture using hidden layers [15,15,1].

Table 8. Neural Network Optimization Results after using Chi-Square

Epoch	***Learning Rate***	***Hidden Layer***	***Akurasi***	***Presisi***	***Recall***
1000	0.01	[4, 4, 1]	96,38%	100%	89,38%

Based on the image above, the results of the neural network architecture using hidden layers [15,15,1].

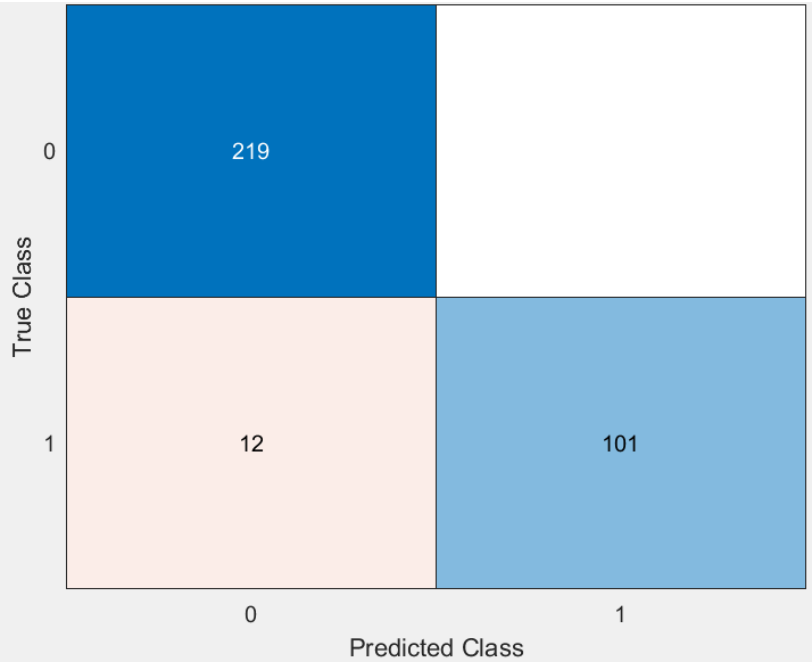


Figure 4. Confusion matrix with 2 hidden layers [4,4,1]

Based on the image above, the results of the confusion matrix using hidden layers [4,4,1].

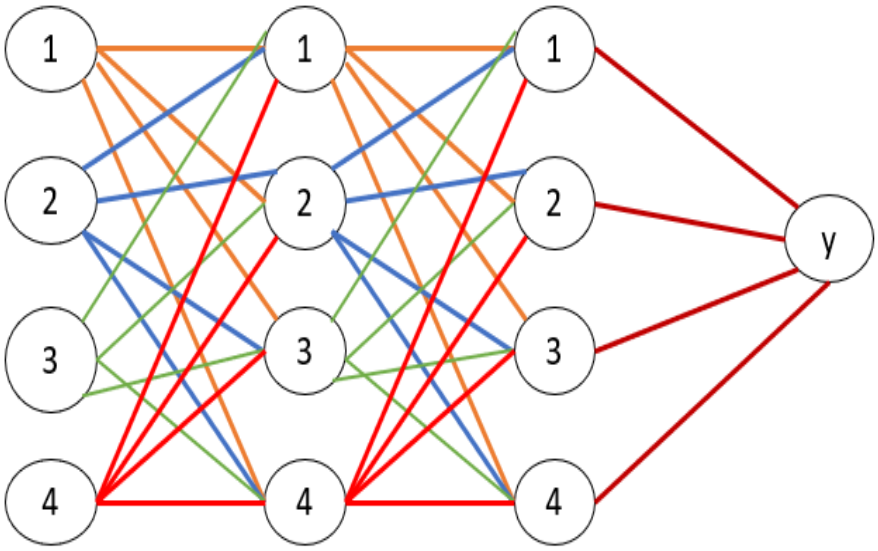


Figure 5. Neural Network Architecture 2 hidden layer [4,4,1]

Comparison graph of experimental results of neural network optimization and neural networks using chi square feature selection.

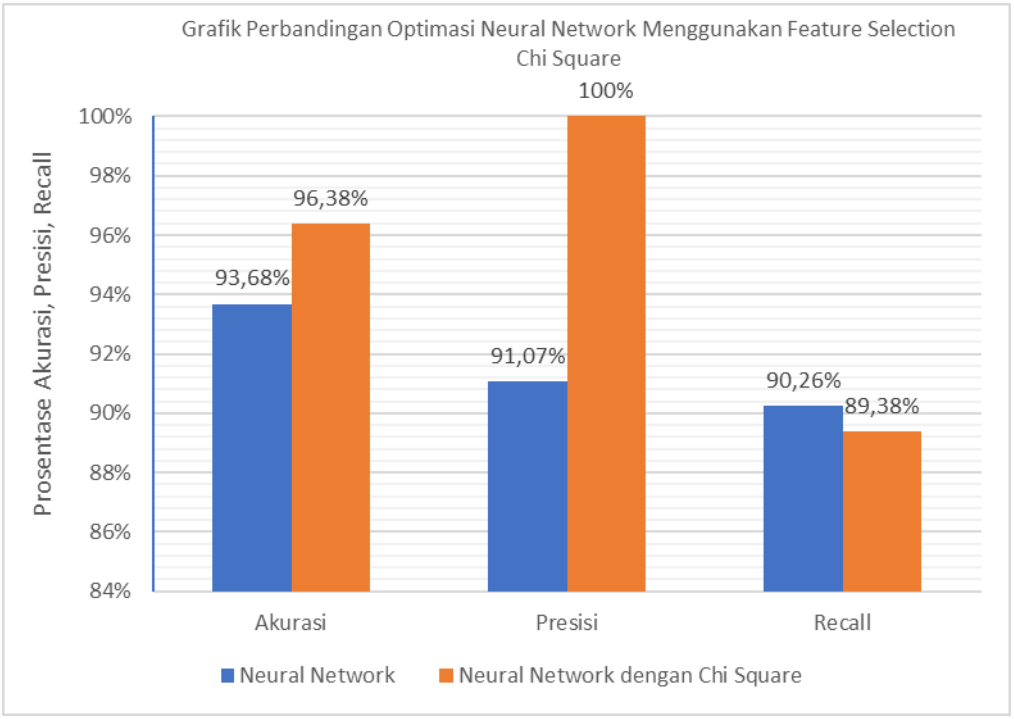


Figure 6. Graph of Neural Network Optimization Results with chi square feature selection

Thus, the use of Neural Network Optimization using feature selection chi-square has a 96.38% greater accuracy value than before using features selection chi square with a 93.68% value.

3. RESULTS AND DISCUSSION

Based on the experimental results, from the initial stage to the evaluation, it can be concluded that the poverty data classification model using the neural network method using feature selection is quite accurate. The results of the classification research using the neural network method obtained an accuracy of 93.68%, with a precision value of 91.07% and a recall value of 90.26% using an epoch of 1000 and a learning rate of 0.01 and a hidden layer [15,15,1] by using 15 features namely X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, and X15. While the results of the neural network optimization classification using the chi square feature selection get an accuracy of 96.38%, a precision value of 100% and a recall value of 89.38% using an epoch of 1000 and a learning rate of 0.01 and a hidden layer [4,4,1] by using 4 features, namely X11, X8, X1, and X9.

4. CONCLUSION

Thus the use of Neural Network Optimization before using the chi-square feature selection has a smaller value compared to the results of Neural Network optimization after the chi-square feature selection is performed. Based on the results of this study, the application of this classification can assist the Government in equitable distribution of aid with the aim of targeting poor families who need it more. As well as to determine the level of inflation that is currently happening in Batang Regency. What is needed as a development suggestion in this system is that research will be more accurate if there are more training data sets. And for future research, the classification to determine the level of poverty in Batang Regency can develop optimization of the neural network method with other methods, so that more accurate results will be obtained.

REFERENCES

[1] E. Sulistyaningrum, "PENDUGAAN AWAL KEMISKINAN RUMAH TANGGA: PENDEKATAN ARTIFICIAL NEURAL NETWORK AL BARKAH TRIANDIKA, Eny Sulistyaningrum, S.E., M.A., Ph.D," 2019.

[2] N. Su, X. An, C. Yan, and S. Ji, "Incremental attribute reduction method based on chi-square statistics and information entropy," *IEEE Access*, vol. 8, pp. 98234–98243, 2020, doi: 10.1109/ACCESS.2020.2997013.

[3] Y. D. Setiyaningrum, A. F. Herdajanti, C. Supriyanto, and Muljono, "Classification of twitter contents using chi-square and K-nearest neighbour algorithm," *Proc. - 2019 Int. Semin. Appl. Technol. Inf. Commun. Ind. 4.0 Retrosp. Prospect. Challenges, iSemantic 2019*, pp. 78–81, 2019, doi: 10.1109/ISEMANTIC.2019.8884290.

[4] Euis Saraswati, Yuyun Umaidah, and Apriade Voutama, "Penerapan Algoritma Artificial Neural Network untuk Klasifikasi Opini Publik Terhadap Covid-19," *Gener. J.*, vol. 5, no. 2, pp. 109–118, 2021, doi: 10.29407/gj.v5i2.16125.

[5] Sarwosri, D. Sunaryono, R. J. Akbar, and R. D. Setiyawan, "Poverty classification using Analytic Hierarchy Process and k-means clustering," *Proc. 2016 Int. Conf. Inf. Commun. Technol. Syst. ICTS 2016*, pp. 266–269, 2017, doi: 10.1109/ICTS.2016.7910310.

[6] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelangggan Tokopedia," *J. Gaussian*, vol. 11, no. 4, pp. 562–571, 2023, doi: 10.14710/j.gauss.11.4.562-571.

[7] D. Ispriyanti, A. Prahutama, M. Mustafid, and T. Tarno, "Klasifikasi Penerimaan Beras

Miskin Di Kota Semarang Menggunakan Algoritma Chisquare Automatic Interaction Detection (Chaid) Dan Classification and Regression Tree (Cart)," *Media Stat.*, vol. 12, no. 1, p. 63, 2019, doi: 10.14710/medstat.12.1.63-72.

[8] "Katalog/Catalog: 1102001.3325," *Batang Dalam Angka*, 2023.

[9] I. Kemiskinan and K. Batang, "Katalog : 3205014.3325," 2022.

[10] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining," *Proc. - 2015 7th Int. Conf. Inf. Technol. Electr. Eng. Envisioning Trend Comput. Inf. Eng. ICITEE 2015*, pp. 420–424, 2015, doi: 10.1109/ICITEED.2015.7408983.

[11] E. A. Kusuma, "Model Neural Network Untuk Identifikasi Variabel Kemiskinan Rumah Tangga Kecamatan Aranio," *Jutisi J. Ilm. Tek. Inform. dan …*, 2018, [Online]. Available: http://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/article/view/292%0Ahttp://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/article/viewFile/292/276

[12] J. Yao, S. Tridandapani, W. F. Auffermann, C. A. Wick, and P. T. Bhatti, "An adaptive seismocardiography (SCG)-ECG multimodal framework for cardiac gating using artificial neural networks," *IEEE J. Transl. Eng. Heal. Med.*, vol. 6, no. August, 2018, doi: 10.1109/JTEHM.2018.2869141.

[13] Z. Liu, W. Gao, Y. H. Wan, and E. Muljadi, "Wind power plant prediction by using neural networks," *2012 IEEE Energy Convers. Congr. Expo. ECCE 2012*, no. August, pp. 3154–3160, 2012, doi: 10.1109/ECCE.2012.6342351.