Comparison of Data Mining Classification Algorithm Performance for Data Prediction Type of Social Assistance Distribution

Moh. Hidayat Koniyo[1] and Made Sudarma[2]
[1]*Department of Informatics Engineering, State University of Gorontalo, Gorontalo, Indonesia*
[2]*Department of Electrical Engineering, Udayana University, Bali, Indonesia*

Keywords: Classification, Decision Tree, Neural Network, Prediction, Evaluation.

Abstract: Data on the distribution of social assistance consisting of 11 types of assistance needs to be optimized through the application of classification algorithms to predict the receipt of types of assistance. Data on aid distribution was obtained from the Department of Social Services of Gorontalo City. The data will then be used to build a classification model with the Decision Tree C4.5 algorithm and Neural Network. Furthermore, it will be evaluated using the confusion matrix method with several testing parameters. The classification model and evaluation process are carried out using WEKA 3.8.3 data mining tools. Evaluation results are then compared and analysed so that the algorithm with the best model and performance is selected based on the accuracy and classification modelling categories on the ROC (Receiver Operating Characteristic) curve, to be used in predicting new data in the form of prospective recipient social assistance data.

1 INTRODUCTION

Data distribution of social assistance in Gorontalo City Government, which consists of 11 types of assistance, needs to be optimized through the application of data mining classification algorithms to predict the receipt of types of social assistance. The data mining classification algorithm used in this research are a decision tree C4.5 and a neural network. The selection of these two algorithms is based on various research results that show the results of performance analysis with a reasonable degree of accuracy in solving several classification problems, including: an efficient and fair scholarship evaluation system can be realized (Wang et al., 2019), classification trees can be used to evaluate (Pradeep and Naveen, 2018), used to construct predictive models (Daoud and Mayo, 2019), used to predict the occurrence of lost circulation (Abbas et al., 2019), produce mood classification type labels (Sudarma and Harsemadi, 2017), used to classify Balinese script features (Sudarma and Surya, 2014).

To optimize the performance of data mining classification algorithms by applying C4.5 and neural networks it is expected to know the performance of each algorithm using the confusion matrix method with several test parameters to predict data type distribution for a certain period. In this paper the performance of the two classification algorithms will be compared, namely C4.5 and neural network using several parameters. The best results are based on accuracy and classification modeling categories on the ROC (Receiver Operating Characteristic) curve, to be used in predicting new data in the form of prospective social assistance data.

2 LITERATURE REVIEW

2.1 C4.5 Algorithm

C4.5 algorithm is a machine learning algorithm that is included in the classification and prediction methods, forming a decision tree that is useful for exploring data and finding hidden relationships, so that information or knowledge from classified datasets can be more easily identified (Breslow and Aha, 1997). To overcome the shortcomings of the decision tree algorithm (ID3) that is too sensitive to work attributes that have many values (Hssina et al., 2014). In a comparative study conducted (Hssina et al., 2014), explaining that the C4.5 algorithm acts similar to ID3 but enhances some ID3 behavior, such as the ability to use continuous data, unknown value data, using attributes with different weights, and the ability to trim trees decision made. At each tree node, C4.5 selects one data

Koniyo, M. and Sudarma, M.
Comparison of Data Mining Classification Algorithm Performance for Data Prediction Type of Social Assistance Distribution.
DOI: 10.5220/0009910003360342
In *Proceedings of the International Conferences on Information System and Technology (CONRIST 2019)*, pages 336-342
ISBN: 978-989-758-453-4

attribute that most effectively divides its sample set into a set of enriched sections in one class or another. The criterion is the acquisition of normalized information that results from the selection of attributes to separate data. The attribute with the highest normalized information acquisition was chosen to make a decision (Korting, 2006).

In the C4.5 algorithm, the gain value is used to determine which variable will be the node of a decision tree (the variable with the highest gain).

$$Gain(A) = Entropi(S) - \Sigma_{i=1}^{k} \frac{|S_i|}{|S|} xEntropi(S_i) \quad (1)$$

This process uses the parameter ”entropy” to measure the level of heterogeneity of the dataset, where the greater the value of entropy, the greater the level of heterogeneity of a data set.

$$Entropi(S) = \Sigma_{j=1}^{k} - pjlog_2pj \quad (2)$$

Information : S = dataset (case) k = number of partitions S pj = probability obtained from Sum (Yes) divided by total cases

2.2 Neural Network Algorithm

Neural Network or better known as ANN (Artificial Neural Network) is a data mining method that is widely used to do classification and prediction (McCulloch and Pitts, 1943). A Neural Network generally consists of input, output, and hidden layer. And one of the most popular algorithms used in learning of ANN is Backpropagation (McClelland et al., 1986). ANN or Artificial Neural Networks (ANN) is a parallel system consisting of many, special non-linear processors, known as neurons (Markopoulos et al., 2016). Like the human brain, they can learn from its examples, they can generalize and fault tolerance, and they can respond intelligently to new triggers. Each neuron is a primary processing unit, which receives one or more external inputs and uses it to produce an output. The whole system is considered parallel because many neurons can implement calculations simultaneously. The most important feature of neural networks is the structure of the neurons that are connected because they determine how the calculations are performed. Starting from the source layer that receives input and the output layer where the input layer is mapped, neural networks can have one or more hidden layers between. Neural networks, known as one or more hidden layers, are multilayer perceptron (MLP). These networks, unlike simple perceptron, are capable of linearly classifying inseparable patterns and can solve complex problems. Examples of ANN with a single hidden layer consisting of four units, six source units, and two output units are shown in Figs.1

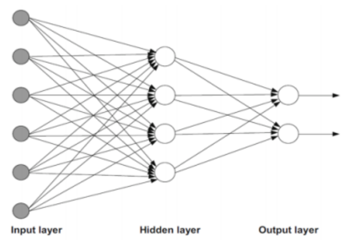


Figure 1: Single Hidden Layer Feed Forward ANN 6-4-2 (Markopoulos et al., 2016)

3 RESEARCH METHOD

3.1 Research Design

Figure 2: Research Design

Data is collected and selected from a collection of operational data, then processed to obtain data with good, complete, and consistent quality. The data that has been pre-processed is determined as a dataset which will then be used to build a classification model with the Decision Tree C.45 and Neural Network algorithm and at the same time be evaluated using the Confusion Matrix method with several test parameters. The classification model and evaluation process are carried out using WEKA 3.8.3 data mining tools. The results of the evaluation are then compared and analysed so that the algorithm with the best model is chosen based on the level of accuracy and classification modelling categories on the ROC (Receiver Operating Characteristic) Curve, to be used in making predictions of new data in the form of prospective social assistance data.

3.2 Datasets

The data used in this study were recipients of social assistance data sourced from the Department of Social Services of Gorontalo City in the database of aid distribution totalling 123 records. Each data record consists of 11 criteria with numeric and string types, namely Trans Code, KKK, Name, Address, Village, Sub-District, Education, Employment, Number of Children, Age, and Type of Assistance.

The data is then pre-processed, and 5 (five) beneficiary data criteria are selected as input attributes and 1 (one) criterion as output or label class attributes (Figure 3).

Data	Attribute	Type	Category
Input	Sub-district	String (Categorical)	Dumboraya, Dungingi, Hulonthalangi, Kota Barat, Kota Selatan, Kota Tengah, Kota Timur, Kota Utara, Sipatana
	Education	String (Categorical)	Tidak Tamat SD, SD
	Employment	String (Categorical)	Tidak Bekerja, Buruh, Petani, Asisten RT
	Number of Children	Numeric	
	Age	Numeric	
Output	Type of Assistance	String (Categorical)	Penyandang Disabilitas, Bantuan Pangan Non Tunai Pusat, Bantuan Bibit Ternak, BPJS Kesehatan (Mandiri), Penerima BNPT Daerah, BPJS Ketenagakerjaan, Bantuan Bibit Pertanian dan Pupuk, Bantuan Modal Usaha, Penerima Rasta, Program Keluarga Harapan, BPJS Kesehatan

Figure 3: Characteristics of Attribute Data.

3.3 Evaluation Measures

Evaluation of the classification results is done by the Confusion Matrix method. Evaluation of the Confusion Matrix produces accuracy, precision, and recall. Accuracy in classification is the percentage of accuracy of data records that are correctly classified instances after testing the classification results (Han et al., 2011). Precision is the proportion of positive predicted cases that are also true positive on actual data, while Recall is the proportion of positive cases that are positively predicted correctly (Powers, 2011).

Correct Classification	Classification as	
	+	-
+	True Positives (A)	False Negatives (B)
-	False Positives (C)	True Negatives (D)

Figure 4: Characteristics of Attribute Data.

$$Accuracy = (A+D)/(A+B+C+D) \tag{3}$$

The performance of classification algorithms can also be analysed through Area ROC (Receiver Operating Characteristic) and PRC (Precision-Recall Curve). The ROC curve is based on the values obtained in the Confusion Matrix calculation, which is between the False Positive Rate (FPR) and the True Positive Rate (TPR).

$$FPR = C/(C+D) \tag{4}$$

$$TPR = A/(A+B) \tag{5}$$

The PRC area is created based on values obtained from the Confusion Matrix calculation, namely Precision and Recall.

$$Precision = A/(A+C) \tag{6}$$

$$Recall = A/(A+D) \tag{7}$$

AUC (area under the curve) is calculated to measure the difference in the performance of the method used. ROC has a diagnostic value (Gorunescu, 2011).

Accuracy Value	Classification Category
0.90 – 1.00	*excellent classification*
0.80 – 0.90	*good classification*
0.70 – 0.80	*fair classification*
0.60 – 0.70	*poor classification*
0.50 – 0.60	*failure*

Figure 5: AUC Classification.

4 RESULT AND ANALYSIS

4.1 Data Model and Evaluation

The classification model and evaluation process carried out with WEKA 3.8.3 data mining tools use two algorithms, namely Decision Tree C4.5, which is implemented into J48 and Neural Network, which is implemented as Multilayer Perceptron. The process of testing the classification results using three test options available in WEKA tools, namely Cross-Validation, Percentage Split, and Use Training Set. For Cross-Validation testing techniques, the selected test parameters are the default parameters (10 folds), five-folds, and 15 folds to analysed whether there is an influence of adding and subtracting the number of folds to the accuracy value. As for the Percentage Split Testing Technique, the chosen test parameters are the defaults (66%), 45%, and 80% to analysed whether there is an influence of the distribution of the amount of training data and test data on the accuracy value. Examples of displaying the results of classification and testing of social assistance distribution datasets using the Decision Tree C4.5 (J48) and Neural Network (Multilayer Perceptron) algorithm with the Use Training Set testing model are shown in Figure 6 and Figure 7.

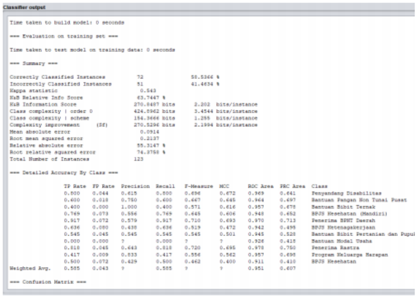


Figure 6: WEKA Display Classification Process using Decision Tree C4.5 (J48) Algorithms with the Use Training Set test option

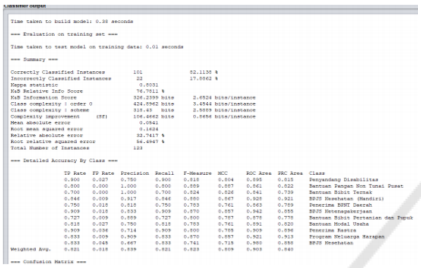


Figure 7: WEKA Display Classification Process using Neural Network Algorithms (Multilayer Perceptron) with the Use Training Set test option

4.2 Comparison and Analysis

Indicators of test results that will be used in the comparison process include accuracy (correctly classified instances), RMSE (Relative Mean Square Error), ROC Area, and PRC Area. The four indicators were tested with three techniques for testing CrossValidation, Percentage Split, and Use Training Set. The results of the comparison can be seen in Figure 8.

Test Options	Parameters	Indicators	C4.5 (J48)	Neural Network (MLP)
Cross-Validation	5-fold cross-validation	Accuracy	13.01%	9.76%
		RMSC	0.3402	0.3586
		ROC Area	0.554	0.517
		PRC Area	0.123	0.146
	10-fold cross-validation (default)	Accuracy	15.45%	10.57%
		RMSC	0.336	0.3597
		ROC Area	0.542	0.505
		PRC Area	0.132	0.129
	15-fold cross-validation	Accuracy	14.63%	10.57%
		RMSE	0.3371	0.3601
		ROC Area	0.538	0.481
		PRC Area	0.133	0.135
Percentage Split	45%	Accuracy	4.41%	10.29%
		RMSE	0.3575	0.3578
		ROC Area	0.497	0.516
		PRC Area	0.125	0.147
	66% (default)	Accuracy	9.52%	14.29%
		RMSE	0.3505	0.3483
		ROC Area	0.568	0.585
		PRC Area	0.180	0.185
	80%	Accuracy	8.00%	12.00%
		RMSE	0.3526	0.3570
		ROC Area	0.543	0.546
		PRC Area	0.265	0.260
Use Training Set		Accuracy	58.54%	**82.11%**
		RMSE	0.2137	**0.1624**
		ROC Area	**0.951**	0.903
		PRC Area	0.607	**0.840**

Figure 8: Comparison of Classification Model Test Results.

Based on the results of comparison of test data in Figure 8, it is known that the best Classification Model chosen for use in predicting data on prospective social assistance recipients is the Classification Model produced by the Neural Network Algorithm which was built through the Use Training Set testing technique, with the highest accuracy value (82.11 %), the lowest RMSE value (0.1624), and the highest PRC Area number (0.840), even though the highest ROC Area value obtained is generated by the Classification Model produced by the Decision Tree C4.5 Algorithm (0.951). But if it is measured using AUC (the area under the curve) for ROC, then the level of diagnosis produced by the ROC Area of the two Classification Models (Decision Tree C4.5 and Neural Network) are both in the category of excellent classification (0.90 - 1).

As for the results of the analysis of the parameter changes made on the Cross-Validation and Percentage Split Testing Techniques, it shows that the addition or reduction of the number of folds on the Cross-Validation will result in a decrease in the accuracy value, except for the Neural Network algorithm, increasing the number of folds results in a fixed or not influence the value of accuracy. As for the Percentage Split, the addition or reduction of the number of datasets that are divided into training data and test data results in a decrease in the accuracy value of the two algorithms. But the accuracy generated by these two testing techniques, either using default parameters or the results of testing parameter changes, results in values that are much lower than the accuracy values generated through the Use Training Set testing technique. Further analysis of the classification results using the Decision Tree Algorithm C4.5 with the Use Training Set testing technique, can be seen through the Tree visualization shown by Figure 9 and the formed Rule.

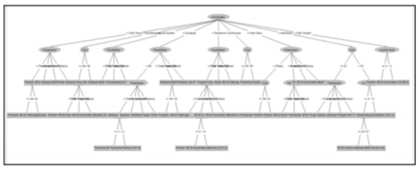


Figure 9: Tree Visualization

Based on the rule formed from the tree, it is known that the attribute that becomes the root as the main determinant in the classification process is the "Sub-district" attribute, then at the next second-level followed by the attribute "Occupation" if the beneficiary is located in the sub-district of Kota Timur and Kota Utara, the "Education" attribute if the recipient is located in the sub-district of Kota Selatan, Dungingi, and Sipatana, the "Age" attribute if the beneficiary is located in the sub-district of Hulonthalangi, Dumboraya & Kota Barat, and the attribute "Number of Children" if the beneficiary is located in sub-district of Kota Tengah. The rules formed from the results of the model classification using the Decision Tree C4.5 (J48) algorithm are as follows:

Kecamatan = Kota Timur
| Pekerjaaan = Petani
||Usia <= 44: Penyandang Disabilitas (2.0/1.0)
||Usia > 44: BPJS Ketenagakerjaan (2.0/1.0)
| Pekerjaaan=Buruh: Penerima BPNT Daerah (3.0/1.0)
| Pekerjaaan = Assiten RT: BPJS Kesehatan (Mandiri) (3.0/2.0)
| Pekerjaaan = Tidak Bekerja: Bantuan Bibit Pertanian dan Pupuk (4.0/2.0)
Kecamatan = Hulonthalangi
| Usia <= 59
||Pendidikan = SD: Penerima BPNT Daerah (5.0/2.0)
||Pendidikan = Tidak Tamat SD: BPJS Kesehatan (Mandiri) (3.0/1.0)
||Pendidikan = Tidak Sekolah: BPJS Kesehatan (Mandiri) (0.0)
| Usia > 59: Bantuan Pangan Non Tunai Pusat (4.0/1.0)
Kecamatan = Kota Selatan
|Pendidikan = SD: Penerima BPNT Daerah (5.0/3.0)
|Pendidikan = Tidak Tamat SD: Bantuan
|Pendidikan = Tidak Sekolah: Penyandang Disabilitas (2.0/1.0)
Kecamatan = Dungingi
| Pendidikan = SD
||Pekerjaaan = Petani
|||Jumlah Anak <= 3: Penerima BPNT Daerah (2.0/1.0)
|||Jumlah Anak > 3: Penerima Rastra (2.0/1.0)
||Pekerjaaan = Buruh: Bantuan Pangan Non Tunai Pusat (2.0)
||Pekerjaaan = Assiten RT: Bantuan Pangan Non Tunai Pusat (0.0)
||Pekerjaaan = Tidak Bekerja: Bantuan Pangan Non Tunai Pusat (0.0)
|Pendidikan = Tidak Tamat SD
||Usia <= 45: Penerima Rastra (4.0/2.0)
||Usia > 45: Program Keluarga Harapan (2.0)
|Pendidikan = Tidak Sekolah: Bantuan Bibit Pertanian dan Pupuk (1.0)
Kecamatan = Sipatana
| Pendidikan = SD
||Pekerjaaan = Petani: BPJS Ketenagakerjaan (3.0/1.0)
||Pekerjaaan = Buruh
|||Jumlah Anak <= 4: Penerima BPNT Daerah (2.0/1.0)
|||Jumlah Anak > 4: BPJS Kesehatan (Mandiri) (2.0/1.0)
||Pekerjaaan = Assiten RT: BPJS Kesehatan (Mandiri) (1.0)
||Pekerjaaan = Tidak Bekerja: BPJS Kesehatan (Mandiri) (0.0)
|Pendidikan = Tidak Tamat SD: Program Keluarga Harapan (3.0/1.0)
|Pendidikan = Tidak Sekolah: Bantuan Bibit Ternak (1.0)
Kecamatan = Dumboraya
| Usia <= 55: BPJS Ketenagakerjaan (11.0/7.0)
| Usia > 55: Penerima Rastra (2.0)
Kecamatan = Kota Utara
| Pekerjaaan = Petani
||Usia <= 54: Penyandang Disabilitas (4.0/1.0)
||Usia > 54: Penerima BPNT Daerah (2.0)
| Pekerjaaan = Buruh
||Pendidikan = SD: Penyandang Disabilitas (2.0/1.0)
||Pendidikan = Tidak Tamat SD: BPJS Kesehatan (Mandiri) (2.0/1.0)
||Pendidikan = Tidak Sekolah: Penyandang Disabilitas (0.0)

| Pekerjaaan = Assiten RT: Penerima Rastra (1.0)
| Pekerjaaan = Tidak Bekerja: BPJS Kesehatan (Mandiri) (3.0/2.0)
Kecamatan = Kota Barat
| Usia <= 53
|| Pekerjaaan = Petani: BPJS Kesehatan (4.0/2.0)
|| Pekerjaaan = Buruh: Program Keluarga Harapan (1.0)
|| Pekerjaaan = Assiten RT: Bantuan Pangan Non Tunai Pusat (0.0)
|| Pekerjaaan = Tidak Bekerja: Bantuan Pangan Non Tunai Pusat (2.0/1.0)
| Usia > 53
||Jumlah Anak <= 4
|||Usia <= 57: BPJS Kesehatan (Mandiri) (4.0/1.0)
|||Usia > 57: Bantuan Bibit Ternak (3.0)
||Jumlah Anak > 4: Penyandang Disabilitas (3.0/1.0)
Kecamatan = Kota Tengah
|Jumlah Anak <= 3: Penerima Rastra (5.0/2.0)
|Jumlah Anak > 3: BPJS Kesehatan (10.0/6.0)

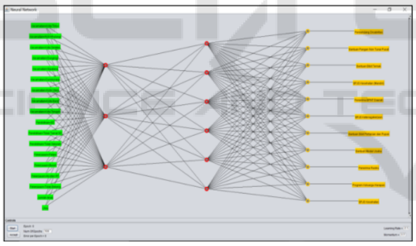
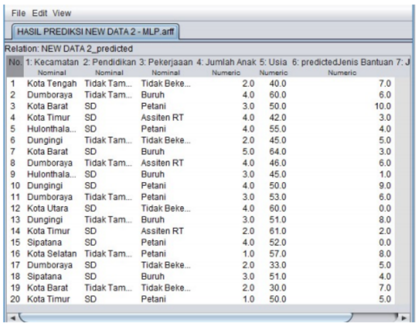


Figure 10: Neural Network Visualization

The implementation of the Neural Network algorithm in the WEKA data mining tools can also be demonstrated by the visualization output space of the Multilayer Perceptron (Figure 10). The visualization was obtained from the results of the construction of a classification model with a testing technique (*use training set*) which produced the best accuracy (82.11%) and had made changes to the default number of hidden layers parameters.

4.3 Prediction

The classification model with the best accuracy is then chosen to be used in predicting new data, namely prospective social assistance data, which in this study were tested with 20 dataset records. The classification results are displayed by the WEKA ARFF Viewer in the form of numerical data, as shown in Figure 11.



File Edit View

HASIL PREDIKSI NEW DATA 2 - MLP.arff

Relation: NEW DATA 2_predicted

No.	1: Kecamatan Nominal	2: Pendidikan Nominal	3: Pekerjaaan Nominal	4: Jumlah Anak Numeric	5: Usia Numeric	6: predictedJenis Bantuan Numeric
1	Kota Tengah	Tidak Tam...	Tidak Beke...	2.0	40.0	7.0
2	Dumboraya	Tidak Tam...	Buruh	4.0	60.0	6.0
3	Kota Barat	SD	Petani	3.0	50.0	10.0
4	Kota Timur	SD	Assiten RT	4.0	42.0	3.0
5	Hulonthala...	SD	Petani	4.0	55.0	4.0
6	Dungingi	Tidak Tam...	Tidak Beke...	2.0	45.0	5.0
7	Kota Barat	SD	Buruh	5.0	64.0	3.0
8	Dumboraya	Tidak Tam...	Assiten RT	4.0	46.0	6.0
9	Hulonthala...	SD	Buruh	3.0	45.0	1.0
10	Dungingi	SD	Petani	4.0	50.0	9.0
11	Dumboraya	Tidak Tam...	Petani	3.0	53.0	6.0
12	Kota Utara	SD	Tidak Beke...	4.0	60.0	0.0
13	Dungingi	Tidak Tam...	Buruh	3.0	51.0	8.0
14	Kota Timur	SD	Assiten RT	2.0	61.0	2.0
15	Sipatana	SD	Petani	4.0	52.0	0.0
16	Kota Selatan	Tidak Tam...	Petani	1.0	57.0	8.0
17	Dumboraya	SD	Tidak Beke...	2.0	33.0	5.0
18	Sipatana	SD	Buruh	3.0	51.0	4.0
19	Kota Barat	Tidak Tam...	Tidak Beke...	2.0	30.0	7.0
20	Kota Timur	SD	Petani	1.0	50.0	5.0

Figure 11: Prediction Results of Prospective Social Assistance Recipients

5 CONCLUSIONS

This research compares two classifier algorithms, namely C4.5 and neural networks, to classify social assistance distribution datasets. Based on the experimental results in this research it can be concluded that from the evaluation results it is known that the Neural Network Algorithm with the Use Training Set testing technique has the highest accuracy compared to the C4.5 Algorithm. Neural Network algorithm which can be used to classify beneficiary data based on the Social Assistance Distribution dataset will undoubtedly make it easier for the government as the policymaker to determine the type of assistance from prospective social assistance data as an effort to optimize the mechanism of social assistance distribution by minimizing subjectivity that can be done by authorized in the management of these activities.

The success rate of the research can be increased by adding data processed in the study and taking data from a variety of beneficiary criteria from various locations. The best algorithm in this research can be compared with other classification methods so that the most accurate algorithm is obtained.

REFERENCES

Abbas, A. K., Al-haideri, N. A., and Bashikh, A. A. (2019). Implementing artificial neural networks and support vector machines to predict lost circulation. *Egyptian Journal of Petroleum*, 28(4):339–347.

Breslow, L. A. and Aha, D. W. (1997). Simplifying decision trees: A survey. *The Knowledge Engineering Review*, 12(01):1–40.

Daoud, M. and Mayo, M. (2019). A survey of neural network-based cancer prediction models from microarray data. *Artificial intelligence in medicine*.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques*, volume 12. Springer Science & Business Media.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hssina, B., Merbouha, A., Ezzikouri, H., and Erritali, M. (2014). A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2):13–19.

Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. *Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil*.

Markopoulos, A. P., Georgiopoulos, S., and Manolakos, D. E. (2016). On the use of back propagation and radial basis function neural networks in surface roughness prediction. *Journal of Industrial Engineering International*, 12(3):389–400.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Pradeep, K. and Naveen, N. (2018). Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4. 5 and naive bayes algorithms for healthcare analytics. *Procedia computer science*, 132:412–420.

Sudarma, M. and Harsemadi, I. G. (2017). Design and analysis system of knn and id3 algorithm for music classification based on mood feature extraction. *International Journal of Electrical and Computer Engineering*, 7(1):486.

Sudarma, M. and Surya, I. W. A. (2014). The identification of balinese scripts' characters based on semantic feature and k nearest neighbor. *International Journal of Computer Applications*, 91(1).

Wang, X., Zhou, C., and Xu, X. (2019). Application of c4. 5 decision tree for scholarship evaluations. *Procedia Computer Science*, 151:179–184.