IJISCS (International Journal of Information System and Computer Science)

Full Paper
eISSN 2598-246X
pISSN 2598-0793

THE EFFECTS OF FEATURE SELECTION METHODS ON THE CLASSIFICATIONS OF IMBALANCED DATASETS

Femi Dwi Astuti[1], Indra Yatini Buryadi[2]
[1,2]Informatic, Universitas Teknologi Digital Indonesia, DIY
[1,2]Raya Janti (Majapahit) Steet No.143, Bantul, DIY

*Corresponding author:
femi@utdi.ac.id[1]
indrayatini@utdi.ac.id[2]

Article history:
Received October 12, 2022
Revised November 26, 2022
Accepted December 8, 2022

Keywords:
Imbalanced Class,
Gain Ratio,
Information Gain,
Naïve Bayes.

Abstract
imbalanced data often results in less than optimal classification. Also, datasets with a large number of attributes tends to make the classification results not too good, and in order get better classification accuracy results, one thing that could be done is to perform pre-processing to select the features to be used in the classification. This research uses information gain and gain ratio feature selection algorithms for the pre-processing stage prior to classification, and Naïve Bayes algorithm for the classification. The test is performed to determine the values of accuracy, precision, recall from the classification process without feature selection; accuracy value with information gain feature selection; accuracy value with gain ratio; and accuracy value with CBFS feature selection. The results are then compared to determine which feature selection algorithm gives the best results when applied to data with imbalanced classes. The results showed that the classification accuracy on the default of credit card client dataset using Nave Bayes algorithm was 64.27%. The information gain feature selection was able to increase the accuracy by 5.27% (from 64.27% to 69.54%), while the gain ratio feature selection was able to increase the accuracy by 14.19% (from 64.27% to 78.46%). In this case, the gain ratio is more suitable for data with greatly varied attribute values.

1.0 INTRODUCTION

Class imbalance happens when there is a significant difference between the number of classes, where the negative class is greater than the positive one[1]. This imbalance has a negative impact on the classification results when the minority class is often misclassified as the majority class because theoretically the majority classifier assumes a relatively balanced distribution [2]. Aside from class imbalance, another problem that often arises is the large number of attributes in the dataset.

The default credit card client dataset is a dataset that stores credit card client data, starting from personal data, history of past payment, delayed payments, and amount of bill statements. This dataset has a relatively large number of attributes. The number of attributs in this dataset is 23. In the default of credit card client dataset, the attribute values vary widely and are divided inti unequal classes. The large number of attributs, especially in unbalanced datasets, can affect the classification performance results[3]. Based on these problems, this study tries to apply feature selection to increace the accuracy value.

In this research, the algorithm that will be used is information gain and gain ratio. The evaluation process is carried out using k-fold cross validation to determine the effect of using feature selection before the classification process with the Naïve Bayes classification method.

The feature selection was chosen because it can overcome the problem of data imbalance in high dimensions data [4][5][6][7]. Several studies have found out that Feature selection could increase the accuracy value of the classification results [8].

Naïve Bayes is one of the classification methods that will be used in this research. Naïve Bayes will be combined with two feature selection methods. With the use of feature selection, it is expected to be able to increase the accuracy of the classification results. The results of using feature selection in classification will be compared between information gain and gain ratio.

2.0 THEORETICAL

2.1. Imbalanced Class &Feature Selection

Imbalanced class is a common problem in machine learning classification process when there is a disproportionate ratio in each class. The types of imbalanced class algorithms are:

a. Undersampling (balancing the dataset by reducing excessive class size)
b. Oversampling, (Balancing the dataset by increasing the size of the rare sample).

Feature selection is one technique most important and frequent used in pre-processing. Pre-processing is a process before the data mining process begins[9]. Main goal from selection feature is choose feature best from whole features used number of method selection feature among other :

a. Information Gain

Information Gain is defined as the effectiveness level of an attribute in classifying data. Mathematically, the information gain of attribute A is written as

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{S} Entropy(S_v)}$$

Description :
A : attribute
V : possible values for attribute A
Values(A) : The set of possible values for attribute A
$|S_v|$: number of samples for the value of v
|S| : total sample data

b. Gain Ratio

Gain ratio (GR) is a modification of the information gain that reduces its bias [9]. Information gain will face problems in handling attributes that have hugely varied values. To solve this problem, one can use another measure, i.e. the gain ratio which can be calculated based on the split information :

$$\boldsymbol{SplitInformation(S,A) = \sum_{i=1}^{c} -\frac{|S_i|}{|S|} Log_2 \frac{|S_i|}{|S|}}$$

Where S is the data sample set, and S1 to Sc are the subsets of the data sample grouped based on the number of variations in the value of attribute A. Next, the gain ratio is formulated as information gain divided by split information.

$$\boldsymbol{GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)}}$$

2.3. Naïve Bayes Classification

Classification is used to assign data objects into a limited number of classes/categories, and can be defined as a process to put data objects into one of the categories (classes) previously defined [10]. The Naïve Bayes Classifier is a classification method rooted in Bayes' theorem. The classification method proposed by British scientist Thomas Bayes that uses probability and statistical methods to predicts future values based on past experience, is known as Bayes' theorem. The main feature of Naïve Bayes Classifier is a very strong (naive) assumption of the independence of each condition/event. This algorithm assumes that object attributes are independent. The probabilities involved in producing final estimations are calculated as the sum of the frequencies from the "master" decision table. Naive Bayes Classifier works very well compared to other classifier models [11]. They reported that "Naïve Bayes Classifier gives better accuracy rate than other classifier models". The advantage of this method is that it only requires a small amount of training data to determine the parameter estimates needed in the classification process. Since it is assumed to be an independent variable, only the variance of a variable in a class is needed to determine the classification, not the entire covariance matrix.

3.0 METHODOLOGY

The research flow to compare various feature selection methods in the classification process of datasets that have imbalanced class can be seen in figure 1.

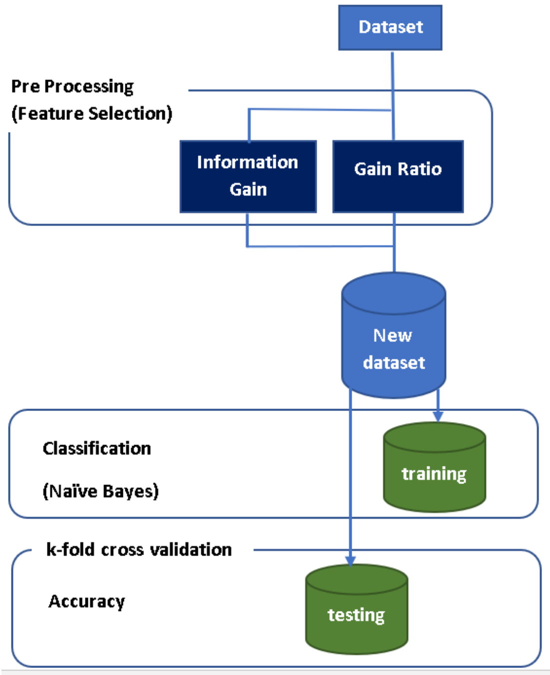

Figure 1. Research Flow

Figure 1 shows the flow of the research, from the collection of the dataset, to the accuracy of the results. The first stage in this research is the process of collecting data to be used as dataset, which is the default of credit card client data taken from UCI machine learning. Having determined the dataset, the pre-processing stage is then carried out for feature selection. The feature selections to be compared are information gain and gain ratio. Having established feature selection, a new dataset will emerge that will then be used for the classification process. The classification was carried out three times, the first was for the original dataset without being subjected to feature selection. The second classification is for dataset from the pre-processing using information gain feature selection, and the third classification is for the dataset from pre-processing using gain ratio feature selection. All three classification processes are carried out using Naïve Bayes method.

Research evaluation/testing was performed by calculating the accuracy value using 10-fold cross validation. The achievement indicator in this research shows different accuracy results between classification with Naïve Bayes only, Naive Bayes accuracy with information gain, and Naïve Bayes with gain ratio.

4.0 RESULTS

The data used in this research is public data from the UCI machine learning repository, i.e., the default of credit card clients. This dataset has 30,000 data records with as many as 23 attributes:

X1 : amount of the given credit (NT dollar)
X2 : Gender, (1: male, 2 : female)
X3 : Education, (1 : graduate school, 2 : university, 3 : high school and 4 : others)
X4 : Marital Status , (1 : married, 2 unmarried dan 3 : other)
X5 : Age
X6 : History of past payment for september 2005
X7 : History of past payment for august 2005
X8 : History of past payment for july 2005
X9 : History of past payment for june 2005
X10 : History of past payment for may 2005
X11 : History of past payment for april 2005

For attribute values X6 to X11, the possible values are -1, 1,2,3,4,5,6,7,8, and 9.

-1 : pay on time
1 : delay payment for one month
2 : delay payment for two month
3 : delay payment for three month
4 : delay payment for four month
5 : delay payment for five month
6 : delay payment for six month
7 : delay payment for seven month
8 : delay payment for eight month
9 : delay payment for nine month

X12 : Amount of bill statement for september 2005
X13 : Amount of bill statement for august 2005
X14 : Amount of bill statement for july 2005
X15 : Amount of bill statement for june 2005
X16 : Amount of bill statement for may 2005
X17 : Amount of bill statement for april 2005
X18 : Amount of previous payment for september 2005
X19 : Amount of previous payment for august 2005
X20 : Amount of previous payment for july 2005
X21 : Amount of previous payment for june 2005
X22 : Amount of previous payment for may 2005
X23 : Amount of previous payment for april 2005

The distribution of data classes from the default data of credit card clients includes:
a. 0 as 6.636 (78%)
b. 1 as 23.364 (22%)

The 0 in the dataset class means the payment default is 'no' and 1 is 'yes'. From the class division, it is clear that the dataset is imbalanced because the majority contain 0 (as high as 78%) which is very high compared with the value of 1 which is only 22%.

Examples of data used in this research can be seen in figure 2.

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0
7	500000	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0
8	100000	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0
11	200000	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0
16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0
17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1
18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0
19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0
21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0
22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1
23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1
24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1
25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0
26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0
27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1
28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0
29	50000	2	3	1	47	-1	-1	-1	-1	-1	-1	650	3415	3416	2040	30430	257	3415	3421	2044	30430	257	0	0
30	50000	1	1	2	26	0	0	0	0	0	0	15329	16575	17496	17907	18375	11400	1500	1500	1000	1000	1600	0	0

Figure 2. Sample data of credit card clients

The study was conducted using rapid miner machine learning tools to see the accuracy of the use of feature selection in unbalanced datasets using naïve bayes classification.

4.1 Testing

The first test was performed for the classification of the default dataset of credit card clients using the Naïve Bayes classification method. The classification is done without feature selection and the results were then tested. The test is performed using k-fold cross validation to determine the accuracy value of the classification results. The accuracy of the classification results is shown in Figure 3.

accuracy: 64.27% +/- 6.11% (micro average: 64.27%)

	true 1	true 0	class precision
pred. 1	1505	2819	34.81%
pred. 0	754	4922	86.72%
class recall	66.62%	63.58%	

Figure 3. Classification accuracy of naïve bayes

Figure 3 shows that the accuracy value is 64.27%, which is considered not too high, so other methods are needed to increase the accuracy value, one of which is by performing feature selection as one of the pre-processing methods on the dataset before performing the Naïve Bayes classification process.

The next test utilizes one of the feature selection methods, i.e., information gain. The dataset used is still the same as the previous one. The weights of the attributes resulted from feature selection process using the information gain method is shown in table 1.

Table 1. Weighting attribute Information Gain

Attribute	Weight
X1	0,010
X2	0,001
X3	0,003
X4	0,001
X5	0,002
X6	0,076
X7	0,060
X8	0,050
X9	0,042
X10	0,040
X11	0,032
X12	0,001
X13	0,001
X14	0,001
X15	0,001
X16	0,001
X17	0,000
X18	0,013
X19	0,013
X20	0,011
X21	0,010
X22	0,006
X23	0,008

From table 1, it is clear that of the 23 attributes used, the X6 attribute has the biggest value of 0.076, while the smallest of all attributes is the X17 with a value of 0.000. Some attributes appear to have relatively small values so that they do not have much effect on the classification process. In this research, the top 12 attributes with biggest values were selected, and based on the results of information gain, the 12 attributes are: X6, X7, X8, X9, X10, X11, X18, X19, X20, X1, X21, and X23.

Having selected the top 12 attributes, those attributes were then classified using Naïve Bayes classification method. The results of the classification using the new dataset can increase the accuracy of the classification results. The accuracy before using feature selection is 64.27% while classification accuracy using information gain is 69.54%. So it is clear that information gain can increase accuracy by 5.27%. The results of accuracy testing in rapid miners are shown in Figure 4.

accuracy: 69.54% +/- 5.88% (micro average: 69.54%)

	true 1	true 0	class precision
pred. 1	1391	2178	38.97%
pred. 0	868	5563	86.50%
class recall	61.58%	71.86%	

Figure 4. Accuraty of naïve bayes with information gain

The next test is performed by using another feature selection method, i.e. the gain ratio. The dataset used is still the same as the dataset for the previous test. The values obtained through the feature selection process using the gain ratio method are shown in table 2.

Table 2. Weighting attribute Gain Ratio

Attribute	Weight
X1	0,028
X2	0,001
X3	0,002
X4	0,001
X5	0,048
X6	0,153
X7	0,101
X8	0,146
X9	0,086
X10	0,146
X11	0,156
X12	0,032
X13	0,045
X14	0,028
X15	0,029
X16	0,030
X17	0,146
X18	0,028
X19	0,028
X20	0,028
X21	0,032
X22	0,025
X23	0,028

From table 2, it is clear that the largest value of the 23 attributes attribute X11 (History of past payment in April 2005) with a value of 0.156. While the smallest value of all attributes is X2 (gender) and X4 (marital status) with a value of 0.001. Some attributes seem to have relatively small values so as to have much effect on the classification process. In this research, the top 12 attributes with largest values were selected, and based on the results of the gain ration, the 12 attributes are X11, X6, X17, X10, X9, X7, X9, X5, X13, X21, X12, and X16.

Having selected the top 12 attributes, those attributes were then classified using Naïve Bayes classification method. The results of classification using the new dataset are shown to increase the accuracy of the classification results. The accuracy before using feature selection is 64.27% while the classification accuracy using gain ratio is 78.46%, so the gain ratio gives an increased accuracy of 14.19%. This value is much higher when compared with the results of information gain feature selection (with an accuracy of 5.27%). This could happen because in theory, Information gain tends to have problems with attributes with greatly varied values. And the attributes of the dataset used in this research vary greatly so that the increase in accuracy is not too significant. The results of accuracy testing in rapid miners are shown in Figure 5.

accuracy: 78.46% +/- 1.31% (micro average: 78.46%)

	true 1	true 0	class precision
pred. 1	935	830	52.97%
pred. 0	1324	6911	83.92%
class recall	41.39%	89.28%	

Figure 5. Accuraty of naïve bayes with gain ratio

Based on the results of the study, it can be seen a comparison of the accuracy of various tests as shown in Table 3 while the accuracy comparison chart can be seen in Figure 6.

Table 3 comparison of accuracy results

Method	accuracy
Naïve Bayes	64,27%
Naïve bayes + Information Gain	69,54%
Naïve Bayes + Gain Ratio	78,46%

Based on figure 6, it can be seen that the highest accuracy is when the classification is carried out by utilizing the feature selection gain ratio method. The lowest accuracy is when the classification is carried out without the use of feature selection.

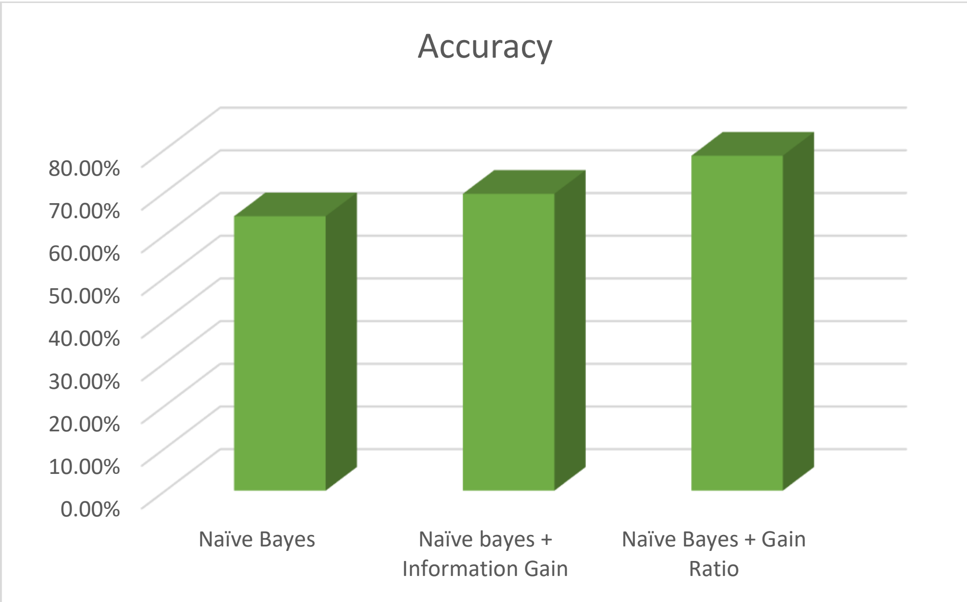

Figure 6. Accuracy Comparison Chart

The presentation of differences in classification accuracy results can be clearer when viewed from the graph. Through figure 6, it can be easily seen the difference in accuracy results between the use of the naïve bayes method alone, the naïve bayes method which is combined with the feature selection of information gain and the naïve bayes method combined with the feature selection of gain ratio. Based on figure 6, it can be seen that the greatest accuracy is the use of the naïve bayes method combined with the selection of the gain ratio feature.

5.0 CONCLUSION

Based on the discussion that has been described, it can be concluded that feature selection is one way that can be used to increase the accuracy value of the classification results. Feature selection is done by selecting the best features in the dataset. After going through several tests using feature selection, it can be said that the best feature selection for attributes whose values vary is the gain ratio. Information gain is only able to slightly increase the value of accuracy because information gain is not suitable for datasets that have varying attribute values. The results of classification accuracy on the default of credit card client dataset using Naïve Bayes are 64.27%. The information gain feature selection can increase accuracy by 5.27% (from 64.27% to 69.54%). The gain ratio feature selection can increase accuracy by 14.19% (from 64.27% to 78.46%). Gain ratio is more suitable for data whose attribute values vary widely

REFERENCES

[1] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2015.

[2] S. Zhang, S. Sadaoui, and M. Mouhoub, "An Empirical Analysis of Imbalanced Data

Classification," *Comput. Inf. Sci.*, vol. 8, no. 1, pp. 151–162, 2015, doi: 10.5539/cis.v8n1p151.

[3] A. Purnomo, M. A. Barata, M. A. Soeleman, and F. Alzami, "Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease," *J. Phys. Conf. Ser.*, vol. 1511, no. 1, 2020, doi: 10.1088/1742-6596/1511/1/012001.

[4] D. Mladenić and M. Grobelnik, "Feature selection for unbalanced class distribution and Naive Bayes," *Proc. Sixt. Int. Conf. Mach. Learn.*, no. January, pp. 258–267, 1999, doi: 10.1214/aoms/1177705148.

[5] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, no. March 2003, pp. 1289–1305, 2003.

[6] Y. Hu *et al.*, "An Improved Algorithm for Imbalanced Data and Small Sample Size Classification," *J. Data Anal. Inf. Process.*, vol. 03, no. 03, pp. 27–33, 2015, doi: 10.4236/jdaip.2015.33004.

[7] D. Tiwari, "Handling Class Imbalance Problem Using Feature Selection," *Int. J. Adv. Res. Comput. Sci. Technol.*, vol. 2, no. 2, pp. 516–520, 2014.

[8] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, 2018, doi: 10.1155/2018/1407817.

[9] P. P. R., V. M.L., and S. S., "Gain Ratio Based Feature Selection Method for Privacy Preservation," *ICTACT J. Soft Comput.*, vol. 01, no. 04, pp. 201–205, 2011, doi: 10.21917/ijsc.2011.0031.

[10] I. Pratama, P. P.-I. JOURNALS, and undefined 2020, "Multiclass Classification with Imbalanced Class and Missing Data," *Ijconsist.Org*, no. September, pp. 1–6, 2020, [Online]. Available: https://ijconsist.org/index.php/ijconsist/article/view/25.

[11] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.