IJIBEC
International Journal of Islamic Business and Economics

International Journal of Islamic Business and Economics
Available at http://e-journal.iainpekalongan.ac.id/index.php/IJIBEC/

Data Attribute Selection with Information Gain to Improve Credit Approval Classification Performance using K-Nearest Neighbor Algorithm

Ivandari, Tria Titiani Chasanah, Sattriedi Wahyu Binabar, & M. Adib Al Karomi

STMIK Widya Pratama Pekalongan
E-mail: Ivandarialkaromi@gmail.com

Article Info

Article history:

Received	: 7 May 2017
Accepted	: 17 May 2017
Published	: 1 June 2017

Abstract

Credit is one of the modern economic behaviors. In practice, credit can be either borrowing a certain amount of money or purchasing goods with a gradual payment process and within an agreed timeframe. Economic conditions that are less supportive and high community needs make people choose to buy goods with this credit process. Unfortunately the high needs sometimes are not in line with the ability to make payments in accordance with the initial agreement. Such condition causes the payment process to be disrupted or also called the term “bad credit”. This research uses public data of credit card dataset from UCI repository and private data that is dataset of credit approval from local banking. The *information gain algorithm* is used to calculate the weights of each of the attributes. From the calculation results note that all attributes have different weights. This study resulted in the conclusion that not all data attributes influence the classification result. Suppose attribute A1 to UCI dataset as well as loan type attribute on local dataset that has information gain weight 0 (zero). The result of classification using K-Nearest Neighbors algorithm shows that there is an increase of 7.53% for UCI dataset and 3.26% for local dataset after feature selection on both datasets.

Keywords:

KNN Accuracy, Feature selection, Credit Approval

1. Introduction

Credit is the economic behavior of modern society that makes a purchase or borrows some money by returning the loan on a regular basis. One of the triggers of society deciding to do credit is the overwhelming desire and unable to resist the desire. The bad loans often occur due to the many needs without balanced with the source of funds owned. The high number of bad loans makes the financial services and banking sector should be more selective in choosing prospective customers. Data 776 customers taken from financial services companies in Indonesia mentioned that 566 of them have problems in payment or credit repayment, while customers who make payment in accordance with the agreement are only as many as 210 people or about 27% of total customers.

The dataset credit approval taken from UCI repository is the data of credit card owners. These data attributes are deliberately kept secret to keep the names of customers and agencies. This dataset is widely used by researchers to test the performance of classification algorithms. This credit approval dataset is a public data that can be downloaded on the page: https://archive.ics.uci.edu/ml/datasets/Credit+Approval. The number of records of this dataset is 690 with 16 attributes one of which is a label attribute. The label attributes say there are 307 customers or 44.5% make credit payments smoothly, while 383 or 55.5% others have bad credit.

The great number of bad credit customers will make the company suffer from financial loss. Preventing bad debts can be done by selecting customers more objectively before the credit approval is given (Maulana & Al Karomi, 2016). Customer selection can be done by taking into account the data from the previous period as the decision making guidance for the next prospective customers. Data mining can analyze old cases to find patterns from data by using pattern recognition techniques such as statistics and mathematics (Larose, 2005). Large data sets can be meaningless if the information or knowledge inside is less or cannot be retrieved. Data mining answers this problem by analyzing large data and then creating a rule, pattern, or a particular model to recognize new data that is not within the rows of stored data (Prasetyo, 2012). Data Mining or often also called *Knowledge Discovery in Database* (KDD) is a field of science that often discusses the pattern of a set of data. A series of processes to gain knowledge or patterns from the data set is called data mining (Ian H Witten Eibe Frank Mark A Hall, 2011).

Data mining is divided into *Supervised Learning* and *Unsupervised Learning* based on learning method (Santosa, 2007). *Supervised Learning* uses data from the past or the training data in the process of calculation, while *Unsupervised Learning* does not. One of the functions of data mining is classification. Classification is part of the supervised learning with one goal attribute or label. One of the best and widely used classification algorithms is K-Nearest Neighbors (Wu et al., 2007). The performance of an algorithm can be affected by the data set and data type used (Amancio et al., 2013). Some algorithm models are powerful for certain data types (Ragab, Noaman, Al-Ghamdi, & Madbouly, 2014) (Patel, Vala, & Pandya, 2014) (Ashari, Paryudi, & Tjoa, 2013). Data attributes greatly affect the performance of the algorithm. The presence of irrelevant attributes can decrease the performance of the algorithm as well as affect the accuracy of the algorithm (Han & Kamber, 2006). The more relevant attributes used in the classification process will improve the accuracy of an algorithm (Maimoon, 2010) (Alpaydin, 2010). The number of irrelevant attributes can degrade the performance of the classification algorithm (Karegowda, Manjunath, & Jayaram, 2010).

Attempts to sort the attributes according to their interests are mostly done to make the dataset more mature. This effort is known as feature selection. This stage is one of the pre-processing classification steps by eliminating the irrelevant attributes in the dataset. This stage will reduce the irrelevant attributes in order to improve the algorithm accuracy. One of the popular and widely used feature selection algorithms is *information gain* (Alkaromi, 2014) (Azhagusundari

& Thanamani, 2013). Good *information gain* is used for the selection of attributes especially in handling high-dimensional data (Koprinska, 2010). The number of attributes that are not relevant in addition to affecting the accuracy level can also hinder the computation process.

This research applies *information gain* algorithm to select the credit approval dataset feature to improve the performance of KNN classification algorithm.

2. Research Methods

This research uses experimental research method. Figure 1 shows the framework of the study. In this research two different datasets are used to know the effect of the application of *information gain* to the result of KNN algorithm classification.

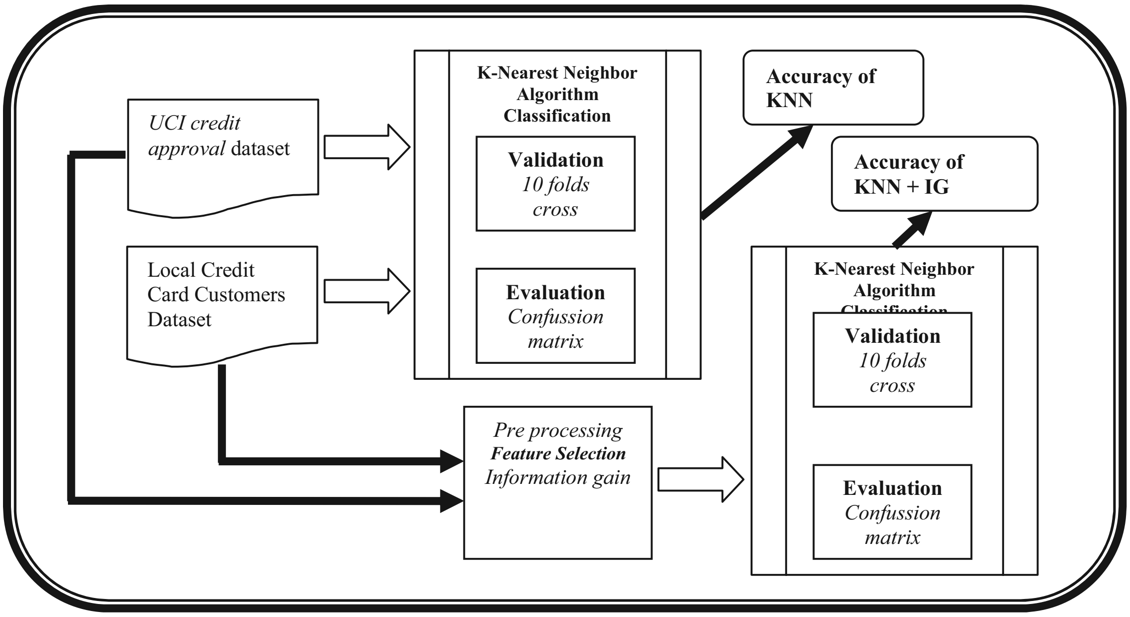


Figure 1. Research Framework

2.1 Data Collection Method

Data collection is the first stage of the research. The data used in this research is the credit approval dataset with 14 data attributes and 766 records. The label attribute on this data is the credit status with the stuck and smoother variant. Table 1 presents the metadata of the credit approval dataset.

Table 1. Metadata of credit card customers

Role	Attribute Name	Type	Statistics	Range	Missing
label	credit status	binominal	mode = BAD (556), least = GOOD (210)	BAD (556), GOOD (210)	0
regular	sex	binominal	mode = P (462), least = L (304)	P (462), L (304)	0
regular	age	integer	avg = 29.161 +/- 263.166	[-7162.000 ; 1043.000]	1
regular	amount of loan	numeric	avg = 2712482.631 +/- 9995602.067	[83333.330 ; 228655000.000]	0
regular	jkw	integer	avg = 18.961 +/- 32.076	[1.000 ; 679.000]	0
regular	amount of installment per month	numeric	avg = 233391.702 +/- 548968.221	[0.000 ; 10350000.000]	0

regular	type of loan	polynominal	mode = 100 (766), least = 100 (766)	100 (766)	0
regular	type of loan	polynominal	mode = 301 (720), least = 302 (5)	301 (720), 302 (5), 303 (6), 304 (6), 305 (29)	0
regular	bi economics sectors	integer	avg = 6013.046 +/- 216.196	[6000.000 ; 9990.000]	1
regular	col	polynominal	mode = 1 (600), least = 2 (166)	1 (600), 2 (166)	0
regular	bi debtor class	polynominal	mode = 874 (757), least = 834 (1)	874 (757), 876 (8), 834 (1)	0
regular	Biguarantor class	polynominal	mode = 000 (519), least = 835 (1)	875 (229), 000 (519), 800 (8), 874 (9), 835 (1)	0
regular	nominative balance	numeric	avg = 2007385.712 +/- 8711282.360	[-4000000.000 ; 209404092.000]	0
regular	principal arrears	numeric	avg = 790085.298 +/- 4139216.644	[0.000 ; 91612122.240]	0
regular	interest arrears	numeric	avg = 87717.084 +/- 568231.776	[0.000 ; 11000000.000]	0

Source : Bank credit card customers in Indonesia

Note : Processed with Rapid Miner

This study also uses public datasets obtained from UCI repository. Table 2 shows the metadata of the credit approval data set.

Tabel 2. Metadata of ***credit approval***

Role	Atribute Name	Type	Statistics	Range	Missing
label	L	binominal	mode = - (383), least = + (307)	+ (307), - (383)	0
regular	a1	binominal	mode = b (468), least = a (210)	b (468), a (210)	12
regular	a2	polynominal	mode = ? (12), least = 30.83 (1)	? (12), 22.67 (9), 20.42 (7), 18.83 (6), 19.17 (6), 20.67 (6), 22.5 (6), 23.58 (6), 24.5 (6), 25.0 (6), 23.0 (5), 23.08 (5), 23.25 (5), 27.67 (5), 27.83 (5), 33.17 (5), 20.0 (4), 20.75 (4), 22.08 (4), 22.92 (4), 23.5 (4), 24.58 (4), 24.75 (4), 25.17 (4), 25.67 (4), 26.17 (4), ... and 300 more ... , 57.42 (1), 57.58 (1), 57.83 (1), 58.33 (1), 58.42 (1), 58.58 (1), 58.67 (1), 59.5 (1), 59.67 (1), 60.08 (1), 60.58 (1), 60.92 (1), 62.5 (1), 62.75 (1), 63.33 (1), 65.17 (1), 65.42 (1), 67.75 (1), 68.67 (1), 69.17 (1), 69.5 (1), 71.58 (1), 73.42 (1), 74.83 (1), 76.75 (1), 80.25 (1)	0
regular	a3	numeric	avg = 4.759 +/- 4.978	[0.000 ; 28.000]	0
regular	a4	binominal	mode = u (519), least = y (163)	u (519), y (163)	8
regular	a5	binominal	mode = g (519), least = p (163)	g (519), p (163)	8

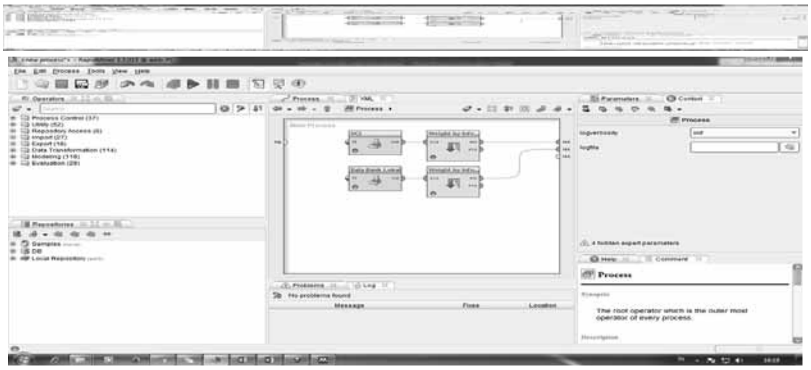
regular	a6	polynominal	mode = c (137), least = r (3)	w (64), q (78), m (38), r (3), cc (41), k (51), c (137), d (30), x (38), i (59), e (25), aa (54), ff (53), j (10), ? (9)	0
regular	a7	polynominal	mode = v (399), least = o (2)	v (399), h (138), bb (59), ff (57), j (8), z (8), ? (9), o (2), dd (6), n (4)	0
regular	a8	numeric	avg = 2.223 +/- 3.347	[0.000 ; 28.500]	0
regular	a9	binominal	mode = t (361), least = f (329)	t (361), f (329)	0
regular	a10	binominal	mode = f (395), least = t (295)	t (295), f (395)	0
regular	a11	integer	avg = 2.400 +/- 4.863	[0.000 ; 67.000]	0
regular	a12	binominal	mode = f (374), least = t (316)	f (374), t (316)	0
regular	a13	binominal	mode = g (625), least = s (57)	g (625), s (57)	8
regular	a14	polynominal	mode = 0.0 (132), least = 202.0 (1)	0.0 (132), 120.0 (35), 200.0 (35), 160.0 (34), 100.0 (30), 80.0 (30), 280.0 (22), 180.0 (18), 140.0 (16), 240.0 (14), 320.0 (14), 300.0 (13), ? (13), 260.0 (11), 220.0 (9), 400.0 (9), 60.0 (9), 340.0 (7), 360.0 (7), 380.0 (5), 108.0 (4), 132.0 (4), 144.0 (4), 232.0 (4), 40.0 (4), 420.0 (4), ... and 121 more ... , 487.0 (1), 49.0 (1), 491.0 (1), 510.0 (1), 515.0 (1), 519.0 (1), 52.0 (1), 523.0 (1), 550.0 (1), 56.0 (1), 583.0 (1), 600.0 (1), 62.0 (1), 640.0 (1), 680.0 (1), 711.0 (1), 75.0 (1), 76.0 (1), 760.0 (1), 840.0 (1), 86.0 (1), 928.0 (1), 93.0 (1), 94.0 (1), 980.0 (1), 99.0 (1)	0
regular	a15	integer	avg = 1017.386 +/- 5210.103	[0.000 ; 100000.000]	0

Source : https://archive.ics.uci.edu/ml/datasets/Credit+Approval

Note : Processed with RapidMiner

2.2 Feature selection

This feature selection stage was done after knowing the importance level of all dataset attributes. Software Rapid Miner is used to perform calculations. Figure 2 presents a rapid miner worksheet with two datasets for the information gain feature selection process.



No problems found

The root operator which is the outer most operator of every process.

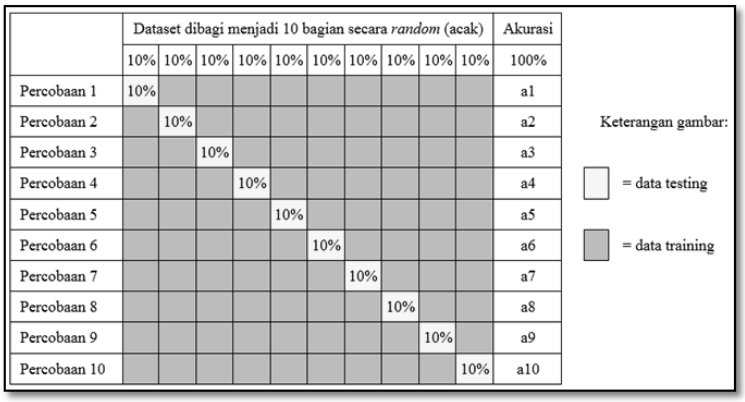
Figure 2. Worksheet of information gain selection feature

2.3 Classification of K-Nearest Neighbors

The process of classification is done by using software Rapid Miner which includes validation and evaluation. The validation is done using cross validation, while the evaluation is done by using *confusion matrix.*

2.3.1 Cross Validation

Cross validation is one way of validating the classification dataset calculations that are widely used by researchers around the world. The process done in the actual validation is to divide the entire data so that each record gets the same portion as training data or data testing. The most recommended data sharing is 10 or often called 10 folds cross validation. 10 folds cross validation is dividing the dataset into 10 parts randomly then using 90% data as data testing and 10% other as training data. This process is repeated up to 10 times until all records get part as data testing. Figure 3 represents the 10 folds cross validation.



	Dataset dibagi menjadi 10 bagian secara *random* (acak)										Akurasi
	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%	100%
Percobaan 1	10%	▨	▨	▨	▨	▨	▨	▨	▨	▨	a1
Percobaan 2	▨	10%	▨	▨	▨	▨	▨	▨	▨	▨	a2
Percobaan 3	▨	▨	10%	▨	▨	▨	▨	▨	▨	▨	a3
Percobaan 4	▨	▨	▨	10%	▨	▨	▨	▨	▨	▨	a4
Percobaan 5	▨	▨	▨	▨	10%	▨	▨	▨	▨	▨	a5
Percobaan 6	▨	▨	▨	▨	▨	10%	▨	▨	▨	▨	a6
Percobaan 7	▨	▨	▨	▨	▨	▨	10%	▨	▨	▨	a7
Percobaan 8	▨	▨	▨	▨	▨	▨	▨	10%	▨	▨	a8
Percobaan 9	▨	▨	▨	▨	▨	▨	▨	▨	10%	▨	a9
Percobaan 10	▨	▨	▨	▨	▨	▨	▨	▨	▨	10%	a10

Keterangan gambar:

□ = data testing

▨ = data training

Figure 3. Representation of 10 folds cross validation

2.3.2 Confusion Matrix

Confusion matrix is one evaluation method of classification algorithm. This matrix compares the dataset of the classification according to the actual dataset by the total number of records of the existing dataset. If the overall classification dataset is the same as the actual data, then the accuracy of the classification algorithm is 100%. Figure 4 is a process of calculating the performance of the algorithm using confusion matrix in software Rapid Miner.

Figure 4. Model of ConfusionMatrix in Rapid Miner application

The calculation of confusion matrix manually can be seen in table 3 below.

Table 3. Confusion Matrix

Classification		*Predicted class*	
		Class: YES	*Class*: NO
Observed class	*Class*YES	***a*** ***True Positive (TP)***	***b*** ***False Negative (FN)***
	*Class*NO	***c*** ***False Positive (FP)***	***d*** ***True Negative (TN)***

Source: (Gorunescu, 2011)

Note: for labels with 2 variants (yes and no)

From table 3. the level of accuracy of an algorithm model can be calculated using the equation as follows:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Notes:

A: the positive classification result with the positive actual class

B: the result of negative classification with the positive actual class

C: Positive classification result with the negative actual class

D: the result of negative classification with negative actual class

3. Results

3.1 The result of calculation of information gain

The result of information algorithm feature selection using Rapid Miner appears when the play button has been pressed. Table 4 is the overall weight of the attributes of the UCI credit approval dataset, while table 5 is the overall weight of the attributes of the credit approval dataset.

Table 4. Weight of UCI credit approval dataset attributes

Attribute	Weight
a1	0.0
a12	5.314335658427709E-4
a13	0.01608464732626536
a4	0.04780338720150725
a5	0.04780338720150725
a3	0.07606674864736256
a7	0.0931005103029748
a6	0.2033924281879285
a8	0.20499991966638934
a15	0.20539914171876603
a10	0.29154532541882355
a11	0.3609740417259771
a14	0.545581315418313
a9	0.7955564024869302
a2	1.0

Source : https://archive.ics.uci.edu/ml/datasets/Credit+Approval

Note : processes with Rapid Miner

Table 5. Weights of customers'credit card dataset attributes

Attribute	Weight
type of loan	0.0
bi_debtor_class	0.003674972187509268
bi_economics_sector	0.005556000889587041
sex	0.006676784556444653
type of loan	0.012385293071103486
age	0.02390718755420234
jkw	0.025697434452412796
nominative_balance	0.054095963366998776

amount_of installment per month	0.14014094129885607
amount of loan	0.20437474672919517
col	0.22414291584752508
bi_guarantor class	0.2396212592889049
interest arrears	0.2842707596587644
principle arrears	1.0

Source : Bank credit card customers in Indonesia

Note : Processed with Rapid Miner

From the result of feature selection we have known the weight of each attribute of both datasets. The importance of all these attributes will be used as a reference in the next process of classification using the KNN algorithm.

3.2 Performance of K-Nearest Neighbors

The results of the KNN classification algorithm using the UCI credit approval dataset and credit card customers are presented in the following tables. Table 6 presents the accuracy of the KNN algorithm with the overall attributes used for the credit approval dataset, while table 7 presents the KNN accuracy for the dataset of credit card customers with all attributes used.

Table 6. The accuracy of KNN for UCI credit approval dataset

K	1	3	5	7	9	11	13	15	17
ACC	68,12	72,75	72,61	74,2	74,35	74,64	**74,93**	**74,93**	74,49

Source : https://archive.ics.uci.edu/ml/datasets/Credit+Approval

Note : Processed with Rapid Miner

Table 7. The accuracy of KNN for credit card customer dataset

K	1	3	5	7	9	11	13	15	17
ACC	**91,52**	**91,13**	90,74	89,96	89,56	89,43	89,3	89,43	89,56

Source : Bank credit card customers in Indonesia

Note : Processed with Rapid Miner

Improved algorithm accuracy occurs after the feature selection has been performed on some data attributes. Table 8 is the level of accuracy of the KNN algorithm in the UCI credit approval dataset after the feature selection is performed using information gain. Table 9 shows an increase of the accuracy of KKN dataset of credit card customers after feature selection is done using information gain.

Table 8. Increased accuracy of credit approval dataset

Number of used attributes	Names of attributes	Weight based on information gain	Accuracy level of KNN
15	a1	0	74.49
14	a12	5.31E-04	74.35
13	a13	0.01608465	74.64
12	a4	0.04780339	74.78
11	a5	0.04780339	74.78
10	a3	0.07606675	75.07
9	a7	0.09310051	75.36

8	a6	0.20339243	75.07
7	a8	0.20499992	75.36
6	a15	0.20539914	76.67
5	a10	0.29154533	82.46
4	a11	0.36097404	**82.46**
3	a14	0.54558132	81.59
2	a9	0.7955564	77.83
1	a2	1	44.49

Source : https://archive.ics.uci.edu/ml/datasets/Credit+Approval

Note : Processed with Rapid Miner

Table 9. Increased accuracy of credit card customer dataset

Number of used attributes	Names of attributes	Weight based on information gain	Accuracy level of KNN
14	type_of loan	0	91.52
13	bi_debtor_class	0.004	91.52
12	bi_economics_sector	0.006	91.52
11	sex	0.007	91.52
10	type of loan	0.012	91.52
9	uage	0.024	91.52
8	jkw	0.026	91.52
7	nominative_balance	0.054	91.52
6	amount_of installment per month	0.14	**94.78**
5	amount_of loan	0.204	93.6
4	col	0.224	92.56
3	bi_guarantor_class	0.24	92.3
2	interest_arrears	0.284	92.42
1	principle_arrears s	1	89.82

Source: Bank credit card customers in Indonesia

Note: Processed with rapid miner

4. Conclusion

Based on the results of the research, there is an increase of KNN class accuracy for both datasets after the feature selection. The highest level of accuracy for the UCI credit approval dataset before the feature selection is 74.93%. Meanwhile, the highest accuracy credit card customer dataset using all attributes is at 91.52%. After performing feature selection using only 5 attributes of the 15 attributes, the KKN accuracy level for the dataset of UCI credit approval rose to 82.46%. While the accuracy level for the dataset of credit card customers by using only 6 of the 14 attributes rose to 94.78%. This study proves that the information gain selection algorithm can improve the accuracy or performance of KNN classification. For the UCI credit approval dataset the accuracy increases at 7.53%, while the data set of credit card customers increases at 3.26%.

5. Acknowledgment

This research is part of a research funded by DIKTI in a Beginners Lecturer Research Scheme for fiscal year 2017.

References

Alkaromi, M. A. (2014). Information Gain untuk Pemilihan Fitur pada Klasifikasi Heregistrasi Calon Mahasiswa dengan Menggunakan K-NN.

Alpaydin, E. (2010). *Introduction to Machine Learning Second Edition*. London: The MIT Press.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. a., & Costa, L. D. F. (2013). A systematic comparison of supervised classifiers. Retrieved from http://arxiv.org/abs/1311.0202v1

Ashari, A., Paryudi, I., & Tjoa, A. M. (2013). Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, *4*(11), 33–39.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature Selection based on Information Gain, (2), 18–21.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques Second Edition*. *Elsevier*. Elsevier.

Ian H Witten. Eibe Frank. Mark A Hall. (2011). *Data Mining 3rd*.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative Study of Attribute Selection using Gain Ratio and Correlation Based Feature Selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Koprinska, I. (2010). Feature Selection for Brain-Computer Interfaces, 100–111.

Larose, D. T. (2005). *Discovering Knowledge in Data: an Introduction to Data Mining*. John Wiley & Sons.

Maimoon. (2010). *Data Mining and Knowledge Discovery Handbook*.

Maulana, M. R., & Al Karomi, M. A. (2016). Sistem Pendukung Keputusan Persetujuan Kredit Menggunakan Algoritma C4.5. *Jurnal IC-Tech*, *Vol. XI No*(1), 29–38. Retrieved from http://jurnal.stmik-wp.ac.id/gdl.php?mod=browse&op=read&id=ictech--muchrifqim-80

Patel, K., Vala, J., & Pandya, J. (2014). Comparison of various classification algorithms on iris datasets using WEKA, *1*(1), 1–7.

Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi menggunakan Matlab*. Yogyakarta: Andi Offset.

Ragab, A. H. M., Noaman, A. Y., Al-Ghamdi, A. S., & Madbouly, A. I. (2014). A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining. *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE ’14*, 106–113. https://doi.org/10.1145/2643604.2643631

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis* (Edisi Pert). Yogyakarta: Graha Ilmu.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2007). *Top 10 algorithms in data mining*. *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2.