The Comparison of Linear Regression Method and K-Nearest Neighbors in Scholarship Recipient

Okfalisa
State Islamic University of Sultan Syarif Kasim Riau
Pekanbaru, Indonesia
okfalisa@gmail.com

Ratika Fitriani
State Islamic University of Sultan Syarif Kasim Riau
Pekanbaru, Indonesia
ratika.fitriani@students.uin-suska.ac.id

Yelfi Vitriani
State Islamic University of Sultan Syarif Kasim Riau
Pekanbaru, Indonesia
yelfivitriani@yahoo.com

Abstract—**The scholarships award for students are often subjective, not transparent, un-measurable, and less precise on target. One of the computer technologies used to process big data such as scholarship recipient problems is data mining. Various methods of data mining can be used to predict the feasibility of data such as K-Nearest Neighbors (KNN) and Linear Regression. This study compares both methods in solving the scholarship recipient problem. The attributes used are Semester period, Grade Point Average (GPA), Statement Letter of Active Student, Letters of Assistance, Student Identity Card, Identity Card, Family Card, Study Result Card, Statement Letter, Bank Account, and Statement Letter of Passed Administration. The variables used in the comparison process include Accuracy, Precision, Recall, Classification Error, Absolute Error, and Root Mean Square Error (RMSE). Data from 8212 scholarship recipients are tested through simulation testing of training data and testing data 90:10, 70:30, 50:50, 30:70, and 10:90. Herein, Rapidminer is used as a tool to view the results of analysis from both methods. As a result, both methods for data simulation 90:10 and 70:30 provide 100% of accuracy, precision, and recall. Meanwhile, KNN from data simulation 50:50, 30:70, and 10:90 provide better performance in accuracy, precision, recall, classification error, absolute error and RMSE than in Linear Regression with comparison of mean differences are 17.79%, 18.1%, 10.83%, 17.79%, and 0.25 respectively. KNN and Linear Regression methods have been successfully applied to classify and cluster the data of scholarship recipients. The result has shown that KNN method is more effective and efficient rather than Linear Regression method. This provides new knowledge contribution. Hopefully, the selection process of scholarship recipients can be implemented much better, transparent, no longer subjective, and right on the target.**

Keywords—Data mining, K-Nearest Neighbors (KNN), Linear Regression, Rapidminer, Accuracy, Precision, Recall, Classification Error, Absolute Error, and Root Mean Square Error (RMSE), Scholarship

I. INTRODUCTION

A scholarship award is an appreciation given to individual scholar in order to continue their education to a higher level. The reward given can be a special access to an institution or financial assistance. Basically, the scholarship provides an income for those who receive it [1]. Usually, it is in form of funds spent for students during the course time in the desired study period. Riau Province Government always offers this scholarship program every year. Unfortunately, the scholarship given to the students is subjective thus a lot of eligible students are not getting the scholarships and vice versa. Emotional factor and relationships with the staffs in charge in this program make it a un-fair condition in the election. Data mining is extracting useful information from large datasets. There are a lot of data problems solved by using data mining techniques such as association, prediction, classification, and clustering. To solve the scholarship recipient problem, the classification and cluster will be done by using data mining techniques. It is conducted by comparing two methods, Linear Regression and K-Nearest Neighbors. Linear Regression is a statistical method used to form a model of the relationship between dependent variables with one or more independent variables. If the number of independent variables is only one, it is called a Simple Linear Regression, [2] else it can be called as Multiple Linear regression models. This model is the most popular data-driven model for their easy application and very well known techniques in parametric methods. Meanwhile, the relationship between dependent and independent variables is far and not true linear provides a poor fit of this method to the data [3, 4].

K-Nearest Neighbors or commonly abbreviated with KNN is a nonparametric method to classify new data whose class is unknown yet and choose data as much as k which is nearest located from new data. Generally,k is determined as an odd number to avoid the appearance of the same amount of distance in the classification process. This method provides a more flexible approach with an explicit form for $f(k)$. This method often more complex to understand and interpret. For some of the observation on data predictor, parametric methods work better [3]. The KNN classifier is one of the most popular neighborhood classifiers in pattern recognition. If compared with Bayes algorithm and other Euclidean distance calculation, K-Nearest neighbor algorithm has better efficiency and performance [5].

Previous research conducted by Sumarlin [6] entitled "Implementation of K-Nearest Neighbors Algorithm as Decision Support of Improvement of Academic Achievement (Peningkatan Prestasi Akademik-PPA) and Students Scholarship (Bantuan Belajar Mahasiswa-BBM) classification" explained that from 227 records of PPA scholarship datasets, the accuracy reached 77.96% , while for BBM scholarship used 183 record datasets, the accuracy reached 97.28%. The combination of PPA and BBM scholarships reached the accuracy of 85.56%. Another study by Mustafidah [7] entitled "Data Mining Regression Model of Motivation Learning's Influence against Student Discipline Level". This research studied the relationship between motivation learning's influences against discipline level in some following classes in Indonesia universities. They used two kinds of the regression model, linear and quadratic. As the result, the linear regression model provided the error of MSE (Mean Square

978-1-5386-5889-5/18/$31.00 ©2018 IEEE
SNPD 2018, June 27-29, 2018, Busan, Korea

Error) equal to 8.13% while quadratic regression model was 8.15%. It can be concluded that Linear Regression is more suitable for this cases. Another study by Arto and Annika [8] proposed the average value of RMSE in KNN is smaller than Linear Regression. When the results were examined within diameter classes, KNN result is less bias than regression model results, especially with extreme values of diameter. The difference between these models as more obvious when the assumed model form is not exactly correct. Furthermore, the implementation and advancement of KNN Model were described by Bahar et.al [9] with the title "Model of Scholarship Type Determinants using KNN Algorithm Combination of Rule Base and Knowledge Base". They found that it was more flexible in determining the status of testing data through this combination. The result from Leidiyana and Hanny [10] found that KNN has an accuracy of 81.46% and includes a very good data classification. It was due to an Area Under Curve (AUC) value was between 0.90-1.00, and equal to 0.98.

Based on these above studies, the comparison of Linear Regression and K-Nearest Neighbors methods toward the classification of scholarship recipients data are studied in this paper. The comparison of accuracy, precision, recall, classification error, absolute error, and RMSE are analyzed to show the discrepancies the models and examined the better used in the classification of the scholarship recipient data.

II. Literature Review

A. Data Mining

Data mining is a process of extracting and identifying information and knowledge in large volumes and how the relationships can be built using statistical, mathematics, artificial intelligence, and machine learning techniques (Turban, et al, 2005) [4, 10]. The process of data mining is performed in knowledge discovery database (KDD) through several stages as shown in Fig. 1 [11]. It includes the selection process-as data selection; pre processing-as cleaning data process from missing values, duplicate data, in-consistency data, and wrong data; transformation-as dimensional reduction with normalization ranges in 0-1. For data normalization, it can be calculated by using the formula below.

$$v^i = \frac{v\text{-}mm_a}{max_a\text{-}min_a}(new\_max_a\text{-}new\_min_a)+new\_min_a \qquad (1)$$

where,

v^i	= new data after normalization
v	= data before normalization
new_max_a	= new max value limit is 1
new_min_a	= new min value limit is 0
max_a	= maximum value in column
min_a	= minimum value in column

The last stage is to form the data mining pattern as a form of interpretation/evaluation. Herein, the Linear Regression method and K-Nearest Neighbors method are applied.

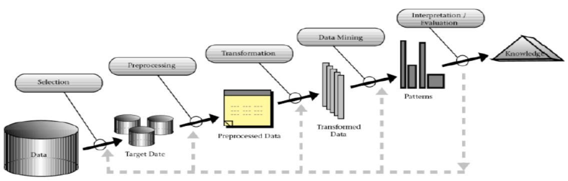


Fig. 1. Stages of Knowledge Discovery in Database –KDD[11]

B. Linear Regression Process

Linear Regression classification method is one of data processing method in the concept of data mining. The steps performed on this method include [6]:

1) Calculating the mean of each attribute:

$$\bar{X} = \sum_{i=1}^{n} x_i \qquad (2)$$

2) Calculating the standard deviation:

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}} \qquad (3)$$

where,
S= deviation standards
x_i= value of x to i
$\bar{x}$= mean
n= number of attributes

3) Calculating the value of a and b_n:Steps that must be done is to find the value of a and b_n in order to obtain linear regression equation.

$$a = \frac{Det(A0)}{Det(A)} \qquad (4)$$

$$bi = \frac{Det(Ai)}{Det(A)} \qquad (5)$$

where:
a = constants
b = regression coefficient value
det= determinant

4) Creating a linear regression equation: After getting the value of a and b_n, then do the calculation by using the equation of linear regression line.

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (6)$$

where:
Y = dependent variable (predicted value)
a = constants
b = slope / correlation coefficient (increase or decrease value)
x = independent variable

5) Partial correlation testing

$$r_{xiy} = \frac{n\sum x_1 y - (\sum y \sum x_1)}{\sqrt{[(n\sum y^2) - (\sum y)^2][(n\sum x_1^2) - (\sum x_1)^2]}} \qquad (7)$$

where:

r_{xiy} = partial correlation test

∑= Total amount of x_iy

6) Calculating the correlation coefficient

$$\sum y^2 = \sum(y^2) - \frac{(\sum y)^2}{n} \tag{8}$$

$$\sum x_iy = \sum(x_iy) - \frac{(\sum x_i)(\sum y)}{n} \tag{9}$$

7) Calculating the relative contribution of both predictors (R^2)

$$R^2 = \frac{JK_{reg}}{\sum y_i^2} \tag{10}$$

where:

JKreg (sum of squares regression)= $a_1 \sum x_{1i}y_i + a_2 \sum x_{2i}y_i$

JKres $= \sum (y_i - \bar{y}_i)^2$

8) Calculating R simultaneous

$$R_{simultan} = \sqrt{R^2} \tag{11}$$

9) Finding the coefficient of determinant's value

$$KP = R_{simultan} \times 100\% \tag{12}$$

C. K-Nearest Neighbors Process

The steps in K-Nearest Neighbors process are explained below [12, 13].

1) Determination n initial data as the initial knowledge-base of the system.

2) Determination of the nearest K's value (neighbor).

3) Preparation of the training data in the form of the criteria's value of new data that the status has not been known.

4) Determination of the status of each training data based on certain rules to generate knowledge-base of the system.

5) Calculation of the distance of each sample of training data against data to be tested (testing data) based on the Euclidian equation:

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \tag{13}$$

where:

d = the distance between the points in training data x and the points in testing data y that will be classified, where$x = x_1, x_2 \ldots, x_i$ and $y = y_1, y_2 \ldots, y_i$

i = represents the value of attribute

n = an attribute dimension.

6) Setting the status of testing data based on the average value of K nearest training data samples.

III. Research Methodology

This research is conducted by following several stages as described in Fig. 2. Stage 1 is the process of problem formulation related to the comparison of both methods and the process of receiving scholarship managed by Riau Province Government as a case study. Furthermore, the literature study is reviewed to understand the basic theory of KNN and Linear Regression. The differences and similarities of both methods are investigated to find the variables that serve as a benchmark comparison between the two methods. As a result, comparable variables are defined as viz.accuracy, precision, recall, absolute error, classification error, and RMSE value. A literature review is studied through various sources and references from books or e-books, articles, national and international journals and proceedings. The data collection by Provincial Government of Riau in 2013 and 2014 consisted of 8,212 data.

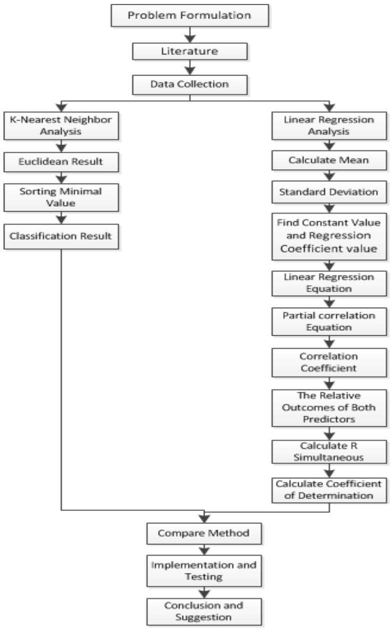

Problem Formulation
Literature
Data Collection
K-Nearest Neighbor Analysis
Euclidean Result
Sorting Minimal Value
Classification Result
Linear Regression Analysis
Calculate Mean
Standard Deviation
Find Constant Value and Regression Coefficient value
Linear Regression Equation
Partial correlation Equation
Correlation Coefficient
The Relative Outcomes of Both Predictors
Calculate R Simultaneous
Calculate Coefficient of Determination
Compare Method
Implementation and Testing
Conclusion and Suggestion

Fig. 2. Stages of Research Methodology

The criteria attribute that determined by agreement with the People's Welfare Bureau of Riau Province are Semester, Grade Point Average (GPA), Statement Letter of Active Student, Letters of Assistance, Student Identity Card, Identity Card, Family Card, Study Result Card, Statement Letter, Bank Account, and Statement Letter of Passed Administration. The collected data come from the data of scholarship recipients who had passed the administrative requirements. The next stage is the analysis. The analysis is performed for overall data and step set based on KNN and Linear Regression algorithms. This process is conducted by using Rapidminer tool to compare the results of both methods through the simulation of training data and testing data including 90:10, 70:30, 50:50, 30:70, and 10:90. The last stage of this research is to obtain conclusion and suggestion for improvement.

IV. Result and Discussion

A. KDD Process

The implementation of the KDD provides several outputs. As a result of data cleaning process, 8,212 data are of missing values from the basic of 21,717 raw data. In this case study, the data integration process is skipped due to its single data performed. Furthermore, based on the interviews and observations in People's Welfare Bureau of Riau Province, 11 attributes on the data are obtained and ready to be used for KDD process. The attributes used are Semester, Grade Point Average (GPA), Statement Letter of Active Student, Letters of Assistance, Student Identity Card, Identity Card, Family Card, Study Result Card, Statement Letter, Bank Account, and Statement Letter of Passed Administration. While other attributes such as birth year, parent's birthplace, SKAK (Statement Letter of Active Students), SPTMBPL (Statement Letter for not receiving another scholarship program), SPKD (Statement Letter of Data Validity), SKTM (Statement Letter for Economic Status), and description only serve as the support attributes. Furthermore, the transformation process is done by changing the data format into normalization ranges from 0-1 based on the formula explained in chapter II and equation (1). The result of transformation can be seen in Fig. 3.

B. *Development of Rapidminer Analysis Series*

The Rapidminer analysis series used are developed in Fig. 4 and Fig. 5.The series is made based on simulation of training data and testing data as 90:10, 70:30, 50:50, 30:70, and 10:90. The performance measurement results are set according to the required comparison variables, viz. accuracy, precision, recall, absolute error, classification error, and RMSE values. The similar process is conducted for Linear Regression as well as described in Fig. 5.

C. *Comparative Analysis of Testing Results*

1) KNN testing results for training data 90% and testing data 10%

As the result of measurement, the confusion matrix for KNN is explained in Table I. Table I shows that the tests performed using 90% training data and 10% testing data on each value of K (K1, K3, K5, K7, and K9) obtained the results with True Positive (TP) is 401, False Negative (FN) is 0, False Positive (FP) is 0, and True Negative (TN) is 422 with 100% accuracy, 100% Precision, 100% Recall, 0% error classification, absolute error is 0.000, and root mean squared error is 0.000.

TABLE I. CONFUSION MATRIX KNN 90:10

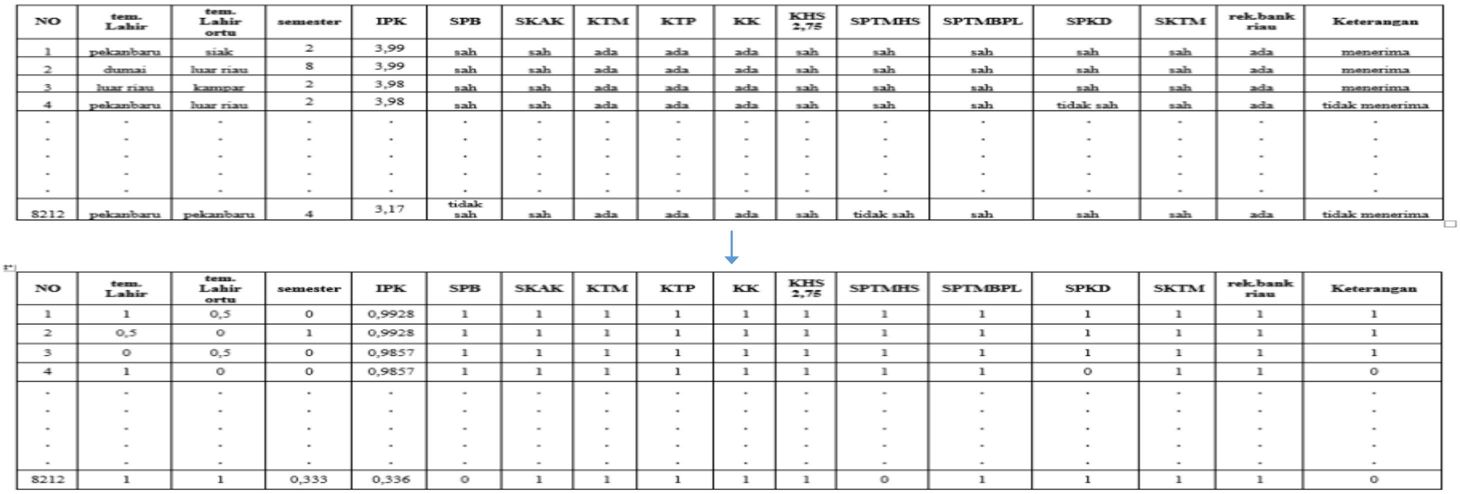
K Value	TP	FP	FN	TN	Accuracy	Precision	Recall	Classification error	Absolute error	Root mean squared error
K 1	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 3	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 5	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 7	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 9	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000

2) KNN testing results for training data 70% and testing data 30%

The confusion matrix result with KNN method for 70:30 can be seen in Table II. It explains that KNN test diagram on K1, K3, K5, K7, and K9 provides 100% accuracy, 100% precision, 100% Recall, and 0.00% of classification error.

TABLE II. CONFUSION MATRIX KNN 70:30

K Value	TP	FP	FN	TN	Accuracy	Precision	Recall	Classification error	Absolute error	Root mean squared error
K 1	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 3	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 5	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 7	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 9	401	0	0	422	100%	100%	100%	0%	0.000 +/- 0.000	0.000



NO	tem. Lahir	tem. Lahir ortu	semester	IPK	SPB	SKAK	KTM	KTP	KK	KHS 2,75	SPTMHS	SPTMBPL	SPKD	SKTM	rek.bank riau	Keterangan
1	pekanbaru	siak	2	3,99	sah	sah	ada	ada	ada	sah	sah	sah	sah	sah	ada	menerima
2	dumai	luar riau	8	3,99	sah	sah	ada	ada	ada	sah	sah	sah	sah	sah	ada	menerima
3	luar riau	kampar	2	3,98	sah	sah	ada	ada	ada	sah	sah	sah	sah	sah	ada	menerima
4	pekanbaru	luar riau	2	3,98	sah	sah	ada	ada	ada	sah	sah	sah	tidak sah	sah	ada	tidak menerima
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8212	pekanbaru	pekanbaru	4	3,17	tidak sah	sah	ada	ada	ada	sah	tidak sah	sah	sah	sah	ada	tidak menerima

NO	tem. Lahir	tem. Lahir ortu	semester	IPK	SPB	SKAK	KTM	KTP	KK	KHS 2,75	SPTMHS	SPTMBPL	SPKD	SKTM	rek.bank riau	Keterangan
1	1	0,5	0	0,9928	1	1	1	1	1	1	1	1	1	1	1	1
2	0,5	0	1	0,9928	1	1	1	1	1	1	1	1	1	1	1	1
3	0	0,5	0	0,9857	1	1	1	1	1	1	1	1	1	1	1	1
4	1	0	0	0,9857	1	1	1	1	1	1	1	1	0	1	1	0
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8212	1	1	0,333	0,336	0	1	1	1	1	1	0	1	1	1	1	0

Fig. 3. Transformation of data normalization

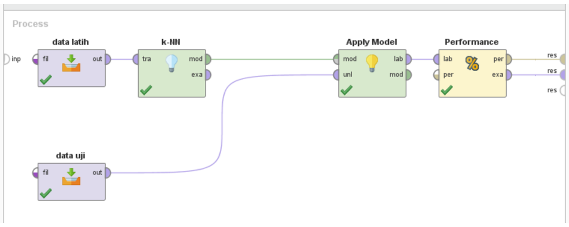


Fig. 4. Rapidminer series for K-NN process

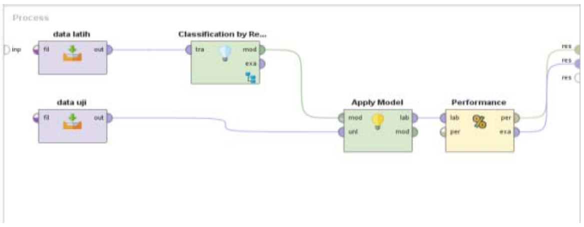


Fig. 5. Rapidminer series for linear regression process

3) KNN testing results for training data 50% and testing data 50%

The confusion matrix of KNN method for 50:50 can be seen in Table III. It provides K1 with True Positive (TP) is 802, False Negatif (FN) is 0, False Positif (FP) is 1,and True Positif (TP) is 3303 with accuracy, 99.98%, Precision 99.94%, Recall 99.98%, classification error 0.02%, and root mean squared error 0.016. For K3 finds accuracy 99.22%, Precision 98.08 %, Recall 99.52%, classification error 0.78%, dan root mean squared error 0.073. K5 with similar values of TP, FN, FP and TP provides accuracy 97.57 %, Precision 94.51 %, Recall 98.50 %, classification error 2.41%, dan root mean squared error 0.104. For K7 generates accuracy 97.49 %, Precision 94.49 %, Recall 94.31%, classification error 2.51 %, dan root mean squared error 0.119. Finally K9 provides accuracy in 97.49 %, Precision 94.49 %, Recall 94.31%, classification error 2.51 %, dan root mean squared error 0.126.

TABLE III. CONFUSION MATRIX KNN 50:50

K Value	TP	FP	FN	TN	Accuracy	Precision	Recall	Classifica tion error	Absolute error	Root mean squared error
K 1	802	1	0	3303	99.98%	99.94%	99.98%	0.02%	0.000 +/- 0.016	0.016
K 3	802	32	0	3272	99.22%	98.08%	99.52%	0.78%	0.011 +/- 0.072	0.073
K 5	802	99	0	3205	97.57%	94.51%	98.50%	2.41%	0.016 +/- 0.103	0.104
K 7	802	103	0	3201	97.49%	94.31%	98.44%	2.51%	0.019 +/- 0.117	0.119
K 9	802	103	0	3201	97.49%	94.31%	98.44%	2.51%	0.020 +/- 0.125	0.126

4) KNN testing results for training data 30% and testing data 70%

The confusion matrix of KNN method for 30:70 can be seen in Table IV. It provides K1 with True Positive (TP) is1,205, False Negative(FN) is 0, False Positive (FP) is 0, and True Negative (TN) is 4,543 with 100% accuracy, 100% precision, 100% recall, 0.00% classification error, and root mean squared error is 0.000.

TABLE IV. CONFUSION MATRIX KNN 30:70

K Value	TP	FP	FN	TN	Accuracy	Precision	Recall	Classific ation error	Absolute error	Root mean squared error
K 1	1205	0	0	4543	100%	100%	100%	0%	0.000 +/- 0.000	0.000
K 3	1205	1	0	4542	99.98%	99.96%	99.99%	0.02%	0.001 +/- 0.016	0.035
K 5	1205	10	0	4533	99.83%	99.59%	99.89%	0.17%	0.003 +/- 0.035	0.035
K 7	1205	19	0	4524	99.67%	99.22%	99.79%	0.33%	0.006 +/- 0.053	0.053
K 9	1205	47	0	4496	99.18%	98.12%	99.48%	0.82%	0.008 +/- 0.066	0.066

For K3 provides 99.98% accuracy, 99.96% precision, 99.99% recall, 0.02% classification error, and root mean squared error is 0.035. For K5, it finds 99.83% accuracy, 99.59% precision, 99.89% recall, 0.17% classification error, and root mean squared error is 0.035. K7 produces 99.67% accuracy, 99.22% precision, 99.79% recall, classification error is 0.33%, and root mean squared error is 0.053. Finally, K9 provides 99.18% accuracy, 98.12% precision, 99.48% recall, 0.82% classification error, and root mean squared error is 0.066.

5) KNN testing results for training data 10% and testing data 90%

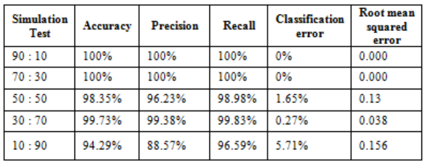
The confusion matrix result with KNN method for 10:90 shown at Table V. It shows that K1 True Positive (TP) is 1204, False Negative (FN) is 0, False Positive (FP) is 3, and True Negative (TN) is 6182, it provides 99.96% accuracy, 99.88% precision, 99.98% recall, 0.04% classification error, and root mean squared error is 0.020. For K3 finds 98.73% accuracy, 96.38% Precision, 99.24% Recall, 1.27% classification error, and root mean squared error is 0.116. K5 provides 92.79% accuracy, 84.66% precision, 95.69% recall, 7.21% classification error, and root mean squared error is 0.184. K7 provides 90.24% accuracy, 81.27% precision, 94.17% recall, 9.76% classification error, and root mean squared error is 0.218. Finally, K9 generates 89.74% accuracy, 80.68% precision, 93.87% recall, 10.26% classification error, and root mean squared error is 0.241.

As the summarization, the average values of KNN from K1 to K9 for the entire simulation can be depicted from Table VI. Table VI explained that the values of accuracy, precision, recall, classification error, and root mean squared error is produced during the simulation test in 30:70 as the highest performance, followed by 50:50, and 10:90 respectively. From Tables III, IV, and V, they find that the smaller values of K will provide the highest values of accuracy, precision, recall, classification error, and root mean squared error. Herein, K1 has the highest performance, it is then followed by K3, K5, K7, and K9 as the lowest one.

TABLE V. CONFUSION MATRIX KNN 10:90

K Value	TP	FP	FN	TN	Accuracy	Precision	Recall	Classificati on error	Absolute error	Root mean squared error
K 1	1204	3	0	6182	99.96%	99.88%	99.98%	0.04%	0.000 +/-0.020	0.020
K 3	1204	94	0	6091	98.73%	96.38%	99.24%	1.27%	0.032 +/-0.112	0.116
K 5	1204	533	0	5652	92.79%	84.66%	95.69%	7.21%	0.059 +/-0.174	0.184
K 7	1204	721	0	5464	90.24%	81.27%	94.17%	9.76%	0.074 +/-0.205	0.218
K 9	1204	758	0	5427	89.74%	80.68%	93.87%	10.26%	0.085 +/-0.225	0.241

TABLE VI. KNN RESULTS

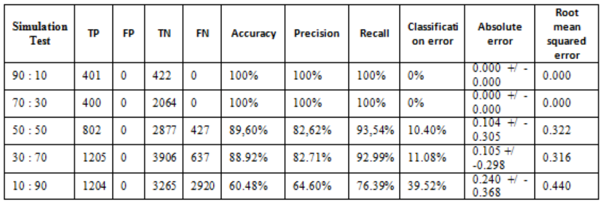


Simulation Test	Accuracy	Precision	Recall	Classification error	Root mean squared error
90 : 10	100%	100%	100%	0%	0.000
70 : 30	100%	100%	100%	0%	0.000
50 : 50	98.35%	96.23%	98.98%	1.65%	0.13
30 : 70	99.73%	99.38%	99.83%	0.27%	0.038
10 : 90	94.29%	88.57%	96.59%	5.71%	0.156

A. *Testing Result of Linear Regression*

By following the equation in Chapter 2, the confusion matrix for Linear Regression is explained in Table VII.

TABLE VII. CONFUSION MATRIX OF LINEAR REGRESSION



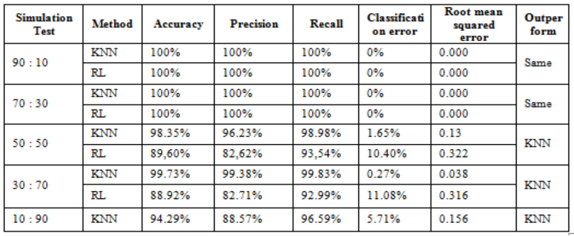
Simulation Test	TP	FP	TN	FN	Accuracy	Precision	Recall	Classificati on error	Absolute error	Root mean squared error
90 : 10	401	0	422	0	100%	100%	100%	0%	0.000 +/ - 0.000	0.000
70 : 30	400	0	2064	0	100%	100%	100%	0%	0.000 +/ - 0.000	0.000
50 : 50	802	0	2877	427	89,60%	82,62%	93,54%	10.40%	0.104 +/ - 0.305	0.322
30 : 70	1205	0	3906	637	88.92%	82.71%	92.99%	11.08%	0.105 +/ -0.298	0.316
10 : 90	1204	0	3265	2920	60.48%	64.60%	76.39%	39.52%	0.240 +/ - 0.368	0.440

It showed that from simulation test in 90:10 and 70:30, the performance of accuracy, precision, recall, classification error, and root mean squared error are 100%. For simulation in 50:50, True Positive (TP) is 802, False Negative (FN) is 427, False Positive (FP) is 0, and True Negative (TN) is 2877,found the performance are 89.60% accuracy, 82.62%precision, 93.54% recall, 10.40% classification error, and Root Mean Squared Error is 0.322. For simulation testing in 30:70, provides 88.92% accuracy, 82.71% precision, 92.99% recall, 11.08% classification error, and root mean squared error is 0.316. Finally for simulation testing in 10:90, it found 60.48%accuracy, 64.60% precision, 76.39% recall, 39.52% classification error, and RMSE is 0.440.

B. *The Results of Comparison Method*

The summarization of two methods analysis is explained in Table VIII.

TABLE VIII. COMPARISON ANALYSIS OF KNN AND LINEAR REGRESSION



Simulation Test	Method	Accuracy	Precision	Recall	Classificati on error	Root mean squared error	Outper form
90 : 10	KNN	100%	100%	100%	0%	0.000	Same
	RL	100%	100%	100%	0%	0.000	
70 : 30	KNN	100%	100%	100%	0%	0.000	Same
	RL	100%	100%	100%	0%	0.000	
50 : 50	KNN	98.35%	96.23%	98.98%	1.65%	0.13	KNN
	RL	89,60%	82,62%	93,54%	10.40%	0.322	
30 : 70	KNN	99.73%	99.38%	99.83%	0.27%	0.038	KNN
	RL	88.92%	82.71%	92.99%	11.08%	0.316	
10 : 90	KNN	94.29%	88.57%	96.59%	5.71%	0.156	KNN

From this table we can compare that KNN and Linear Regression provides the different performance values in simulation test 50:50, 30:70, and 10:90. The performance measurement of accuracy, precision, recall, classification error, and root mean squared error in KNN provides the better values rather than in Linear Regression with a comparison of mean differences are 17.79%, 18.1%, 10.83%, 17.79%, and 0.25 respectively. Based on the values of TP, TN and FN for Linear Regression in Table VII compared to the values for KNN in Table III, IV, and V, the predictions of KNN is more accurate than Linear Regression.

V. CONCLUSION

This paper completed the step process of data mining in classifying and clustering 8,212 scholarship recipients data in Riau Province. To find the right classification methods, the comparison analysis between KNN and Linear Regression is conducted. As the conclusion, KNN provides better performance in accuracy, classification error, weighted mean recall, weighted mean precision, absolute error, and root mean squared error than Linear Regression, especially for simulation testing in 50:50, 30:70 and 10:90. It is also shown that the smaller values of K in KNN will provide a greater value of performance. This supported the statement of Arto and Annika [8] that the average value of RMSE in KNN is smaller than Linear Regression. For better accuracy and performances, KNN method is better applied in classifying and clustering the scholarship recipients data than Linear Regression. Therefore, the selection process can be implemented better, transparent, no longer subjective, and right on the target.

REFERENCES

[1] Larose, Daniel T. 2005. “Discovering Knowledge in Data: An Introduction to Data Mining”. Jonhn Willey & Son, Inc.Bahar, Nidia Rosmawanti. 2014. “Model Penentu Jenis Beasiswa Menggunakan Algoritma K-NN Kombinasi Basis Aturan Dan Basis Pengaturan”. ISSN: 2089-3787.

[2] Susanto, Heri. 2014.”Data Mining untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisplinan dan Prestasi Masa Lalu “. Jurnal Pendidikan Vokasi, Vol 4, Nomor 2. Universitas Negeri Yogyakarta.

[3] Shahab Araghinejad. 2013. “Data Driven Modeling: Using Matlab in Water Resources and Environmental Engineering”. Springer.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2008.”The Elements of Statistical Learning: Data Mining, Inference,and Prediction”. Second Edition. Springer.

[5] Aman Kataria and M.D. Singh. 2013. “A Review of Data Classification Using K-Nearest Neighbour Algorithm”. International Journal of Emerging Technology and Advance Engineering. Vol.3, Issue 6.

[6] Sumarlin. 2015. “Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM”. Jurnal Sistem Informasi Bisnis 01(2015) on-line: http://ejournal.undip.ac.id/index.php/jsinbis.

[7] Mustafidah, Hindayati. 2010. “Model Regresi Data Mining Motivasi Belajar Pengaruhnya Terhadap Tingkat Kedisiplinan Mahasiswa”. JUITA Vol. I Nomor 1.

[8] Haara, arto and Annika kangas. “Comparing K Nearest Neighbours Methods And Linear Regression—Is There Reason To Select One Over The Other”. ISSN 1946-7664.MCFNS 2012.

[9] Bahar, Nidia Rosmawanti. 2014. “Model Penentu Jenis Beasiswa Menggunakan Algoritma K-NN Kombinasi Basis Aturan Dan Basis Pengaturan”. ISSN: 2089-3787

[10] Leidiyana, Henny. 2013. “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor”. Jurnal Penelitian Ilmu Komputer, System Embedded & Logic 1(1) : 65-76 (2013).

[11] Fayyar,E.M., Piatetsky-Shapiro, G., Smyth, P., Ulthurusamy, R.(Eds), 1996. Advances in Knowledge Discovery and Data Mining.Cambridge, MA:MIT Press,573-593.

[12] R.O. Duda and P.E. Hart. 1973. “Pattern Classification and Scene Analysis”, New York: John Wiley & Sons, 1973.

[13] Aman Kataria and M.D. Singh. 2013. “A Review of Data Classification Using K-Nearest Neighbour Algorithm”. International Journal of Emerging Technology and Advance Engineering. Vol.3, Issue 6.