The Analysis of Major Interest State Senior High SchoolthroughLogistic Regressionand K-Nearst Neighbor

Tiani Wahyu Utami[1], TestianaDeniWijayatiningsih[2], Martyana Prihaswati[3], Abdul Karim[4]
tianiutami@unimus.ac.id[1], testiana@unimus.ac.id[2], martyana@unimus.ac.id[3], abdulkarim@walisongo.ac.id

Universitas Muhammaddiyah Semarang,Jl. Kedungmundu Raya No. 18, Semarang, Indonesia[1,2,3]
Universitas Islam Negeri Wali Songo, Jl. Walisongo No 3-5, Semarang, Indonesia[4]

Abstract. Education is a conscious and planned effort to create a learning atmosphere and learning process so that learners are actively developing their potential. At the high school level, learners will follow a major interest. Majors interest is conducted to provide students with opportunities to choose the subject of interest, deepen the subject matter and develop their potential in a flexible range. The State High School 1 Salemopened 2 majors for the continuity of the students' learning process, namely Science andSocial.The factors that influence the learners in majors interest namely internal factors and external factors.The research variable used are the major interest as the response variable and seven predictor variables.The seven predictor variables consist of the math national examination score, science national examination score, language national examination score, the relationship of students and their friends, the relationship of students and their teachers, the relationship of students and their family, and self motivation. The classification analysis method used are binary logistic regression and K-NN.Based on the results obtained from logistic regression, the alleged factor to affect the majors' interest at state Senior High School are the score of Science National Examination and the students' relationship with their friends. The interest classification system of State High School 1 Salem which resulted from binary logistic regression method has a precision is73.7%. The interest classification system of State High School 1 Salem which resulted from the K-NN method has a accuracy of 82.6%. The best classification method for majors' interestatState High School1 Salem is the K-NN method.

.**Keywords:** Senior High School Major, Logistic Regression, K-NN

1 Introduction

Education is a conscious and well-planned endeavor to create a learning atmosphere and learning process so that students actively develop their potential to possess religious spiritual strength, self-control, personality, intelligence, and the skills required by him, society, nation and State (Law of the Republic of Indonesia number 20 year 2003, article 1 paragraph 1). In the process of studying the students can choose the majors in their respective high school. There are high schools that open 3 majors, namely Science, Social, and Language. Besides, there is also high school that only open 2 majors, namely Science and Social. The determination of the opening majors in high school is back in their respective school policies.

ICE 2019, September 28-30, Purwokerto, Indonesia

DOI 10.4108/eai.28-9-2019.2291097

Schools play an important role in developing their students' potential according to their skills or majors [1].

Further, at the high school level, the learners will follow a major interest. Majors interest is conducted to provide students with opportunities to choose the subject of interest, deepen the subject matter and develop a variety of potential in it flexibly according to common basic skills (intelligence), talents, interests and personality characteristics without being constrained by the partition of the majors that are too rigid [2]. In this research the researchers took the research object at the State High School 1 Salem district of Pekalongan Central Java opened 2 majors for the continuity of the learning process, namely Science and Social. The possibility that will happen if the student has an error in the study is a low student learning achievement or can cause a mismatch with the majors that have been chosen by the student or previous students [3].According to [4], the factors that influence the learners in majors interest namely internal factors and external factors. In statistical sciences, many methods can be used to determine the influence of predictor variables on the response variables of a category [5].

Classification is one of the statistical methods to group or classify data that is arranged systematically [6]. Classification problems are often found in everyday life. Whether it's the classification of data in the academic, social, government, and other fields [7]. This classification problem arises when there are a number of measurements consisting of one or several categories that cannot be identified directly but must use a measure [8].

Several classification methods are regression methods of logistic binary and K-Nearest Neighbor. According to Hosmer and Lemeshow [9], the method of logistic regression is a statistical analysis method describing the relationship between a category-scaled response variable that has two or more categories (binaries) with one or more predictor variables. A binary (Variableresponse) is a response variable that is only 1 for the existence of a characteristic and 0 for the absence of such characteristics. K-Nearest Neighbor (K-NN) is a classification method that specifies categories based on the majority of categories in K-Nearest Neighbor [10]. K-NN is done by looking for group K objects in the training data closest to the object in the data testing [11]. In this research will be conducted analysis of the classification of high school in Salem, with logistic regression and K-NN method.

Since that time, the use of logistic regression hasestimation of customer dissatisfaction [12], multivariate logistic regression analysis [13], and quality of logisticregression [14].Previous research including, the study used K-NN to predict student vgraduation on time [15],recognition number of the vehicle plate using K-NN [16]. A classification system is expected to be able to classify all data sets correctly, but it cannot be denied that the performance of a system is not 100% correct so a classification system must also measure its performance [17]. Generally, performance measurements are carried out with a cofusion matrix [18].The comparison of the classifiers and using the most predictive classifier is very important. Each of the classification methods shows different efficacy and accuracy based on the kind of datasets [19].

2 Method

2.1 Data and Research Variable

The data which wereused in this research is the primary data obtained by distributing the questionnaire to the students of Sate High School 1 SalemPekalongan, Central Java. The

total students were 224 students which were consisted of 98 for Social and 126 majoring in Science.This research analyzes data using the R program, logistic regression package and KNN package [20].

Table 1. Research Variable

Variable	Definition
VariableResponse	
Majors interest(Y)	1= Social 2=Science
VariablePredictor	
Math National Examination Score(x_1)	
Science National Examination Score (x_2)	
Language National Examination Score (x_3)	
The relationship of students and their friends(x_4)	1= Low 2= Middle 3 = High
The relationship of students and their teachers(x_5)	1= Low 2= Middle 3 = High
The relationship of students and their family (x_6)	1= Low 2= Middle 3 = High
Self motivation(x_7)	1= Low 2= Middle 3 = High

2.2Research Method

The research analysis steps is:

1. Data retrieval continued with data encoding to become data ready
2. Analyzing data by Binary logistic regression method :
   a. Detemine a multiple hypothesis test or test a model using the G test statistics.
   b. Determine a partial hypothesis test using Wald test statistics.
   c. Analysis of the effect of each variable.
   d. Determine classification result for binary logistic regression.
3. Analyzing data by K-NN method :
   a. Calculate the Euclidean distance between test data and training data,
   b. Determine the group of test result data based on the majority label from the nearest neighbor.
   c. Determine the accuracy of the method using Confussion matrix.
4. Comparing the classification results of both methods.

3Results and Discussion

3.1.Logistic Biner Regression

Regression logistic model which had been created:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (1)$$

with :

$$g(x) = -20{,}4 + 0.015x_1 - 0.062x_2 + 0.32x_3 + 1{,}206x_{4.1} + 0{,}987x_{4.2} - 0.361x_{5.1} - 0.88x_{5.2} + 0.271x_{6.1} + 0{,}792 + 0.008x_{7.1} + 22.354x_{7.2}$$

Furthermore,it continued by testing the significance of the parameters either together or each of the predictor variables.

1.Likelihood Ratio Test

Hipotesis

H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9$(predictor variable does not affect the model together)

H1 :minimum there is one$\beta_j \neq 0$,withj=1,2,...,9.(predictor variable affects the model together)

Significance level:α=5%

Statistical Test: $G = -2\ In \frac{likeli\ hood\ \ tanpa\ \ variabel\ \ bebas}{likeli\ hood\ \ dengan\ \ variabel\ \ bebas}$ (2)

Test Criteria: $H_0$is rejected if G >χ^2 (0,05;9)

The decision:because G=219,451>16,919 χ^2 (0,1;7)so, $H_0$is rejected

The conclusion:so, at significance level5%,it can be concluded that predictor variable influences the model together.

2.Wald Test

Hipotesis

H_0: $\beta_j = 0$(variable jdoes not affect the model)

H_1: $\beta_j \neq 0$,untukj=1,2,...,9.(variable j affects themodel)

Significance level:α=5%

Statistical Test:$W = \frac{\beta_j}{Se\ (\beta_j)}$ (3)

Test Criteria: $H_0$is rejected if$W > \chi^2$ (0,1;9)=3,481

Wald score for each variable can be seen as follows.

Table 2. Wald Test

Predictor Variable	Wald (*W*)	Sig	Decision
Math National Exam Score(x_1)	0.290	0.590	accepted Ho
Science National Exam Score(x_2)	3.688	0.055	rejected Ho
English National Exam Score(x_3)	0.575	0.448	accepted Ho
The relationship between students and their friends(x_4)			
Low($x_{4.1}$)	0.897	0.044	rejected Ho
Middle($x_{4.2}$)	6.276	0.012	rejected Ho
The relationship between students and their teachers(x_5)			
Low($x_{5.1}$)	0.184	0.668	accepted Ho
Middle($x_{5.2}$)	2.009	0.156	accepted Ho
The relationship between students and theirfamily (x_6)			
Low($x_{6.1}$)	0.125	0.723	accepted Ho
Middle($x_{6.2}$)	2.581	0.108	accepted Ho
Self Motivation(x_7)			
Low($x_{7.1}$)	0.00	1.00	accepted Ho
Middle($x_{7.2}$)	0.00	0.997	accepted Ho

Conclusion: Based on Table 2, the equivalent of 10% significance is concluded that the variables X2 and x4 affect the model whereas the variables x1, x3, X5, X6 and X7 do not affect the model. Factors that influence the major interest of statehigh school students Salem I is the science national examination score and students' relationship with their friends.

Furthermore, the establishment of the final model uses an influential variable on the model. The final models obtained are:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

with :

$$g(x) = 1{,}785 - 0.027x_2 + 0{,}72x_{4.1} + 0.783x_{4.2}$$

3.2. Logistic Regression Clasification

By calculating the probability of each observation, then it was obtained the classification result for binary logistic regression methods:

Table 3. Clasification Result

	Predicted	
Observed	Social Major	Science Major
Social Major	48	50
Science Major	9	117
Overal Percentage	73.7 %	

Table 3exposes that there are 48 students majoring in Social that are predicted to enter the Social Major, 50 students of Social Major who are predicted to enter the Science Major, 9 students majoring in Science, which is predicted to study in Social Science and 117 students of Science majoring in Science. Prediction error of 26.3% and accuracy of prediction of 73.7%. It can be deduced accuracy classification with Logistic Regression method of 73.7%.

3.3. K-NN Clasification

This study used 224 data, 80% data used as training data and 20% used as data testing. To determine whether the student enters where to use 3 nearby data so that the value K = 3. To measure the accuracy of the method used Confussion matrix. Table 4 follows the Matrix confussion table for the K-NN method:

Table 4. Confussion Matrix

	Predicted	
Observed	Social Major	Science Major
Social Major	64	34
Science Major	5	121
Overal Percentage	82.6 %	

Table 4 explicates that there are 64 students majoring in Social that are predicted to enter the Social Major, 34 students of Social who are predicted to enter the Science Major, 5 students majoring in Science and is predicted to study in Social Major and 121 students of Science majoring in Science Majors. The error prediction is 17.4% and theaccuracy of prediction is 82.6%. All in all, the accuracy of classification with K-NN method in the School interest data of Salem majors is 82.6%.

4Conclusion

To sum up the analysis that has been done, there is a conclusion that both using binary logistic regression method, the alleged factor to affect the interest in state high school majors is the Science National Examination score and the relationship of the students with their friends. The interest classification system of State High School 1 Salem which resulted from binary logistic regression method has a precision is73.7%. The interest classification system of State High School 1 Salem which resulted from the K-NN method has a accuracy of 82.6%. The best classification method for interest in Salem High School is the K-NN method.The suggestion for this research is the addition of predictor variables that affect the response variable so that it can increase the accuracy value of the model.

References

[1] Kementrian Pendidikan dan Kebudayaan, Modul Pelatihan Implementasi Kurikulum 2013 untuk Guru BK atau Konselor (2013).

[2] Sudiarto, A. Pedoman Arah Peminatan Draf 2 (2013)[online]. http://www.akursudianto.com/wp-content/uploads/2019/06/pedoman-arah-peminatan-draf-2.pdf

[3] Departemen Pendidikan Nasional, Panduan Penilaian Penjurusan Kenaikan Kelas dan Pindah Sekolah. Direktorat Pendidikan Menengah Umum, Jakarta (2004).

[4] Syah, M. Psikologi Pendidikan dengan Pendekatan Baru. Bandung: Remaja Rosdakarya (2009)

[5] Chiang, L. H., dan Pell, R. J. Genetic Algorithms Combined with Discriminant Analysis for Key Variable Identification. Journal of Process Control, 14, 143-155 (2004).

[6] Baby, N & P.L.T. Customer Classification And Prediction Based On Data Mining Technique. International Journal of Emerging Technology and Advanced Engineering, 2(12), 314-18 (2012).

[7] Liao, S. H., Chu, P. H., & Hsiao, P. Y. Data mining techniques and applications - A decade review from 2000 to 2011. Expert Systems with Applications, 39(12), 11303–11311 (2012).

[8] Ferdousy, E.Z., Islam, M.M. & Matin, M.A. Combination of Naive Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. Computer Science and Information Science, 6(3), 48-56 (2013).

[9] Hosmer, D. W. and Lemeshow S. Applied Logistic Regression. United States of American: Sons Inc (2000)

[10] Liu, B. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Berlin: Springer (2007)

[11] Wu X, Kumar V. The Top Ten Algorithms in Data Mining. New York:CRC Press. (2009)

[12] Shahin, A. and Janatyan, N. Estimation of Customer Dissatisfaction Based on Service Quality Gaps by Correlation and Regression Analysis in a Travel Agency. International Journal of Business and Management Vol. 6, No. 3. (2011)

[13] Tanboga IH, Kurt M, Isik T, Kaya A, Ekinci M, Aksakal E, et al. Assessment of multivariate logistic regression analysis in articles published in Turkish cardiology journals. Turk Kardiyol Dern Ars 2012;40:129-34. (2012)

[14] Kumar, et al.: Quality of logistic regression in Indian journals Indian Journal of Public Health, Volume 60, Issue 2 (2016)

[15] Banjarsari, M.A, et. al, Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4 , Jurnal Ilmu Komputer (KLIK) Volume 02, No.02 (2015)

[16] Hidayah, M. R., Aklis, I. & Sugiharti, E. (2017). Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification. Scientific Journal of Informatics, 4(1), 66 – 75 (2017).

[17] Bujlow, T., Riaz, T., & Pedersen, J. M. A method for classification of network traffic based on C5.0 machine learning algorithm.International Conference on Computing, Networking and Communications, ICNC'12, 237–2419 (2012).

[18] Gorunescu, Florin. (2011). Data Mining: Concepts and Techniques. Verlag berlin Heidelberg: Springer (2011).

[19] Y.S. Kim, Comparision of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size,Journal of Expert Systems with Application, Elsevier, pp. 1227-1234 (2008).

[20] Trevino,V. & Falciani,F., GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics, 22, 1154–1156 (2006)