International Journal of Informatics and Information Systems
Vol. 1, No. 1, September 2018, pp. 1-8 ISSN 2579-7069

Application of Data Mining Classification Method for Student Graduation Prediction Using K-Nearest Neighbor (K-NN) Algorithm

Mohammad Imron [a,1,*], Satia Angga Kusumah [a,2]

[a] Informatics Engineering Study Program, STMIK AMIKOM Purwokerto, Central Java, Indonesia
[1] imron@amikompurwokerto.ac.id* , [2]p.duryudana@gmail.com
* corresponding author

Abstract

The student graduation rate is one of the indicators to improve the accreditation of a course. It is needed to monitor and evaluate student graduation tendencies, timely or not. One of them is to predict the graduation rate by utilizing the data mining technique. Data Mining Classification method used is the algorithm K-Nearest Neighbor (K-NN). The data used comes from student data, student value data, and student graduation data for the year 2010-2012 with a total of 2,189 records. The attributes used are gender, school of origin, IP study program Semester 1-6. The results showed that the K-NN method produced a high accuracy of 89.04%.

Keywords: data mining; classification; K-NN algorithm; graduation; timely

1. Introduction

The college is one of the education levels that is considered as the last gate for students to gain knowledge before they finally involve themselves in the workforce. Institutions of higher education should improve the quality of service and satisfy the students as well as the public space around them in order to compete with other colleges. Higher Education Accreditation by BAN-PT (Higher Education National Accreditation Body) is one of the parameters in determining the quality of universities and courses in Indonesia. The student Graduation rate is an indicator of increasing the accreditation of a course. Therefore, it is necessary to monitor and evaluate the student graduation tendencies, on time or not.

Stmik Amikom Purwokerto is one of the higher education institutions in the field of management and computer science. In the Regulation Of The Stmik Amikom Purwokerto Education year 2009 in Chapter I Article 1 paragraph 2 mentioned that the undergraduate program (S1) regular is an academic education program after secondary education, which has a study burden of at least 144 SKS (unit credit Semester) and as many as 160. The duration standard of the undergraduate study program (S1) is 8 semesters, but many students are found to graduate beyond the scheduled ones.

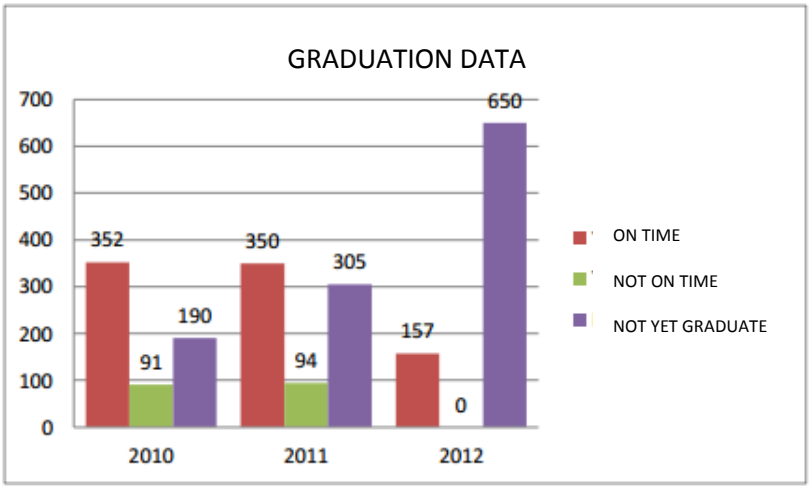

Fig. 1 Student Graduation Chart

From the above data, it can be seen that only about 50% of the total students of each generation graduated on time. It is undoubtedly necessary for monitoring and evaluation of the student graduation rate. Because if left unchecked and the graduation rate will have an impact on the value of accreditation, and can be an obstacle for Stmik Amikom Purwokerto to go to a higher level that is from high school to institute.

Research to be done is to analyze the value of accuracy generated by the classification method K-Nearest neighbor (K-NN) because K-Nearest neighbor is a strict algorithm against the training of data that has much noise. Also, K-Nearest Neighbor is more effective when the data training is significant[1][2].

Based on the explanation above, can be formulated the problem is "how to implement data mining techniques useful as information to know the graduation rate of students in the Stmik Amikom Purwokerto using the algorithm K-Nearest Neighbor (K-NN)." The research aims to know the level of accuracy that has been submitted By the K-Nearest Neighbor (K-NN) algorithm in predicting the graduation rate of students at Stmik Amikom Purwokerto. The applicative benefits that can be obtained from this research are expected to help determine the level of student graduation so that it can be a monitoring and evaluation material to improve the quality of the campus and can be used as a reference in conjunction with the expert system and decision-making system.

2. Research Methods

2.1. Tools and Research Materials

a. Hardware requirements
   1) Laptop
      a. Intel (R) Celeron (R) CPU 877 @ 1.40 GHz
      b. 4GB DDR3 RAM
      c. Hard drive 500Gb
      d. LED Cinecrystal 14 inches
   2) Canon MP287 Printer

b. Software requirements (software)
   1) Operating System : Windows 7 Ultimate
   2) Creation of the Report : Microsoft Office 2010
   3) Creation of datasets : Notepad+ + and Microsoft Excel 2010
   4) Data Mining : Waikato Environment Knowledge Analysis (WEKA)

c. Research Materials

The materials used in this study were obtained from student data, student Value data, and graduation data of students of Stmik Amikom Purwokerto year 2010 to 2012.

2.2. Research Concept

This framework is the steps that will be taken in resolving the issue that will be addressed. The framework in this study looks like in figure 2 below:

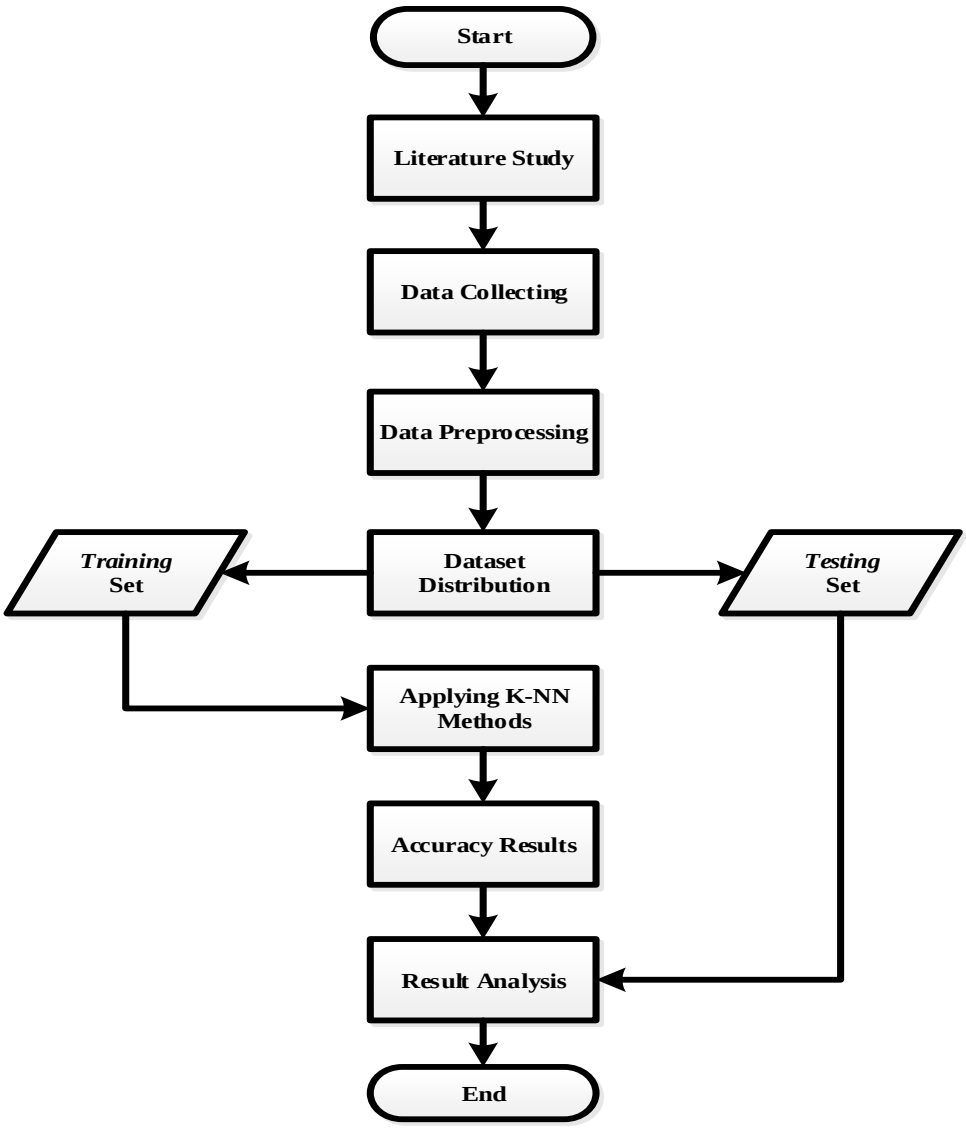

Fig. 2 Research Frameworks

The following describes the Research framework shown in Figure 2 above:

a. Literature study

In this study, the literary studies of the authors conducted a review of the journals or other relevant sources of the library relating to this study, then the source of the library was used as a reference for Support in the research process.

b. Data Collection

At the stage of the author's data collection using student graduation data in 2010-2012 to be used as research data. Student Data obtained from 2010 to 2012 as many as 2.189 Students.

c. Preprocessing Data

In the preprocessing phase of the data, researchers perform several processes to obtain a clean dataset of missing values and inconsistent data so that the dataset can be recognized and managed using WEKA.

d. Dataset distribution

Once data is obtained, the process of dividing the dataset and subsequent data will be divided into two, namely training data and data testing.

e. Application of the K-Nearest Neighbor (K-NN) method

Data that has been divided into two then done process of counting using method K-Nearest Real (K-NN), the application that is used in the calculation process is Waikato Environment Knowledge Analysis (WEKA)[13].

f. Accuracy results

From the process of implementing the K-Nearest Neighbor (K-NN) method that has been done before, obtained the result of accuracy that then from the results of accuracy will be analyzed whether the algorithm used to provide the best accuracy results or not in Classification of student graduation data on time and promptly.

g. Results analysis

The results of the analysis phase will conclude from the implementation of the K-Nearest Neighbor (K-NN) Classification method of the student graduation data. Then the accuracy results will be compared with the accuracy results of the implementation of the Decision Tree classification method that has been analyzed in previous research[14]1[15.

3. Results and Discussion

3.1. Literature Study

In the literary study phase, researchers conducted a review of the relevant journals, books, and other library resources related to the study. Subsequently, reviewed library sources are used as a reference to support the research process.

3.2. Data Collection

Researchers collected the data needed in this study. Data used from Stmik Amikom Purwokerto, which is data of student graduation from the year 2010–2012 obtained from Baak Stmik Amikom Purwokerto, student Value data of the year 2010 – 2012 from semester 1 – 6 derived from the Baak Stmik Amikom Purwokerto, and student Data of the year 2010 – 2012 obtained from the Front Office of Stmik Amikom Purwokerto.

3.3. Data Preprocessing

a. Data cleanup and integration

The data cleanup process is done so that the data obtained is relevant data according to the needs and because not all the attributes in the table will be used[4]. A critical data cleanup is done to improve performance in the mining process. The data cleanup can be done by deleting the incomplete data and deleting the unused attributes[5].

b. Data transformation

In the Data transformation phase, changes in the type of gender, course, and IP attributes are changed. Once the data transformation process is done, the last step of preprocessing data is to convert the dataset from an Excel file to CSV or ARFF format in order to be recognized as a data source on WEKA[6]. However, before being saved as an ARFF file format, it is known that the existing dataset is still the original dataset that still mixed so that the dataset will need to be 2, i.e. data that will be used as the data training and data to be used as data testing[7].

3.4. Dataset Distributions

Datasets that have been created after going through several stages of preprocessing data amounted to 1,746 Records. From 1,746 this record, the label class with a timely and in timely value amounted to 667 records consisting of students of the Generation year 2010 and 2011, and students who have not graduated in the 1,079 records consisting of students in class years 2010 up to 2012.

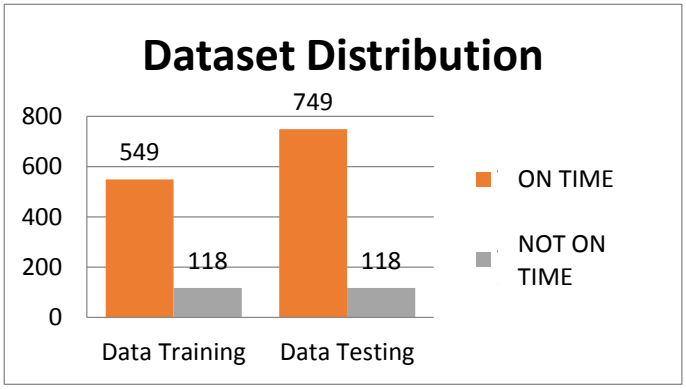

Fig. 3. Dataset Distribution Chart

3.5. Application of K-NN Method

After the distribution of the dataset is done, process mining or testing dataset using the method or algorithm that has been determined and in this study used the algorithm K-Nearest Neighbor (K-NN)[8][9]. The mining process is done by implementing the K-Nearest Neighbor (K-NN) method and using the intelligent system provided by the WEKA tool. Tests will be conducted on each dataset that has been shared in the previous discussion[10][11].

a. K-Fold Cross-Validation

This process is a process to get the K-optimal using the method K-Fold Cross-Validation. The Fold used is 10. Then The value K to be used is k = 1.

b. Accuracy Test

1) Testing Data Training

Testing of the training data resulted in the highest accuracy in test k = 1, which is 88.16%, the meaning of the 667 records known that as many as 588 records were predicted correctly and the 79 records were mispredicted.

Table 1. Classifier Output from Data Training

No.	Specifications	Value
1	*Instances*	667
2	*Attributes*	10
3	*Test mode*	*Evaluate training data*
4	*Time is taken to test model*	0.11 *seconds*
5	*Correctly Classified Instances*	588 (88.16%)
6	*Incorrectly Classified Instances*	79 (11.84%)
7	*Kappa statistic*	0.4629
8	*Mean Absolute Error*	0.1594
9	*Root mean squared error*	0.2818
10	*Relative Absolute Error*	54.62%
11	*Root Relative squared error*	73.86%
12	*Total Number of instances*	667
13	*Confusion Matrix*	A B ← *classified as* 546 3 = Time 76 42 = No
14	Value k =	1
Detailed Accuracy By class:		

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.644	0.878	0.995	0.933	0.533	0.919	0.980	TepatWaktu
	0.356	0.005	0.933	0.356	0.515	0.533	0.919	0.727	Tidak
Weighted Avg.	0.882	0.531	0.888	0.882	0.859	0.533	0.919	0.935	

2) Testing Data Testing

Testing of data testing resulted in an accuracy of 89.04%, which means that from 867 Records It is known that as many as 772 records are predicted to be correct and 95 records are predicted to be incorrect.

Table 2. Classifier Output from Data Testing

No.	Specifications	Value
1	*Instances*	667
2	*Attributes*	10
3	*Test mode*	*The used supplied Test set*
4	*Time is taken to test model*	0.14 *seconds*
5	*Correctly Classified Instances*	772 (89.04%)
6	*Incorrectly Classified Instances*	95 (10.96%)
7	*Kappa statistic*	0.415
8	*Mean Absolute Error*	0.176
9	*Root mean squared error*	0.2998
10	*Relative Absolute Error*	66.27%
11	*Root Relative squared error*	86.78%
12	*Total Number of Instances*	867
13	*Confusion Matrix*	A B ← *classified as* 730 19 = Time 76 42 = No
14	Value k =	1
Detailed Accuracy By class:		

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.975	0.644	0.906	0.975	0.939	0.443	0.884	0.979	TepatWaktu
	0.356	0.025	0.689	0.356	0.469	0.443	0.884	0.544	Tidak
Weighted Avg.	0.890	0.560	0.876	0.890	0.875	0.443	0.884	0.919	

c. K-NN algorithm

The steps for calculating the K-Nearest Neighbor (K-NN) method are as follows:

1) Specifying the K parameter
2) Calculating the distance between the data to be evaluated with all training
3) Sorting the distance formed
4) Determine the closest distance to the order K
5) Pairing a compatible class
6) Look up the number of classes from the closest neighbor and set the class as the data class to be evaluated.

Formula → $d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$

Description:

X_1 = Sample Data

X_2 = Test data or *Data Testing*

i = Data variables

D = Distance

p = Data Dimension

Examples of the manual calculation process are as follows:

$$D_1 = \sqrt{\begin{matrix}(3-2)^2 + (2-3)^2 + (3-2)^2 \\ +(2-2)^2 + (2-2)^2 + (2-2)^2\end{matrix}} = 1{,}73$$

Calculations are done by calculating the distance Euclidean distance from each training data to the testing data. After calculating Euclidean distance, then specify K classification and take the largest value based on the specified K is 1. Furthermore, the results of classification k are determined by order of the ranking and retrieved 1 data that has the highest Euclidean value[12].

From the overall calculation of the distance between the training data amounting to 667 data, the calculation results are given a range and sorted from the smallest to the largest. Once sorted, then seen the classification result arising from the calculation of the highest distance or follow the value of the first rank.

From the results of manual calculations that researchers have done generated data with a "timely" label as many as 772 records whereas with the label "No" amounted to 95 records. This means that the resulting accuracy can be calculated as follows:

Accuracy percentage = (Total correct prediction/Total data) x 100%

= (772/867) x 100%

= 89.04%

3.6. Accuracy Results

From the test results of the DataSet with 3 attempts that have been done in the previous discussion, obtained the highest accuracy result as follows:

Table 3. Data Accuracy Test Results Training & Data Testing

Type of Data	**Amount of Data**	**Method**	**Accuracy**
Data Training	667 *Record*	Lazy – IBK	88.16%
Data Testing	867 *Record*	Lazy – IBK	89.04%

3.7. Result Analysis

Research that has been done using the graduation dataset consisting of data training and data testing, by implementing the K-optimal method on the 2 types of datasets resulted in the acquisition of accuracy value in the training data of 88.16 % in tests with a value of k = 1 while testing data testing with a value of k = 1 resulted in an accuracy of 89.04%.

From the research that has been done by the authors, it can also be analyzed that the value of k affects the accuracy value of each dataset test. This is evidenced by the value of accuracy generated in test tests to 1 to 3, where each experiment used different k values of 1, 3, and 5. In test testing of data, training resulted in the highest accuracy value in the implementation of value K = 1 which is 88.16% while the lowest accuracy value is generated on the test with a value of k = 5, then it can be concluded that the accuracy value Highest possible implementation of the smallest K value.

4. Conclusions and Suggestions

4.1. Conclusion

a. From the total combined original data amounted to 2,189 Records, then done cleaning and data Integration obtained as many as 1,736 Records. From 1,736 Records divided into 2 datasets that are data training and data testing, data training amounted to 667 Records from students from 2011-2011 who have passed both on time and not time, and data testing amounted to 867 Records from 2010-2012 students who have not graduated.

b. From Testing conducted by the author of the training data with the amount of 667 Records obtained the highest accuracy of 88.16% in the 1st test with a value of k = 1, meaning of 667 Records as many as 588 Records is the total predictions is correct. Then D to test that has been done by the authors against the data testing with the amount of data 867 record generates an accuracy of 89.04%, meaning of 867 records as much as 772 record is the total prediction right.

4.2. Suggestions

a. Try using Other algorithms like Naive Bayes, Neural Network, or any other.

b. Try to make a comparison with more algorithms in order to be generated the best level of accuracy that can later be used as a reference to developing a system that can be used to predict the graduation rate of students at Stmik Amikom Purwokerto.

c. Try many more attributes and records in data mining processing.

d. It takes a high level of precision and a perfect data cleanup so that no noise occurs.

References

[1] F. Gorunescu, "Data Mining Concept Model and Techniques". Berlin:Springer, 2011.

[2] Akbarinia, R., Pacitti, E., & Valduriez, P. (2007). Processing top-k queries in distributed hash tables. InEuro-Par(pp. 489–502).

[3] Balke, W. -T., Nejdl, W., Siberski, W., & Thaden, U. (2005). Progressive distributed top-kretrieval in peer-to-peer networks. InICDE.

[4] Bao, J., Zheng, Y., & Mokbel, M .F. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. InProceedings of the 20th international conference on advances in geographic information systems (SIGSPATIAL), 2012.

[5] Bast, H., Majumdar, D., Schenkel, R., Theobald, M., & Weikum, G. (2006). Io-top-k:index-access optimized top-k query processing. In VLDB(pp. 475–486).

[6] Chang, Y.-J., Liu, H.-H., Chou, L.-D., Chen, Y.-W.,& Shin, H.-Y. (2006). A general architecture of mobile social network services. In7th International conference on mobile data management (MDM).

[7] Chen, L., Zeng, W., & Yuan, Q. (2013). A unified framework for recommending items groups and friends in social media environment via mutual resource fusion. Expert Systems with Applications, 40(8), 2889–2903.

[8] Cao, L., & Krumm, J. (2009). From GPS traces to a routable road map. InProceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems.

[9] Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines.ACM Transactions on Intelligent Systems and Technology, 2(3), 1–27.

[10] Condie, T., Conway, N., Alvaro, P., Hellerstein, J.M., Elmeleegy, K., & Sears, R. (2010). Mapreduce online. InNSDI(pp. 313–328).

[11] Dean, J., & Ghemawat, S. (2010). Mapreduce: A flexible data processing tool. Communication of the ACM, 53(1), 72–77.

[12] Dong, Z. -B., Song, G. -J., Xie, K. -Q. & Wang, J. -Y. (2009). An experimental study of large-scale mobile social network. In Proceedings of the 18th international conference on world wide web (WWW '09)(pp. 1175–1176).

[13] Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011). Crowddb: answering queries with crowdsourcing. In SIGMOD conference(pp. 61–72).

[14] Getoor, L., & Diehl, C. P. (2005). Link mining: A survey.SIGKDD Explorer Newsletter, 7(2), 3–12.

[15] Ilyas, I. F., Beskales, G., & Soliman, M. A. (2008). A survey of top-k query processing techniques in relational database systems. Computing Survey, 40(4), 11:1–11:58.