Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbors for Predicting Thesis Graduation

Achmad Solichin
Faculty of Information Technology
Universitas Budi Luhur
Jakarta, Indonesia
achmad.solichin@budiluhur.ac.id

***Abstract*— Thesis is one of the evaluations of learning for students. In Universitas Budi Luhur (UBL), especially in the Informatics Department, the thesis is one of the requirements for graduating students to obtain a Bachelor of Computer degree. In each semester, the number of Informatics Department students who take thesis is around 200-300 students. The problem that is still faced is that student graduation in the thesis is not optimal. Student failures in the thesis are allegedly related to several technical and non-technical factors. In this study, an analysis using data mining algorithms was carried out to determine the factors that influence student graduation in the thesis. The dataset obtained from the Informatics Department students who took a thesis in the 2016/2017, and 2017/2018. In order to obtain the right classification method, this research was tested with three classification methods, namely Decision Tree, Naïve Bayes, and k-Nearest Neighbors (kNN). The results of the comparison of the values of accuracy, precision, and recall indicate that the kNN algorithm has advantages, so this method is chosen to predict graduation. In this study also developed an application for predicting graduation of students' thesis by applying the kNN classification method. The test results showed an accuracy of 78.20%, precision of 80.32%, and recall of 96.49%. This research is expected to be useful for improving the service quality of student thesis.**

Keywords*— *data mining, student thesis, kNN, Naïve Bayes, decision tree, comparison

I. INTRODUCTION

The Informatics Department is part of the Faculty of Information Technology, Budi Luhur University, which is quite large with a large number of students. In the 2017/2018 academic year, the number of active students in Informatics Department is 1,288 students. The thesis is one of the evaluations of learning for students. In the Budi Luhur University, especially in the Informatics Department, the thesis is one of the requirements for graduating students to obtain a Bachelor of Computer degree. The thesis has a weight of 6 credits and is done structured for one semester. A supervisor chosen by the student assists every student who takes the thesis.

The process of managing the thesis at the Informatics Department is currently utilizing various information systems so that all thesis data has been appropriately stored in the database. The data should be used to evaluate and make decisions for managers of Departments and Faculties. However, until now, the transactional data has not been utilized properly.

Therefore, in this study, the proposed analysis model of the thesis graduation rate uses data mining algorithms. This study also aims to determine the factors that influence student graduation in the thesis. Data mining is a technique for tracking existing data to build a model, then using that model to recognize other data patterns that are not in the stored database [1]. One of the data mining techniques is the classification technique. Classification is a process of discovering models or functions that describe and distinguish data classes or concepts that aim to be used to predict classes of objects whose label class is unknown [2]. Data classification consists of two steps process, first is learning (training phase), where classification algorithms are made to analyze training data and then represent in the form of classification rules. The second process is classification, where test data is used to estimate the accuracy of the classification rule [2].

Several previous studies, especially those relating to the prediction of graduation in the academic field, have been carried out. Research by [3] predicts the possibility of new students completing their studies on time using data mining analysis. The study took a case study at STMIK Dipanegara Makassar. The algorithm used is the k-Nearest Neighbor (kNN). The attributes used are national exam scores, gender, religion, department, and province. The data used were 541 from alumni data with the distribution of classes from 2004 to 2010. The design of the application uses the PHP 5.0 programming language and MySQL database. The results of these studies are known to be a close relationship between the new case and the case that already exists in the data warehouse so that it can be a reference for predicting the graduation of a new student whether they can finish their studies on time or not [3].

Research by Kamagi and Hansun implemented the C4.5 algorithm to predict whether students can graduate on time, fast, late, or drop out of school. The attributes used to make predictions are grade point, gender, school origin, type of graduation, and a number of credits passed. The training data used were alumni data for the 2007 and 2008 classes. As for the research data, alumni data were used in 2009 with 100 data records. The study used Microsoft Visual Studio 2012 and Microsoft Excel 2013 to develop an application. The results of the study successfully predicted students with an accuracy of 87.5% [4]. Similar research was also carried out by Ridwan et al., but using the Naïve Bayes algorithm [5]. The research aims to evaluate the academic performance of students in the second year and classify them of students who have the potential to graduate on time or not. The attributes used in the study were gender, school origin, entry point, national exam value, parental salary, and GPA from the first four semesters. The sample data used is alumni data from 2005-2009, with a

total of 100 samples. The test results show the accuracy of the Naïve Bayes classification algorithm of 83%.

Meanwhile, Defiyanti has researched to analyze and predict student learning performance based on demographic variables, academic data, and student economics [6]. The study compared three data mining algorithms, namely decision tree, naïve Bayes, and artificial neural network (ANN). The attributes used are age, gender, ethnicity, school origin, number of credits taken, GPA, tuition fees, student status, parental income, personal income, and residence. The results of the study indicate that the algorithm that has the best accuracy is Naïve Bayes. In terms of speed, the Naïve Bayes algorithm also has the fastest time compared to the other two algorithms.

TABLE I. SOME OF THE CLASSIFICATION ALGORITHM IN VARIOUS STUDIES ACADEMIC RESEARCH AREA

#	Algorithm	Used in
1	Decision Tree / C4.5	[4], [6]–[11]
2	kNN	[3], [6], [11]–[13]
3	Naïve Bayes	[5], [9], [11], [14]–[18]
4	ANN	[6], [9]
5	J48	[8]
6	ID3	[11], [19]
7	SVM	[20]
8	Logistic Regression	[9]
9	Apriori	[21]

Table 1 shows the classification algorithms in various studies on the academic research area. In this study, we compared three classification algorithms that were most widely used by previous researchers (See Table 1), as well as the ten best algorithms according to [22]. The classification algorithms compared are Decision Tree, Naïve Bayes, and K-Nearest Neighbors. The dataset used in this study was obtained from informatics engineering students who took the thesis in the 2016/2017 school year and 2017/2018. In addition to making comparisons, the best algorithms are implemented in a prototype.

II. METHODOLOGY

A. Research Framework

In conducting research, structured steps are needed so that the research objectives can be adequately achieved. The stages of the research become a reference in conducting research. Figure 1 shows our research framework. Based on the research problems described in the previous section, we studied various previous studies related to the application of classification algorithms in the academic field. After conducting a literature study, based on various studies using various classification algorithms, as summarized in Table 1, this study compared three classification algorithms. The best classification algorithm is determined based on testing with measures of performance accuracy, precision, and recall. Furthermore, the selected algorithm is implemented in a prototype to predict student thesis graduation. The performance of the classification algorithm is again tested using the confusion matrix.

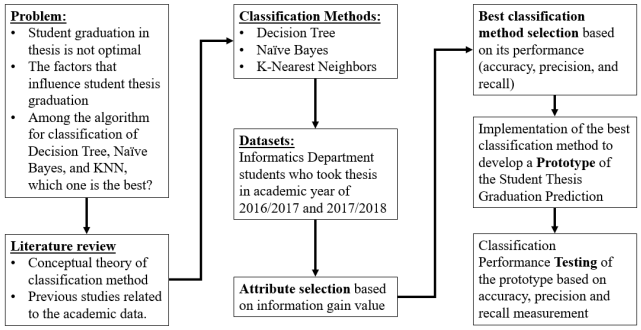

Fig. 1. Research Framework

B. Data Acquisition

In this study, data were obtained directly from the IT department of UBL. Data was taken based on data of students taking thesis in the 2016/2017 odd semester, even semester 2016/2017, odd semester 2017/2018, and even semester 2017/2018 in the Informatics Engineering Study Program.

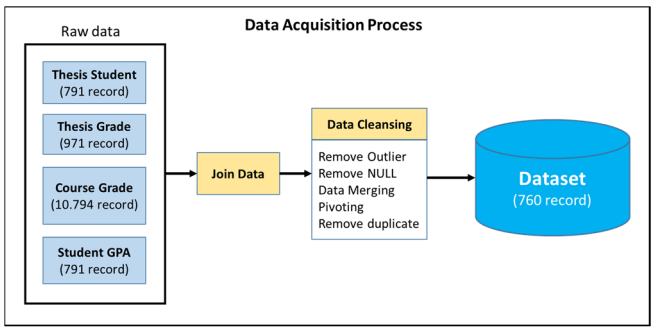

Fig. 2. Data acquisition process

Figure 1 presents the research data acquisition process to obtain the research dataset. Some source data drawn from the UBL database are (1) a list of thesis students, (2) a list of thesis grades, (3) grades courses, and (4) student GPA. All source data is combined into one dataset; then the data cleaning process is carried out. The data cleaning process is needed to ensure the dataset is by the data format needed.

The data cleaning process includes the following processes:

1. *Remove outlier*, to eliminate data that has a value that is wrong and out of reach of reasonable values.
2. *Remove NULL*, to delete empty records or fields.
3. *Data merging*, to combine multiple records with data that has the same meaning. For example, because students are allowed to take courses several times, then the value of a student can be stored in several records. The data must be combined so that a single value data is obtained for each course.
4. *Pivoting*, to change the dimensions of the course value data from row to column.
5. *Remove duplicate*, to eliminate multiple data that has the same value and meaning.

The data acquisition process produced a research dataset of 760 records. Table 2 displays descriptive statistics of the research dataset for each semester. Of all 760 records, 616 students passed the thesis (81%), and 144 students failed (19%). Meanwhile, Figure 2 shows the distribution of student thesis grades grouped by grade A, B, C, D, and F.

TABLE II. DESCRIPTIVE STATISTICS OF THE DATASET

#	Academic Year	Semester	Num of Students	Passed	Failed
1	20162017	Odd	155	135	20
2	20162017	Even	229	173	56
3	20172018	Odd	198	153	45
4	20172018	Even	178	155	23
		TOTAL	**760**	**616**	**144**

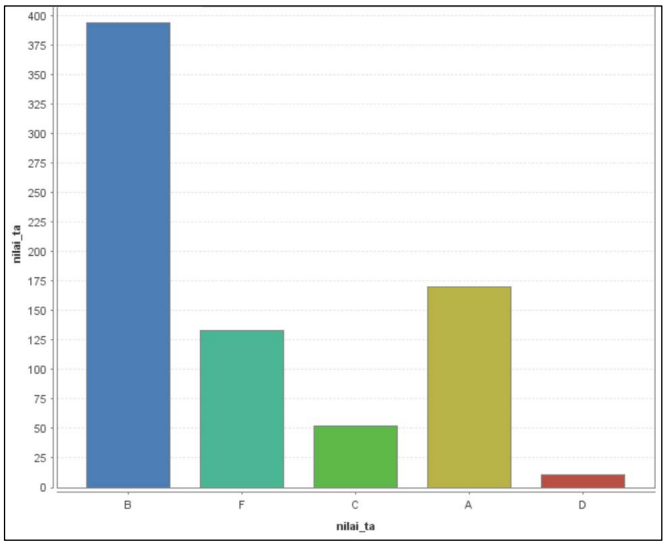

Fig. 3. The distribution of student thesis grades

Meanwhile, Figure 3 presents the distribution of student grade point average (GPA). From the graph, it can be seen that most students have a GPA in the range of 3.0 to 3.4 on a scale of 0-4.

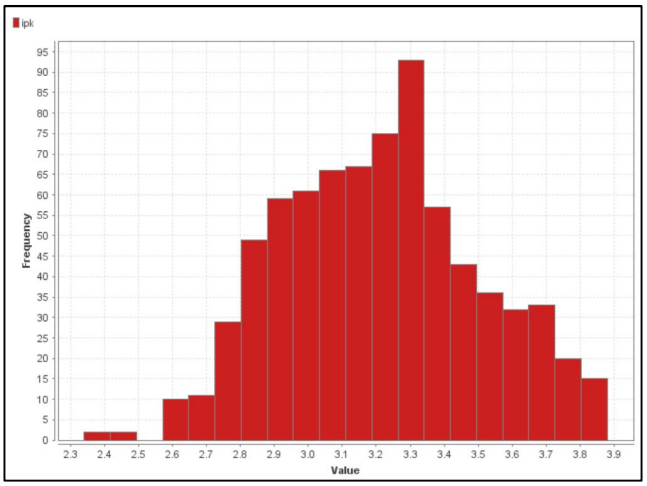

Fig. 4. The distribution of student GPA

C. *Attribute Selection*

In this study, the selection of graduation and grade assignments was made to simplify the number of attributes and eliminate attributes that did not have a significant effect. Table 3 lists the initial attributes of the study, which are thought to have a significant influence on thesis graduation. Attribute selection is based on the information gain value of each attribute. Information gain [23] is calculated statistically based on the entropy of each attribute to the target class entropy. The formula of the information gain is shown at (1).

$$G(D,t) = -\sum_{i=1}^{m} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{m} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{m} P(C_i|\bar{t})\log P(C_i|\bar{t}) \quad (1)$$

Where is C is a set of data, in which there is not the feature t. The value of G(D,t) is higher; t is more useful for the classification for C. This t should be selected. If the higher value of G(D,t) is wanted, it should make the value of P(t) and P(t) smaller.

TABLE III. THE ATTRIBUTE

#	Course ID	Attribute	Range of Values
1	KP002	Algorithm & Data Structure 1	A, B, C, D, E, F
2	KP003	Algorithm & Data Structure 2	A, B, C, D, E, F
3	KP041	Computer Networks	A, B, C, D, E, F
4	MI028	Calculus 1	A, B, C, D, E, F
5	KP066	Data Communication	A, B, C, D, E, F
6	TA001	Internship	A, B, C, D, E, F
7	IF015	Software Project Management	A, B, C, D, E, F
8	KP229	Computer Organization	A, B, C, D, E, F
9	PG061	Object-Oriented Programming	A, B, C, D, E, F
10	PG117	Web Programming	A, B, C, D, E, F
11	KP342	Software Engineering 1	A, B, C, D, E, F
12	KP164	Operating System	A, B, C, D, E, F
13	KP181	Theory of Language & Automata	A, B, C, D, E, F
14	UM031	Budi Luhur Concepts	A, B, C, D, E, F
15	GPA	Grade Point Average	0,0 – 4.0
16	Remedial	The number of courses repeats	0 – 8
17	Thesis Grade	Thesis grades [LABEL]	A, B, C, D, F
18	Graduation	Thesis graduation [LABEL2]	LULUS, GAGAL

Table 4 displays a list of research attributes to analyze the graduation of student thesis assignments, which are sorted by information gain value. The higher the information gain value, the more the attribute affects the determination of graduation of the student's final assignment. If seen from the list, it turns out that the GPA attribute has the highest information gain value. Thus, the GPA attribute is the attribute that most influences the graduation of Final Project students. After the GPA attribute, sequentially are the attributes of the value of Object-Oriented Programming (PG061), Number of courses that repeat (remedial), and the Theory of Language & Automata (KP181). Meanwhile, the attribute that has the lowest information gain value is the value of the Computer Organization (KP229) course.

TABLE IV. THE VALUE OF INFORMATION GAIN FOR ALL RESEARCH ATTRIBUTES

#	Course ID	Attribute	Gain
1	GPA	Grade Point Average	1,000
2	PG061	Object Oriented Programming	0,567
3	remedial	The number of courses repeats	0,417
4	KP181	Theory of Language & Automata	0,390
5	KP342	Software Engineering 1	0,366
6	PG117	Web Programming	0,329
7	TA001	Internship	0,290
8	UM031	Budi Luhur Concepts	0,131
9	IF015	Software Project Management	0,129
10	KP041	Computer Networks	0,100
11	KP002	Algorithm & Data Structure 1	0,078
12	MI028	Calculus 1	0,067
13	KP164	Operating System	0,062
14	KP003	Algorithm & Data Structure 2	0,048
15	KP066	Data Communication	0,020
16	KP229	Computer Organization	0,000

III. RESULT AND DISCUSSIONS

In this section the classification test results are explained using algorithms Decision Tree, Naïve Bayes, and k-Nearest Neghbor. Furthermore, it is also presented the results of the comparison of the three algorithms. Finally, it explains the implementation of the selected classification method in a prototype. In addition, the test results of the selected classification algorithm are explained.

A. Classification Testing with Decision Tree

Based on selected attributes and information gain values in Table 4, a decision tree or decision tree can be arranged, as shown in Figure 5. The attribute with the highest info gain value is GPA, which is the root node of the decision tree.

```
ipk > 2.405
|   ipk > 2.755: LULUS {LULUS=601, GAGAL=127}
|   ipk ≤ 2.755
|   |   ipk > 2.585
|   |   |   KP003 = A: GAGAL {LULUS=0, GAGAL=1}
|   |   |   KP003 = B: GAGAL {LULUS=0, GAGAL=4}
|   |   |   KP003 = C
|   |   |   |   mengulang > 6: GAGAL {LULUS=0, GAGAL=1}
|   |   |   |   mengulang ≤ 6
|   |   |   |   |   ipk > 2.745: GAGAL {LULUS=0, GAGAL=1}
|   |   |   |   |   ipk ≤ 2.745
|   |   |   |   |   |   KP066 = A
|   |   |   |   |   |   |   ipk > 2.710: GAGAL {LULUS=0, GAGAL=3}
|   |   |   |   |   |   |   ipk ≤ 2.710: LULUS {LULUS=1, GAGAL=0}
|   |   |   |   |   |   KP066 = B: LULUS {LULUS=1, GAGAL=0}
|   |   |   |   |   |   KP066 = C
|   |   |   |   |   |   |   ipk > 2.715: LULUS {LULUS=4, GAGAL=0}
|   |   |   |   |   |   |   ipk ≤ 2.715
|   |   |   |   |   |   |   |   MI028 = B: LULUS {LULUS=2, GAGAL=0}
|   |   |   |   |   |   |   |   MI028 = C
|   |   |   |   |   |   |   |   |   PG061 = B: LULUS {LULUS=1, GAGAL=0}
|   |   |   |   |   |   |   |   |   PG061 = C
|   |   |   |   |   |   |   |   |   |   PG117 = B: GAGAL {LULUS=0, GAGAL=4}
|   |   |   |   |   |   |   |   |   |   PG117 = C
|   |   |   |   |   |   |   |   |   |   |   KP002 = B: GAGAL {LULUS=0, GAGAL=1}
|   |   |   |   |   |   |   |   |   |   |   KP002 = C: LULUS {LULUS=1, GAGAL=0}
|   |   ipk ≤ 2.585: LULUS {LULUS=5, GAGAL=0}
ipk ≤ 2.405: GAGAL {LULUS=0, GAGAL=2}
```

Fig. 5. Textual Decision Tree

Furthermore, based on the decision tree formed, testing the value of accuracy with the research dataset. The test results showed an accuracy of 75%, precision of 82.27%, and recall of 88.15%. Table 5 presents the confusion matrix from the results of testing the classification with the Decision Tree.

TABLE V. CONFUSION MATRIX OF THE CLASSIFICATION TESTING WITH DECISION TREE

	true PASSED	true FAILED	class precision
pred. PASSED	543	117	**82.27%**
pred. FAILED	73	27	**27.00%**
class recall	**88.15%**	**18.75%**	

Furthermore, based on the decision tree formed, testing the value of accuracy with the research dataset. The test results showed an accuracy of 75%, precision of 82.27%, and recall of 88.15%. Table 5 presents the confusion matrix from the results of testing the classification with the Decision Tree.

B. Classification Testing with Naïve Bayes

Figure 6 is a testing model for the classification of the Naïve Bayes method modeled on Rapidminer. Moreover, Table 6 presents the results of testing the Naïve Bayes classification method using a dataset that has been obtained. The test results produced an accuracy value of 73.16%, precision of 83.01%, and recall of 84.09%.

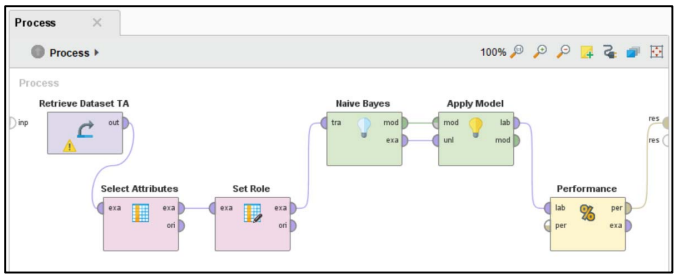

Fig. 6. Classification Model with Naïve Bayes

TABLE VI. CONFUSION MATRIX OF THE CLASSIFICATION TESTING USING NAÏVE BAYES WITH K-FOLD VALIDATION (K = 10)

	true LULUS	true GAGAL	class precision
pred. LULUS	518	106	**83.01%**
pred. GAGAL	98	38	**27.94%**
class recall	**84.09%**	**26.39%**	

C. Classification Testing with kNN (k-Nearest Neighbor)

Furthermore, a final assignment test for students is done with the k-Nearest Neighbor classification method modeled on Rapidminer. Table 7 presents the results of testing the Naïve Bayes classification method using a dataset that has been obtained. With a value of k = 5, the test produces an accuracy value of 80.39%, precision of 81.43%, and recall of 98.21%.

TABLE VII. CONFUSION MATRIX OF THE CLASSIFICATION TESTING USING KNN WITH K-FOLD VALIDATION (K = 10)

	true LULUS	true GAGAL	class precision
pred. LULUS	605	138	**81.43%**
pred. GAGAL	11	6	**35.29%**
class recall	**98.21%**	**4.17%**	

D. Comparison of the Classification Method

Based on the tests that have been conducted on the dataset to determine student graduation, the accuracy of each

classification method is obtained. Table 8 presents a comparison of the value of accuracy between the Decision Tree, Naïve Bayes, and k-Nearest Neighbor classification methods.

TABLE VIII. METHOD OF THE CLASSIFICATION METHOD USING DECISION TREE, NAIVE BAYES, AND KNN

Measure	Decision Tree	Naïve Bayes	k-Nearest Neighbor
Accuracy	75,00%	73,16%	**80,39%**
Precision	82,27%	**83,01%**	81,43%
Recall	88,15%	84,09%	**98,21%**

E. Implementation of the KNN Method

Based on the results of the comparison of the Decision Tree, Naïve Bayes and kNN classification methods as presented in Table 8, it can be concluded that the kNN method has the highest value of accuracy and recall. Therefore, this method is implemented in the application to determine the graduation of final-year students in the Informatics Engineering study program.

Figure 7 presents the main view of the prediction system for graduating the final assignment of Informatics Engineering students developed using the kNN algorithm. To make predictions, the user is asked to enter the grade of the course, how many times to repeat the course, and also the grade point average (GPA) of the student. Then asked to include the dataset as training, and the value of k used. The value of k must be of an odd value so that it can present the results of the prediction correctly. After the user presses the "PREDICTION" button, then the system will display the prediction results on the right of the screen.

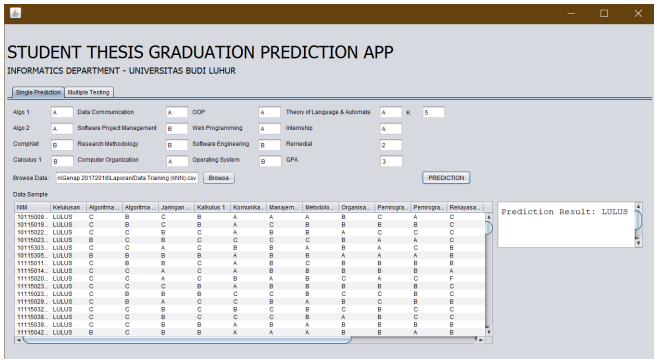

Fig. 7. Thesis Graduation Prediction System of Informatics Engineering Students with KNN Algorithm

TABLE IX. RESULTS OF ACCURACY, PRECISION, AND RECALL TESTING ON GRADUATION PREDICTION APPLICATIONS WITH KNN ALGORITHM

k	Accuracy (%)	Precision (%)	Recall (%)
3	75.00	80.88	90.16
5	77.63	80.56	95.08
7	78.29	80.27	96.72
9	79.61	80.54	98.36
11	78.95	80.00	98.36
13	78.95	80.00	98.36
15	78.95	80.00	98.36
Mean	**78.20**	**80.32**	**96.49**

Based on the results of testing on the prediction application of student graduation applying the kNN classification algorithm, the values of accuracy, precision, and recall were obtained as presented in Table 9. The test results showed that the average value of accuracy was 78.20%, precision 80.32%, and recall 96.49%. These results are quite good but can still be improved in future studies.

IV. CONCLUSIONS

This study compared three classification algorithms applied to predict student thesis graduation. Some conclusions can be drawn from this study. Based on the gain values of each attribute included in this study, it was concluded that several attributes sequentially affect the graduation of students' thesis are: Grade Point Average, Grade of Object Oriented Programming course, Number of repeating subjects, Grade of the Theory of Language and Automation, Software Engineering 1, Web Programming, Internship, Budi Luhur Concepts, Software Project Management, Computer Networks, Algorithms & Data Structures 1, Calculus 1, Operating Systems, Algorithms & Data Structures 2, Data Communications, and Computer Organizations.

Based on the tests that have been done, it shows that the KNN algorithm has a better level of accuracy, precision, and recall than another two algorithms. The kNN classification algorithm produces an accuracy of 80.39%, precision of 81.43%, and recall of 98.21%. This study has succeeded in implementing the KNN data mining algorithm to predict graduation of a student thesis. An application developed using the Java programming language. The test results showed an average accuracy value of 78.20%, precision of 80.32%, and recall of 96.49%.

ACKNOWLEDGMENT

We would like to thank Universitas Budi Luhur for supporting this research. This research was also funded by Universitas Budi Luhur through contract number A/UBL/DRPM/000/092/05/18.

REFERENCES

[1] E. Prasetyo, *Data Mining – Konsep dan Aplikasi Menggunakan MATLAB*, 1st ed. Yogyakarta: Andi Offset, 2012.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., vol. 1. Morgan Kaufmann Publishers, 2012.

[3] M. S. Mustafa and I. W. Simpen, "Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus: Data Akademik Mahasiswa STMIK Dipanegara Makassar)," *Creat. Inf. Technol. J. (CITEC Journal)*, vol. 1, no. 4, pp. 1–10, 2014.

[4] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *Ultim. Vol. VI, No. 1 | Juni 2014*, vol. VI, no. 1, pp. 15–20, 2014.

[5] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013.

[6] S. Defiyanti, "Perbandingan : Prediksi Prestasi Belajar Mahasiswa Menggunakan Teknik Data Mining (Study Kasus Fasilkom Unsikas)," *Konfrensi Nas. Sist. Inf.*, no. March 2014, p. 11, 2014.

[7] M. Ridwan, "Sistem Rekomendasi Proses Kelulusan Mahasiswa Berbasis Algoritma Klasifikasi C4.5," *J. Ilm. Inform.*, vol. 2, no. 1, pp. 105–111, 2017.

[8] Andri, Y. N. Kunang, and S. Murniati, "Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang," in *Seminar Nasional Informatika 2013 (semnasIF 2013)*, 2013, pp. 56–63.

[9] K. Hastuti, "Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif," in *Seminar Nasional Teknologi Informasi & Komunikasi Terapan*, 2012, pp. 241–249.

[10] R. T. Shita and N. Marliani, "Aplikasi Data Mining Dengan Metode Classification Berbasis Algoritma C4.5," in *Seminar Nasional Sistem Informasi Indonesia*, 2013, pp. 517–521.

[11] D. Gustian, A. F. Rahmawati, Titin, R. R. Putra, and P. Anisa, "Comparison of Classification Data Mining in Process Majors Students," in *2018 4th International Conference on Computing, Engineering, and Design (ICCED)*, 2018, pp. 4–9.

[12] H. Leidiyana, "Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor," *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[13] A. Rohman, "Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa," *Neo Tek.*, vol. 1, no. 1, 2015.

[14] Diana, "Sistem Pendukung Keputusan Menentukan Lokasi Usaha Waralaba Menggunakan Metode Bayes," *J. MATRIK*, vol. 19, no. 1, pp. 41–52, 2017.

[15] D. L. Fithri and E. Darmanto, "Sistem Pendukung Keputusan untuk Memprediksi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes," in *Prosiding SNATIF Ke-1*, 2014, pp. 319–324.

[16] S. Salmu and A. Solichin, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta," in *Seminar Nasional Multidisiplin Ilmu (SENMI) 2017*, 2017, pp. 701–709.

[17] A. Saleh, "Penerapan Data Mining Dengan Metode Klasifikasi Naïve Bayes Untuk Memprediksi Kelulusan Mahasiswa Dalam Mengikuti English Proficiency Test," in *Konferensi Nasional Sistem Informasi (KNSI) 2015*, 2015, pp. 1–7.

[18] K. W. Haryanto and R. A. Saputra, "APLIKASI PREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER (NBC) (STUDI KASUS : DI STMIK YADIKA BANGIL)," *J. SPIRIT*, vol. 10, no. 1, pp. 5–12, 2018.

[19] D. Himawan, "Aplikasi Data Mining Menggunakan Algoritma ID3 Untuk Mengklasifikasi Kelulusan Mahasiswa Pada Universitas Dian Nuswantoro Semarang," Universitas Dian Nuswantoro, 2011.

[20] O. Somantri and S. Wiyono, "Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," *Sci. J. Informatics*, vol. 3, no. 1, pp. 34–45, 2016.

[21] I. Kurnawan, F. Marisa, and D. Purnomo, "Implementasi Data Mining dengan Algoritma Apriori untuk Memprediksi Tingkat Kelulusan Mahasiswa," *J. Teknol. Manaj. Inform.*, vol. 4, no. 1, pp. 204–209, 2018.

[22] X. Wu *et al.*, "Top 10 algorithms in data mining," in *Knowledge and Information Systems*, 2008, vol. 14, no. 1, pp. 1–37.

[23] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm," *2012 Int. Conf. Comput. Sci. Electron. Eng.*, vol. 2, pp. 355–358, 2012.