Evaluation of university accreditation prediction system

M B Musthafa[1,*], N Ngatmari[1], C Rahmad[2], R A Asmara[2] and F Rahutomo[2]

[1] Electrical Engineering Department, State Polytechnic of Malang, Malang, Indonesia
[2] Information Technology Department, State Polytechnic of Malang, Malang, Indonesia

*ichsancomp@gmail.com

Abstract. Higher education accreditation body (Badan Akreditasi Nasional Perguruan Tinggi/ BAN-PT in Indonesian) is an Indonesian body with the main task to assess the quality of Indonesian university. The assessment result is called accreditation, which has 5 years of validation time. In order to monitor the quality, University has an internal mechanism which is called an internal quality assurance system (Sistem Penjaminan Mutu Internal/ SPMI in Indonesian). Usually, SPMI assesses the quality periodically, one or two times each year. This process needs much effort, i.e. time, manpower, and financial cost. Sometimes, internal auditor of the university does not have sufficient knowledge, as much as BAN-PT assessor. This condition causes a lack of assessment accuracy, then causes the quality of SPMI itself. On the other hand, University has abundant of condition data, saved in higher education database (HEDB). This paper proposes to exploit the availability of data in this case. Therefore university able to monitor the quality by machine learning process periodically without much effort as manual SPMI process. Furthermore, this paper evaluates two machine learning methods, i.e. naive Bayes and K-Nearest Neighbor (K-NN). This proposal exploits several data: student, academic, admission, and alumna. K-NN and naive Bayes work in registrant and capacity ratio, student registration ratio, average student Grade Point Average (GPA) in late five years, and on-time graduation scale. The experiment results show the average accuracy of naive Bayes and KNN are 70% and 95.2% respectively.

1. Introduction

In Indonesia, university quality benchmark is called accreditation, which is evaluated by BAN-PT. The accreditation assessment is done in every 5 years. Accreditation is one of important parameter to classify the quality of university, especially private universities [1]. University has an internal mechanism which is called as SPMI. This process needs much effort, i.e. time, manpower, and financial cost. The time effort such as filling in the accreditation form manually. The form involves much data, therefore it needs many employees to handle it. Moreover it involves person (assessor) to assess the form. The availability of assessor between one university and the other is different. The other cases are no assessor available in a university with appropriate capability toward specific study program. Because different types of study program are assessed in a different way. Doing the SPMI process need so much cost in the term of stationery and assessor fee.

On the other hand, University has abundant of condition data, saved in HEDB database. HEDB is a data storage system managed by the center of data and information (Pusat Data dan Informasi/ Pusdatin in Indonesian), the ministry of technology research and higher education (Kementerian Riset Teknologi dan Pendidikan Tinggi/ Kemenristekdikti in Indonesian). The data is accurate because the academic data reporting process is done regularly, conducted twice a year or each semester. Data update can be done as well in every academic cycle, such as short semester. University able to use the data to monitor the accreditation quality periodically by machine learning process. So the university can evaluate and aware of the accreditation value without much effort as manual SPMI.

From this description, this research aims to exploit HEDB data for management information and improvement. This research transforms the data into a data warehouse. The appropriate data warehouse scheme is described in our previous publication [2]. The other publication [3] describes several ideas to use the warehoused data with several computer science and statistical approaches. This paper discusses further in data utilization by management with a classification technique for study program accreditation. The methods evaluated in here are KNN and naive Bayes, because some previous research on knowledge discovery data in field education, the best accuracy is naïve bayes and KNN [4-6]. That algorithm are work based on student, academic, admission and alumna data.

Several kinds of research have been done with data mining methods in academic and university data [7-9]. This paper discusses to exploit academic data in HEDB, which have not been discussed by the other researchers. Furthermore, this paper evaluates classification value of accreditation study program in university using data mining methods base on HEDB by comparing the performance of naive Bayes and K-NN classifier. Comparison is done to get the best accuracy for deployment purpose in real situation.

2. Literature review

The accreditation is assessed regularly every 5 years with fluctuating values [10]. Accreditation is carried out by an independent institution who have been appointed by the government, namely BAN-PT [10]. The university itself can conduct internal evaluation by SPMI, sometime the value of accreditation by SPMI misses far from the results assessed by BAN-PT. there are 2 types of accreditation: Institutional accreditation and study program accreditation. Institutional accreditation shows the quality of educational programs and quality of alumna in the university. Study program accreditation shows the quality of educational programs and quality of alumna in the scope study program of university.

CRISP-DM is a systematic process to implement a data mining project. This research uses the CRISP-DM technique which consists of 6 steps [11]: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Data mining about academic data has been worked by many researchers for understanding, classification or prediction. For example, the researchers in [9] using K-NN and multivariate adaptive regression spline (MARS) for classification value of accreditation in public elementary school. The purpose of that research is comparing two classification algorithm. The evaluation metric of the two classification methods are statistic press's Q, APER, specificity, and sensitivity. The evaluation results show that using MARS method is better than K-NN method. Some researchers [8], work to improve relative accreditation methods based on data mining for higher education. The purpose of that research is applying IT in the accreditation system, therefore the process is run faster and more efficient. Some researches work to control the quality of education and accreditation by predicting period of student's study [7].

3. Method

This section explains the method to classify accreditation value of study program. Figure 1 explains the steps of data classification for accreditation prediction value. Data is obtained from HEDB through a web service application into one of the data warehouse scheme [2]. The data is processed from HEDB to obtain a dataset with attributes: registrant and capacity ratio, student registrant and pass test ratio, average of student GPA in late five years, and value of accreditation. The data is split into 2 types: independent and dependent variable. Independent variables of the attributes are registrant and capacity ratio, student registrant and pass test ratio, average of student GPA in late five years. The dependent variable is value of accreditation history, which is collected from university quality assurance unit. The data is split into 2 parts: training data and testing data. Further detail of this paper research methods is described in the following subchapters.

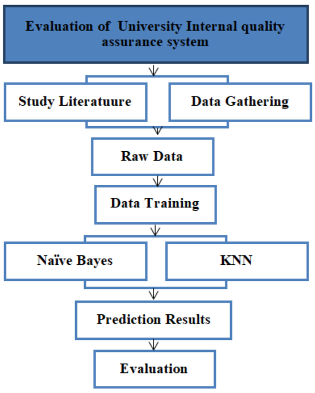

Figure 1. Method to classifying the value of accreditation.

3.1. Business understanding

The purpose of HEDB's data mining is to find data patterns. This data mining is carried out to find data patterns of study program accreditation, so the higher education management able to conduct early detection of study programs quality based on existing transaction data. The expectation is with this monitoring and evaluation, the poor condition will be detected and improvement can be delivered, therefore the results of accreditation will be better.

3.2. Data understanding

Figure 2 shows the process of data collection and analytics from HEDB [2]. The dataset is obtained from HEDB at XYZ University in all undergraduate study programs. The dataset is collected from 2007-2018 academic calendar year. The data is obtained through the HEDB feeder web service application. This web service is installed in university. The data is transformed into data warehouse by integration service, especially developed with Python code. The data warehouse is implemented with MySQL database management system.

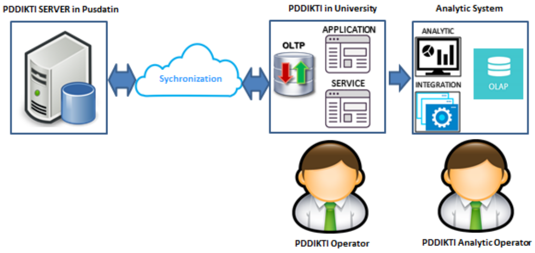

Figure 2. System architecture.

3.3. Data preparation

Selection and processing the data is needed for modeling with several machine learning algorithms. Therefore, the algorithms can be evaluated in order to obtain maximum accuracy result in accordance with the desired target. The data to be processed from HEDB are student, academic, admission and alumna data. Table 1 shows the data examples.

Table 1. Dataset from HEDB.

No	Admissions ratio	Registrants ratio	Average of GPA	Value of accreditation
1	45,04	100	2,73	A
2	71,1	100	3,60	B
3	48,2	100	2,96	B
4	71,33	100	2,75	A
..	..	..	..	..
..	..	..	..	..
80	81,4	100	2,81	A

3.4. Modeling

The dataset modelled by k-NN and naïve bayes classifier, KNN uses neighbor classification as the predicted value of the new instance. K value of KNN means the closest k-data from the test data. Distance parameter commonly used is the euclideance distance. Euclideance distance between two points [12].

$$d(x_1, x_2) = \sum_{i=1}^{n}(x_{1i} - x_{2i})^2 \tag{1}$$

Determining value of K can effect value of accuracy. Naïve bayes classifier to predict the future of opportunities based on experience that has accured. That formula is [12]:

$$P\,(X|H) = \frac{P(X|H)P(H)}{P(X)} \tag{2}$$

This paper uses Python with SciKit Learn library to perform the algorithm. The evaluation strategy uses stratified k-fold cross-validation with 3 and 5 for k value.

3.5. Evaluation

Evaluation is done by the confusion matrix. This evaluation produces accuracy, precision, and recall values [13,14]. The accuracy is a percentage of the number of the records that are classified correctly by the algorithm and the data is true [12]. The precision is amount of positive categorized data categorized correctly / total data classified correctly. The recall is amount of data classified positively / total testing data classified positively. The data will be modeled in two algorithms: K-NN and naive Bayes. This paper evaluates the best algorithm for this case based on the results of the accuracy metric.

3.6. Deployment

Creating prototypes from the results of modeling that produces the best accuracy values using python which retrieves data automatically from the database.

4. Result and discussion

In this experiment, training data is used to produce a model where testing data is used to test the model. Table 2 shows naive Bayes classifier results of accuracy, precision, and recall from several experiments evaluated by k-fold with value of k equals to 3 and 5. The results show the best average accuracy is in k=5 experiment with accuracy 70%.

Table 2. Accuracy, precision, and recall with naive Bayes.

k-fold	**Accuracy (%)**	**Precision (%)**			**Recall (%)**		
		A	**B**	**C**	**A**	**B**	**C**
3	70	22	100	67	36	79	80
	70	30	93	100	56	76	86
	62	11	87	100	20	72	67
Average	67.4	21	93.3	89	37.3	75.6	77.6
5	75	20	100	100	33	82	100
	69	33	100	0	50	78	0
	75	20	100	100	33	82	100
	56	14	88	100	35	67	67
	75	40	89	100	57	80	80
Average	70	25.4	95.4	80	41.6	77.8	69.4

Figure 3 shows the results of tuning parameter with value of k between 1 and 70. Based on the graph in the figure, the best accuracy of K-NN where k is 1

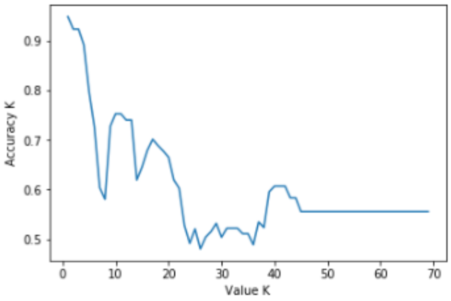

Figure 3. Tuning parameter of K-NN with k value between 1 and 70.

Table 3. Accuracy, precision, and recall with K-NN value of k=1.

k-fold	**Accuracy (%)**	**Precision (%)**			**Recall (%)**		
		A	**B**	**C**	**A**	**B**	**C**
3	100	100	100	100	100	100	100
	89	91	85	100	87	88	100
	96	88	100	100	93	97	100
Average	95	93	95	100	93.3	95	100
5	100	100	100	100	100	100	100
	94	88	100	100	93	89	100
	88	100	75	100	86	86	100
	100	100	100	-	100	100	-
	94	80	100	100	89	95	100
Average	95.2	93.6	95	100	93.6	94	100

Table 4. Accuracy, precision, and recall with K-NN value of k=2

k-fold	Accuracy	Precision (%)			Recall (%)		
	(%)	A	B	C	A	B	C
3	100	100	100	100	100	100	100
	89	91	85	100	87	88	100
	92	100	88	100	89	94	100
Average	93.66	97	91	100	92	94	100
5	100	100	100	100	100	100	100
	75	88	100	25	93	67	40
	88	100	75	100	86	86	100
	100	100	100	-	100	100	-
	94	80	100	100	89	95	100
Average	91.4	93.6	95	81.25	93.6	89.6	85

Table 5. Accuracy, precision, and recall with K-NN value of k=3.

k-fold	Accuracy	Precision (%)			Recall (%)		
	(%)	A	B	C	A	B	C
3	96	89	100	100	94	97	100
	89	91	85	100	87	88	100
	81	62	88	100	67	86	100
Average	88.66	80,6	91	100	82.6	90.3	100
5	100	100	100	100	100	100	100
	75	88	100	25	93	67	40
	88	100	75	100	86	86	100
	100	100	100	-	100	100	-
	81	40	100	100	57	87	100
Average	88.8	85.6	95	81.25	87.2	88	85

Table 3, 4, and 5 show the results of accuracy, precision and recall from several experiments evaluated by k-fold with the value of k equals to 3 and 5 respectively. The experiment results show that the best accuracy value is 95.2%, where K is 1. The number K=1 will make the classification result feel bad because it only takes into account the nearest neighbor or the closest characteristic record, but if the number of K is too much produce cryptic classifications [15].

5. Conclusion

This paper has exposed the possibility of HEDB data exploitation for university management purpose, especially accreditation monitoring and prediction system. This paper also evaluates the usage possibility of several data mining/ machine learning techniques in this issue, i.e. K-NN and naive Bayes classifier. The experiment results show the average accuracy of naive Bayes and KNN are 70% and 95.2% respectively. The results of this data mining method can be used to monitor the accreditation prediction value every semester because the data transactions in HEDB is refreshed in every semester as well. The data in HEDB database affects the results of accreditation prediction.

Acknowledgments

Authors would like to thanks DRPM Kemenristekdikti who funded this research by master thesis research scheme. The research is held in the master of applied science in electrical engineering department, state polytechnic of Malang. We also grateful for providing facilities in the process of doing it.

References

[1] Sinaga A S and Girsang A S 2017 University Accreditation using Data Warehouse *J. Phys. Conf. Ser.* **801** p 12030

[2] Rahutomo F, Rahmad C, Musthafa M B and Ngatmari N 2019 Desain Skema Data Warehouse PDDIKTI sebagai Pendukung Keputusan Perguruan Tinggi *INOVTEK Polbeng-Seri Inform.* **4** pp 90–100

[3] Ngatmari N, Musthafa M B, Rahmad C and Asmara R A 2019 Pemanfaatan Data Pddikti Sebagai Pendukung Keputusan Manajemen Perguruan Tinggi *JTIIK*

[4] Wati M, Indrawan W, Widians J A and Puspitasari N 2017 Data mining for predicting students' learning result *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)* pp 1–4

[5] Durairaj M and Vijitha C 2014 Educational data mining for prediction of student performance using clustering algorithms *Int. J. Comput. Sci. Inf. Technol.* **5** 4 pp 5987–5991

[6] Enriko I K A, Suryanegara M and Gunawan D 2016 Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters *J. Telecommun. Electron. Comput. Eng.* **8** 12 pp 59–65

[7] Peling I B A, Arnawan I N, Arthawan I P A and Janardana I G N 2017 Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm *Int. J. Eng. Emerg. Technol.* **2** 1 pp 53–57

[8] Tastimur C, Karakose M and Akin E 2016 Improvement of relative accreditation methods based on data mining and artificial intelligence for higher education *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)* 1–7

[9] Merluarini B, Safitri D and Hoyyi A 2014 Perbandingan analisis klasifikasi menggunakan metode k-nearest neighbor (K-NN) dan multivariate adaptive regression spline (MARS) pada data akreditasi sekolah dasar negeri di kota semarang *J. Gaussian* **3** 3 pp 313–322

[10] BANPT 2008 *Buku I: Naskah Akademik Akreditasi Program Studi Sarjana* Jakarta: BAN-PT

[11] Larose D T and Larose C D 2014 *Discovering knowledge in data: an introduction to data mining* (John Wiley & Sons)

[12] Han J, Pei J and Kamber M 2011 *Data mining: concepts and techniques* (Elsevier)

[13] Ricardo B Y and Berthier R N 2011 *Modern information retrieval: the concepts and technology behind search* (New Jersey, USA: Addi-son-Wesley Prof.)

[14] Manning C D, Raghavan P and Schütze H 2008 Text classification and naive bayes *Introd. to Inf. Retr.* **1** 6

[15] Indrayanti I, Sugianti D and Al Karomi A 2017 Optimasi Parameter K pada Algoritma K-nearest Neighbour untuk Klasifikasi Penyakit Diabetes Mellitus *Pros. SNATIF* 823–829