Journal of Telematics and Informatics (JTI)
Vol.7, No.1, March 2019, pp. 1~9
ISSN: 2303-3703

The Prediction Of National Exam Scores Of Junior High School Students Using K-Nearest Neighbor (K-NN) Algorithm

Pranoto Wibowo, Imam Much Ibnu Subroto, Sri Arttini Dwi Prasetyowati
Magister Teknik Elektro, Universitas Islam Sultan Agung, Semarang 50112
e-mail: prant_wb@yahoo.co.id; imam@unissula.ac.id; arttini@unissula.ac.id

ABSTRACT

The prediction of the acquisition of national exam scores for Junior High School (JHS) students is intended to know the results of the student's national exam early when students take the national examination. Knowledge gained from the results of this prediction will be important information for the school to take appropriate steps so that the acquisition of student national exam scores can be improved even better. The acquisition of student national exam scores is low and there is no prediction model that is used to predict the achievement of student national exam scores is a problem that needs to be addressed. This paper propose a predicting student's national exam scores for four national exam subjects (INDONESIAN, ENGLISH, MATHEMATICS and SCIENCE) using K-Nearest Neighbor (k-NN) as a prediction method and compare it with Decission Tree method. The results of the study showed that the prediction k-NN model had better performance than the prediction model of Decission Tree. Performance results obtained by evaluating using derivatives of the confussion matrix terminology to determine the value of accuracy, sensitivity (recall), and precission each subjects. To measure the performance of predictive methods used the value of accuracy in each method and each subject. The greater the accuracy value (max 1), then the better performance of the prediction model used. Performance of k-NN in average accuracy=0.85, precision=0.87, recall=0.91 is better than Decission Tree method performance with accuracy=0.82, precision=0.85, and recall=0.89.

Key words : *UN, k-NN, Decission Tree, Prediction, Accuracy*

1. Introduction

National Examination who called UN is the activity to measure the achievement of student competence in some certain subjects in group science subject and technology in order to assess achievement of national education standart [1]

National examinations are held during teaching and learning activities has been followed by students that lasted three years in level of education followed. Implementing national examination for JHS level there are four subjects tested, they are INDONESIAN, ENGLISH, MATHEMATICS and SCIENCE. In the last three years, the value of national examinations as the level of basic and middle education are not used as the determinant of graduation but as the bases for register in school selection of a higher level. For junior high school level like in JHS 13 TEGAL so the final exam score are used to register in Senior/ Vocational High School. The value of National Exam results are used to: a) one of the graduation requirements from the education unit; b) one of the considerations in the selection to register the next level of education; c) quality mapping; and d) guidance and assistance for quality improvement [2].

For the past three years, the national exam score of JHS 13 Tegal students are below average 60,00 or D (LESS) thus influencing the value of the success of the school and the students themselves. This condition needs attention and needs to be solved So that the expected condition is the fullfillment of the school's target, that is by obtaining the national exam average scores above 60,00 or C (ENOUGHT) .

To get a minimum national exam score of 60,00 or C an action or activity is needed so that student are encouraged to learn and get maximum score. Besides national exam preparation activities, the school have to be able to map students who have the potential get score below 60,00 or C. So that the mapping carried out by the school gets maximum results, it is necessary

to predict the value of the national exam for nine grade students as early detection from school to anticipate students with score below 60,00 or D.

The solution for the school regarding the value of the national examinations, an effective and efficient method is needed to predict nasional exam score, so the author proposed k-Nearest Neighbor (k-NN) as a solution to predict the national exam score of JHS 13 Tegal students. K-NN method is very affective and has been used to predict achievement students and determination of Senior High School students major.

This k-NN model is expected to predict students who get national exam scores below 60,00 or D. The result of this prediction are expected to be preliminary information or early detection so that the school can take policy towards students when going up to nine grade especially by giving more attention to students who are predicted to get national exam score below 60,00 or D. With this research model, it is expected that the school can immediately find out early and can take the policy so that the national exam score of students can be increased.

From the background of the problem there are still many students of JHS 13 Tegal who get national exam score below 60,00 or D and school also doesn't yet have a prediction model to detect of student national exam scores early as initial information in improving students achievement of national exam score, to fill school target, so that the formulation of the problem in this study is :

" How does the national exam value prediction model of JHS 13 Tegal students use the k-NN algorithm?".

Related Study

National Examination who called UN is activities for measuring the performance of graduates in certain saubject nationally by referring to the Standard of Graduet Competence. In JHS level, there are four subjects in final examination, they are INDONESIAN, ENGLISH, MATHEMATICS and SCIENCE. Final Examination in Junior High School level is held in nine grade after completing all the educational process up to the school exam. National exam score who called UN score is the value obtained by students results of the final exam taken. [4].Un score for junior high school level consist of INDONESIAN, ENGLISH, MATHEMATICS and SCIENCE subject's scores which is written in the certificate of National Examination Results (NER). The NER is used as a basis for registering for High or Vocational School level.

Refer to [4] the mining data has several equivalents, likes knowledge discovery or pattern recognition. Both terms actually have their respective accuracy. Terms of knowledge discovery right to use because the main purpose of mining data is indeed to get knowledge that is still hidden in the chunks of data. Pattern recognition term is right for use because the knowledge that is about to be explored is in the form of a pattern that might also need to be explored in the chunks of data that are being faced. If in this paper the term data mining is used, this is based more on the more popular term in the activity of extracting data knowledge..

According to [5], Data mining is a term used to describe knowledge discovery in the data base. Data mining is process that uses statistical techniques, mathematics, artificial intelligence and learning machine to extract and identify useful information and related knowledge from a variety of database. Data mining is process that uses statistical techniques, mathematics, artificial intelligence and learning machine to extract and identify useful information and related knowledge from a variety of database. Mining data is divided into several groups based on the tasks performed, they are [4] :1. Estimation, estimation is almost the same as clarification, except the target variable is estimatedmore in numerical direction than in the direction of the category. The model is built using a complete record that provides the value of the target variable as a predictive value. Then, in the next review the estimated value of the target variable is based on the value of the prediction variable. As an example, will be estimated of systolic blood pressure inhospital patients based on the age of the patient, gender, body weight indext, andblood sodium levels. The relationship between systolic blood pressure and the value of the predictive variable in the learning process will produce an estimation model. The estimation model produced can be used for other new cases. 2. Prediction, prediction is amost the same as the clarification and estimation, except that in the prediction the value of the results will be in the future. 3. Clarification, in clarification there is a variable target for categories, for example, income clasification can be separated into three categories, they are high income , medium income and low income. 3. Clustering, clustering is a record grouping, observation or pay attention and form a class of of objects that have similarities 4. Cluster, cluster is collection of records that have similarities with

each other and have an incompatibility with records in other clusters. 5. Association, association duty in mining data is find attributes that appear at one time. In the business word commonly called Shoping basket analysis.

Nearest Neighbor is a pproach to finding cases by calculating the closeness between new cases and old cases. That is based on matches weight from a number of features [4]. k-NN is done by looking for groups of k object in the closest training data (similar) to objects on new data or testing data. k-NN is one method used to predict the output variable with a classification approach. On the classification approach, The dataset is divided into training data and testing data. k-NN uses the measurement of proximity /similarity between testing data and training data by calculating distance. Then k-NN will choose the number of training data closest to the test data which most appear in predicting the output variable.

Steps used in calculating k-NN prediction is 1. Specify the parameter k (number of closest neighbors), 2. Calculate the square of Eucliden (query instance) distance training data gaints testing data, 3. Sort the Euclidean distance of the training data from the smallest.

Euclidean distance is a measure of the proximity or similarity between training data and testing data.

If the training data is in the form X_{Bi} (X_{B1}, X_{B2}, X_{B3}, …., X_{Bn}) and testing data X_{Ai} (X_{A1}, X_{A2}, X_{A3}, …, X_{An}) then the Euclidean distance equation is as follows :

$$D_{AB} = \sqrt{\sum_{i=1}^{n} (X_{A,i} - X_{B,i})^2} \quad (1)$$

k-NN has several advantages, which are resilient to training data that has a lot of noise and is effective when training data is large. This is the reason researchers to take k-NN as their research method.

2. Research Method

This first research begins with the literature search phase, then identification and data collection, then initial processing, the the method used, experiment and testing, and the last is evaluation and conclusion. 1. Literature search. This stage is to look for literature from books and previous research journals about predictions, *data mining* methods using *k-Nearest Neighbor* algorithms. 2. Identification and data collection. At this stage identification of the research that will be carried out and collecting data in accordance with research. On data collection, The data used is the grade data of JHS 13 Tegal generations 2015/2016 to 2017/2018 obtained from the dapodik application and the admission archive of new students at the school. The number of data from JHS 13 Tegal graduates for three years is 645 data with attributes including: No, Name, value of National Standarized School Exam of Elementary School (NSSEES), report card value of semester 1 dan 2 in grades 7 and 8, and National Exam (NE) score for each subjects of INDONESIAN, ENGLISH, MATHEMATICS and SCIENCE. The data is collected from existing documents and used as initial data. 3. Data preparation and selection. Prepare for the data that has been obtained such as looking at the table structure that is in the database. Data selection is done because not all tables and data in the database are related to the research conducted., so only data relating to the research will be used. In this study what is needed is data with numerical attributes or variables. From the total of 645 collected data, it was then re-selected to 400 data to be used for this study. After the data has been obtained, the next step is to select data from the existing database, to use data easier because what will used is in the form of attribute/ value variable or numerical based on subjects, then on the dataset selected variables/ attributes that are suitable to be used as input and output data that is the value of NE JHS, value of NSSEES (INDONESIAN, MATHEMATICS, dan SCIENCE subjects), report card value of NE subjects in grade 7 and 8 (semester 1 to 2).

To make the research easier, the value is coverted according to [7], and shown in table 3.1 the list of value conversions.

Table 1. The list of value conversions

VALUE	CONVERTION	PREDICATE	CRITERIA
91 – 100	4	A	EXCELENT
75 – 90,9	3	B	GOOD
60 – 74,9	2	C	ENOUGHT

0 – 59,9	1	D	LESS

After selecting data and determining input and output data, the next step is convert the value according to table 3.1, the next step is determining training data and testing data. Training data and testing data were taken from datasets prepared for each NE subjects. Of the 400 data that have been prepared, then training data is made in the amount of 250 and testing data is 150 data. 4. Method used. In this study the method used is *k-Nearest Neighbor* to predict the achievement of national exam of JHS 13 Tegal. From the attributes that exist algorithm *k-Nearest Neighbor*, where the attributes used up to n attributes will be used *Euclidean* formula calculations for the four UN subjects. 5. Experiment and testing. For experiments and testing using student datasets which are processed into datasets for all four subjects. From the dataset of each subject that has been thereafter made an experiment using the k-NN prediction model. In the experiment k-NN model, the training model is made training data as much as 250 data and testing data as much as 150 data then carried out the experiment three times for the value of k that is different that is each k=3 k=5 k=7 and k=9. 6.Evaluation. To evaluate the method used in this study used method *Confusion Matrix*[8].

Table 2. *Confusion Matrix* Table

		Actual Value	
		TRUE	**FALSE**
Predictive Value	**TRUE**	TP *(True Positive)*	FP *(False Positive)*
	FALSE	FN *(False Negative)*	TN *(True Negative)*

Description of table 3.6. :TP (*True Positive*): the results of the prediction are in accordance with the actual values which are the same as true . TN (*True Negative*): the result of the prediction is false and the actual value is also false. .FP (*False Positive*): the result of the prediction is true, but the actual value is also false. FN (*False Negative*): the result of the prediction is false, but the actual value is also true. Some evaluation methods can be derived from *confusion matrix*, they are:

1. Accuracy. Accuracy is the result of calculating all correct predictive values divided by the entire data. The best accuracy value if the accuracy value is equal to 1,0 and the worst value is 0,0.

$$Akurasi = \frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

2. Precision. Precision is calculated from the total number of true positive predictive values divided by the total number of correct class predictions. The best precision value is 1,0 while the worst is 0,0..

$$Presisi = \frac{TP}{TP+FP} \quad (3)$$

3. Recall (*Sensitivity / SN*). Recall or *sensitivity* Recall is calculated from the number of true positive predictions divided by the total number of positive classes The best SN value is 1,0 and the worst value is 0,0.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Software used in helping yhis research is Weka (*Waikato Environment for Knowledge Analysis)*, Weka is an integrated software thath contaions the implementation of data mining methods. Weka was developed by Wakaito University, New Zealand uses the Java programming language[8].

4. Result and Discussion

To get a prediction model, at the experimental stage carried out through three stages, they are: 1. Dividing the data that has been prepared into two, they are training data of 250 (62.5%) and *testing* data of 150 (37.5%). 2.Experiment using the *Decision Tree* method as a comparison. 3. Experiment using *k-Nearest Neighbor (k-NN).*For *k-NN* method using two data sources, they

are: a. data with converted values and predicate NE value b. data with original values data predicate NE values. 4. Experimental results were evaluated using *Confusion Matrix*.
Data values for each subject are shown in the table 3 to tabel 6.

Table 3. Value data of INDONESIAN subject

NO	STUDENT'S NAME	**INDONESIAN**					
		NSSEESIND	IND71	IND72	IND81	IND82	**NE IND**
1	AFDOL FAIZIN	64.0	73	75	78	75	68.0
2	ANDRE LAKSONO	68.0	75	74	78	74	70.0
3	ANGGITA IKA FITRIYANI	66.0	77	80	82	85	76.0
…	……..	….	…	…	…	…	…
399	TAUFIQ NAZAR MAULANA	42.0	75	76	76	76	46.0
400	WINDI AGUSTIN	64.0	76	75	76	77	64.0

Table 4. Value data of ENGLISH subject

NO	STUDENT'S NAME	**ENGLISH**				
		ENG71	ENG72	ENG81	ENG82	**NE ENG**
1	AFDOL FAIZIN	77	73	80	85	40.0
2	ANDRE LAKSONO	74	70	74	72	60.0
3	ANGGITA IKA FITRIYANI	76	75	83	80	66.0
…	…..	…	…	…	…	…
399	TAUFIQ NAZAR MAULANA	73	75	74	76	38.0
400	WINDI AGUSTIN	74	74	74	74	46.0

Table 5. Value data of MATHEMATICS subject

NO	STUDENT'S NAME	**MATHEMATICS**					
		NSSEESMATH	MATH71	MATH72	MATH81	MATH82	**NE MATH**
1	AFDOL FAIZIN	55.0	72	70	74	76	40.0
2	ANDRE LAKSONO	65.0	73	77	71	77	47.5
3	ANGGITA IKA F	62.5	70	74	71	75	52.5
…	….	…	…	…	…	…	…
399	TAUFIQ NAZAR	50.0	79	73	74	74	42.5
400	WINDI AGUSTIN	55.0	73	74	80	76	45.0

Table 6. Value data of SCIENCE subject

NO	STUDENT'S NAME	**SCIENCE**					
		NSSEESC	SC71	SC72	SC81	SC82	**UN SC**
1	AFDOL FAIZIN	52.5	76	73	80	76	40.0
2	ANDRE LAKSONO	62.5	70	84	76	75	65.0
3	ANGGITA IKA FITRIYANI	50.0	77	72	80	76	55.0

…	….	…	…	…	…	…	…
399	TAUFIQ NAZAR MAULANA	47.5	72	71	73	73	37.5
400	WINDI AGUSTIN	52.5	71	72	74	75	40.0

The results of the experiments predicting national exam scores for all subjects using weka for decission tree and k-NN method displayed in the table 4.5

Table 7. Recapitulation of the accuracy of national exam value prediction using Weka

METHOD	INDONESIAN	ENGLISH	MATHEMATICS	SCIENCE
	%	%	%	%
Decission Tree	48	**86.6667**	98.6667	95.3333
k-NN (Convertion_Predicate)				
k=3	44	**86.6667**	98.6667	94
k=5	50.6667	**86.6667**	98.6667	95.3333
k=7	50.6667	**86.6667**	98.6667	95.3
k=9	52.6667	**86.6667**	98.6667	95.3333
k-NN (Value_Predicate)				
k=3	50	85.3333	98.6667	**96**
k=5	60	84.6667	98.6667	95.3333
k=7	**56.6667**	85.3333	**99.3333**	95.3333
k=9	54.6667	86	**99.3333**	**96**

From the table 7 it can be seen for the *k-NN* method that the accuracy value is better than the decission tree method at each sibjects except for ENGLISH which is the value of decission tree accuracy same with k-NN in all k by using the converted value data and predicates of 86.6667 %. For *k-NN* method the data of value and predicate at INDONESIAN, the accuracy value of 56.6667 % with k=7, MATHEMATICS subject the accuracy value of 99.3333 % with k=7 and k=9, and SCIENCE subject the accuracy value of 96 % with k=3 and k=9.
Figure 1 to 4 showed the comparation of the accuracy value method based on each subjects.

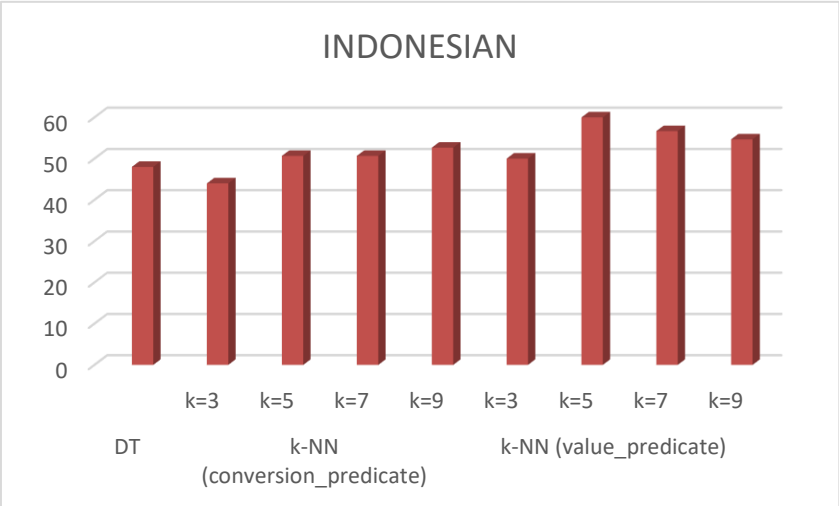

Figure 1. The comparation of prediction method accuracy NE score of INDONESIAN Subject

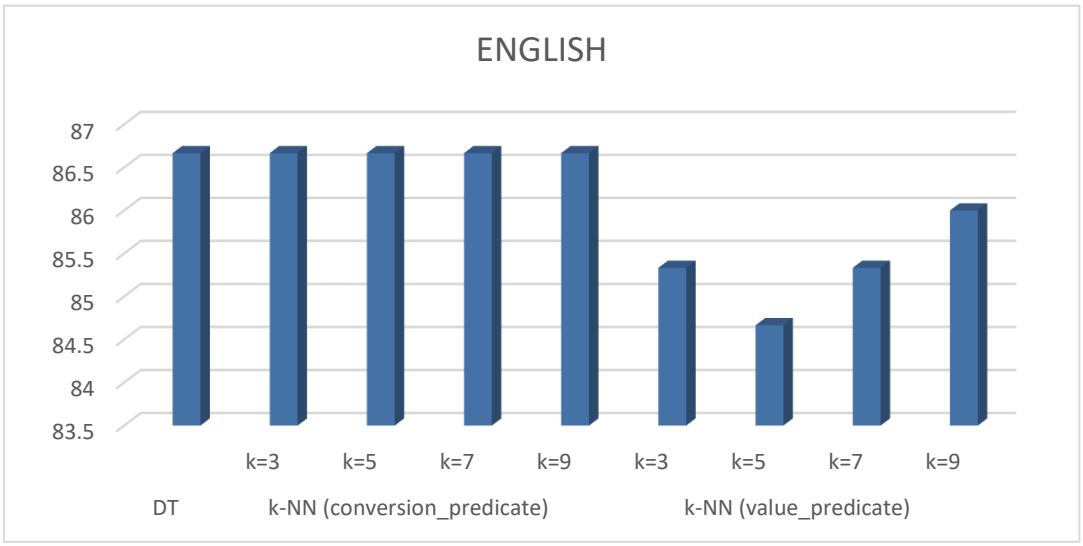

Figure 2. The comparation of prediction method accuracy NE score of ENGLISH Subject

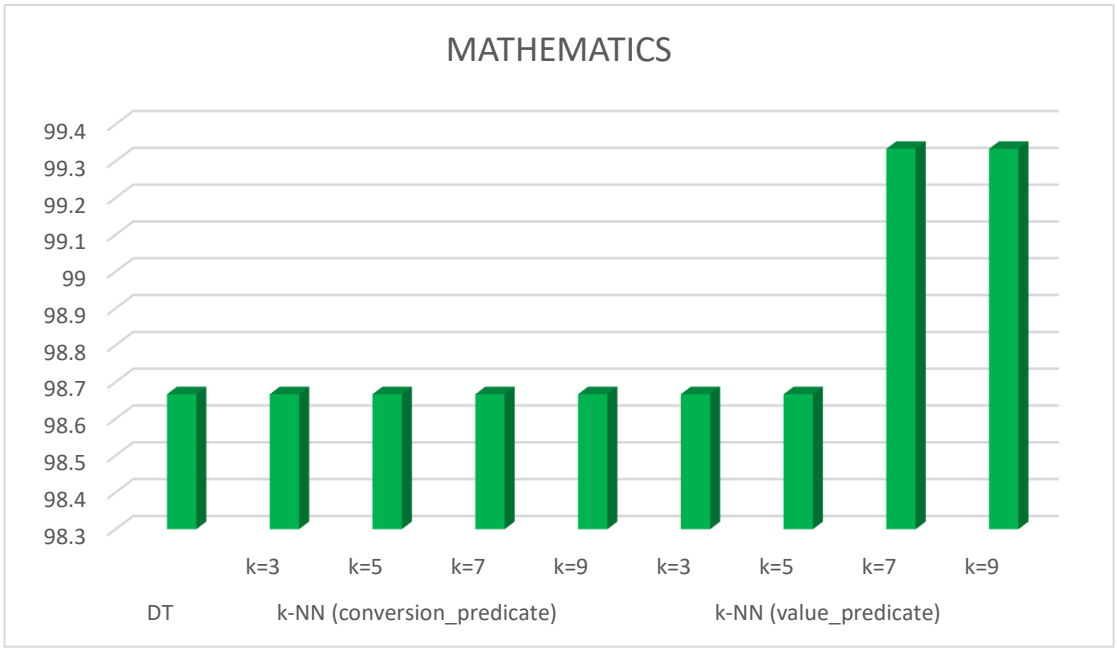

Figure 3. The comparation of prediction method accuracy NE score of MATHEMATICS Subject

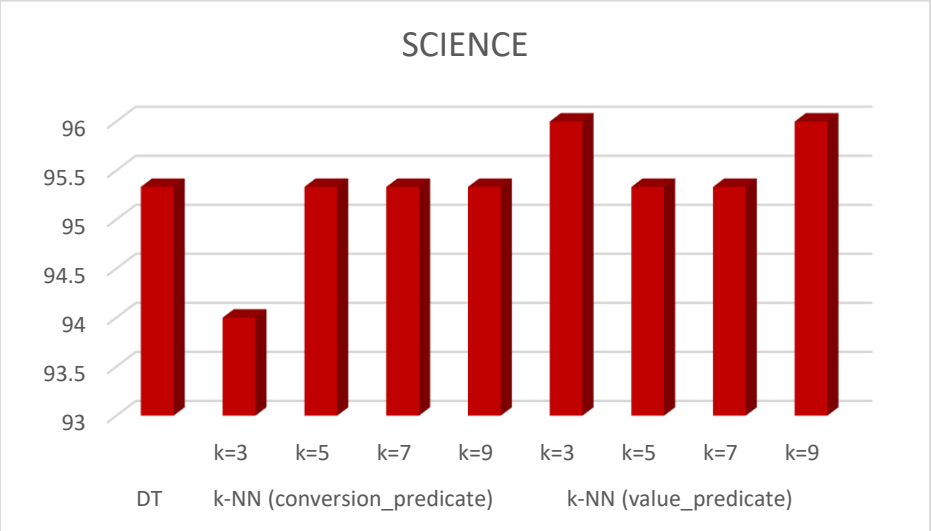

Figure 4. The comparation of prediction method accuracy NE score of SCIENCE Subject

Based on the result of the experiment using the dicession tree and k-NN method the result can be evaluated using derivative of the confussion matrix terminology to find out value of accuracy, sensitivity and precission in each subjects.

Table 8. Predict evaluation of Final Exam score

METODE	k	NILAI	INDONESIAN			ENGLISH			MATHEMATICS			SCIENCE		
			ACC	PRE	REC	ACC	PRE	REC	ACC	PRE	REC	ACC	PRE	REC
k-NN	3		0.44			0.86			0.98			0.94		
(Convertion		B		0.49	0.41									
Predicate)		C		0.4	0.48									
		D		0.59	0.44		0.87	1		0.99	1		0.95	0.97
	5		0.5			0.86			0.98			0.95		
		B		0.58	0.49									
		C		0.45	0.6									
		D		0.59	0.41		0.87	1		0.99	1		0.95	1
	7		0.5			0.86			0.98			0.95		
		B		0.58	0.49									
		C		0.45	0.6									
		D		0.59	0.41		0.87	1		0.99	1		0.95	1
	9		0.53			0.86			0.98			0.95		
		B		0.63	0.49									
		C		0.47	0.6									
		D		0.59	0.41		0.87	1		0.99	1		0.95	1
k-NN	3		0.5			0.86			0.98			0.96		
(Value		B		0.54	0.39									
Predicate)		C		0.48	0.52		0.2	0.05					0.5	0.2
		D		0.51	0.59		0.88	0.98		0.99	1		0.97	1
	5		0.6			0.84			0.98			0.95		
		B		0.62	0.59									
		C		0.57	0.65		0.17	0.06						
		D		0.63	0.57		0.88	0.97		0.99	1		0.95	1
	7		0.56			0.85			0.99			0.95		
		B		0.56	0.49									
		C		0.53	0.64		0.2	0.05		1	0.5			
		D		0.62	0.61		0.88	0.99		0.99	1		0.95	1
	9		0.48			0.86			0.99			0.96		
		B		0.58	0.46									
		C		0.43	0.54		0.25	0.06		1	0.5		1	0.2
		D		0.49	0.43		0.88	0.99		0.99	1		0.96	1
DT			0.48			0.86			0.98			0.95		
		GOOD		0.58	0.46									
		ENOUGH		0.43	0.54									
		LESS		0.49	0.43		0.87	1		0.99	1		0.95	1

Note : *ACC : Accuracy; PRE: Precission; REC: Recall*

Data presented at table 8 show that in general k-NN method has greater value of accuracy than decission tree method. For ENGLISH, MATHEMATICS and SCIENCE subject have greater of accuracy while INDONESIAN subject shows smaller accuracy value.

5. Conclusion

Aplication of k-NN method used to predict the students of JHS 13 Tegal exam score has better performance than using decisssion tree method. Average value of accuracy for INDONESIAN subject with k-NN method k=5 is 0.6 and *decission tree* method is 0.48. Average value of accuracy for ENGLISH subject with *k-NN* method k=9 is 0.86 and *decission tree* method is 0.86. Average value of accuracy for MATHEMATICS subject with *k-NN* method k=9 is 0.99 and *decission tree* method is 0.98. Average value of accuracy for SCIENCE subject with *k-NN* method k=9 is 0.96 and *decission tree* method is 0.95.

Bibliography

[1] Kemdikbud, 2013. *Permendikbud RI No 66 Tahun 2013.* Jakarta : Kemdikbud
[2] BSNP, 2013. *Prosedur Operasi Standar Penyelenggaraan Ujian Nasional SMP/MTs, SMPLB, SMA/MA, SMALB, SMK/MAK serta Pendidikan Kesetaraan Program Paket B/Wustha, Program Paket C, dan Program Paket C Kejuruan Tahun Pelajaran 2013/2014*. Jakarta: BSNP
[3] BSNP, 2016. *Prosedur Operasi Standar Penyelenggaraan Ujian Nasional SMP/MTs, SMPLB, SMA/MA, SMALB, SMK/MAK serta Pendidikan Kesetaraan Program Paket B/Wustha, Program Paket C, dan Program Paket C Kejuruan Tahun Pelajaran 2016/2017*. Jakarta: BSNP
[4] Susanto, S. & Suryadi, D., 2010. *Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data*. Yogyakarta : Andi Offset
[5] Kusrini & Luthfi, E.T.,2009. *Algoritma Data Mining*. Yogyakarta : Andi Offset
[6] Turban, E.,dkk.2005. *Decicion Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset
[7] Kemdikbud, 2014. *Permendikbud RI No 104 Tahun 2014*, Jakarta: Kemdikbud
[8] Adinugroho, Sigit; & Sari, Yuita Arum. 2018. *Implementasi Data Mining Menggunakan WEKA*. Malang: UB Press