Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions

Windu Gata, Hasan Basri
Computer Science
Post Graduate STMIK Nusa Mandiri Jakarta
Jakarta, Indonesia
windugata@nusamandiri.ac.id,
hasanbasrisukses@gmail.com

Rais Hidayat, Yuyun Elizabeth Patras
Education Management
Pakuan University
Bogor, Indonesia
rais72rais@gmail.com, yuyunpatras64@gmail.com

Baharuddin Baharuddin
Faculty of Islamic Religion
Universitas Muhammadiyah Makassar
Makassar, Indonesia
afinyeyen@yahoo.com

Rhini Fatmasari
Education Management
Universitas Terbuka
Jakarta, Indonesia
riens@ecampus.ut.ac.id

Siswanto Tohari
Faculty of IT
Budi Luhur University
Jakarta, Indonesia
siswantobl@gmail.com

Nia Kusuma Wardhani
Faculty of Economic
Mercubuana University
Jakarta, Indonesia
nia.kusuma@mercubuana.ac.id

***Abstract*—The online game is a game which is currently booming and interest ranging from children, teens, to adults. Online games can create a sense of opium to the people who play it. Online games become a new problem for the students, because online games make learning impaired concentration. The learning achievements can be measured from the value of report cards. The challenge on this research can be carried out using a method of classification for predicting learning achievements using algorithms of classification i.e. Naïve Bayes, Random Forest, and C4.5. After the third comparison algorithm, then the prediction results obtained by learning achievements. Naïve Bayes algorithm proved that value the accuracy and value of the AUC 69.18% of 0.771 contains the classification, fair for the random forest algorithm accuracy 66.34% and AUC values of 0.738 contains the classification, fair as for algorithm C4.5 65.65% accuracy and value of the AUC of 0.686 including into poor classification. From these results it can be concluded that the naïve bayes algorithm has higher accuracy compared with the random forest algorithm and C4.5, visible difference in accuracy between the naïve bayes with random forest of 2,84%, whereas the difference between the naïve bayes with C4.5 of 3,53%. Naïve bayes algorithm is thus able to predict achievement students can study better.**

Keywords*—*online games; learning achievement; naïve bayes algorithm; random forest; C4.5

I. INTRODUCTION

Each activity performed by a child without any supervision from parents will impact negatively, as well with the activity of playing games online that are point of view may spend a lot of time, so that on previous research indicates that online gaming can cause adiksi on everyone who played it. This happens because the online game gives adiksi to everyone who ever tried and often play online games. As a result, with a sense of opium that had been attached to the child or the person, raises curiosity to play it back.

Since the beginning of the emergence of the online game that increasingly expanded and diversified, starting from nintendo, sega, and online games and being a trend in recent years. The online game is certainly very easily playable by anyone with an internet connection. Now that online games can be accessed through hand-held phones without having to come into the Café.

The presence of online games has resulted in several conflicts among online game players. Online gaming addiction on the research conducted by Syahran found that results of research of a adiksi video game experts in America, Mark Griffiths of Nowingham Trent University, at an early age children find almost a dozen years a third play online games every day, which is more worrying about 7% of his playing for at least 30 hours/week, this is caused because the child aged

12-18 average year often play online games with surfing the internet that does not protected from bad information. Of expert psychology in America, David Greenfield, found about 6% of internet users are experiencing online gaming addiction. [1].

Other research described that online gaming is able to give the interest of all persons without any age limit, regardless of that online games can provide a negative value. This is because online games can make sense of addiction to play it [2].

Reliance online games that are currently being experienced by adolescents, resulting in that many teens not concentration in the learning process, so nothing much to learn in the achievements of each semester. This is demonstrated by the large number of teenagers are addicted to and forget the time when playing the game online [3].

Generally there is no online games which provide educational value for the gamers, therefore a gamers tend to be inactive at the moment processed so that the achievements of the study experienced a decrease [4].

Variety of ways conducted to find out the cause of anything that effects of adiksi games online against the achievement of learning. Therefore, with the rapid development of technology and knowledge-based computer systems, it has become a part of a study that the researchers must be involved in every area of computer science. This research was conducted to help solve problems by using classification of data mining to determine the predictions of student learning achievement. Need for a method that can process data-data is already collected from the results of data collection conducted in this research.

Based on previous studies, this research is conducted by means of implementing data mining methods to compare and naïve bayes, random forest, and C4.5 to figure out algorithms that have the highest accuracy, in terms of These predictions adiksi game online on the achievements of the study that has never been done on previous research.

II. Related Research

On the research of the influence of online game against the aggressive behavior of teenagers in Samarinda explained that the change in behavior caused by teens online games played [5]. The relationship of playing games with the motivation of junior high school students in district of Bacolod West very significant relationship since online games can create learning concentration distracted [6]. Research adiksi online game in indonesia shows that junior high school students spent much time playing online games [2]. Online gaming addiction trigger impact on withdrawal, aggressiveness, the problem of interpersonal relationships and lead to psychological [7]. in other literature explaining that online gaming is not considered to have a negative impact [8]. Online gaming makes the first State high school students 1 Kuta has decreased achievement [9]. appears a question, whether there is a relationship of the liveliness of the freedom of Association, the guidance of parents, and discipline against the achievement of learning [10].

This research more interesting with a combination of education and computer science that is data mining. Some related research include: Handling Imbalanced Data in Customer Churn Prediction Using Combined Sampling and Weighted Random Forest [11]. The old prediction studies students with a method of random forest. Hypertension Prediction System Using Naive Bayes Classifier [12]. Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine, and Naive Bayes Classification [13]. Prediction of Timeliness Graduation of Students Using Naïve Bayes: A Case Study at Islamic State University Syarif Hidayatullah Jakarta [14]. Student Academic Performance Evaluation Using Naïve Bayes Algorithm (Case Study: Fasilkom Unilak) [15]. Naive Bayes Method For Prediction Of Graduation (Case Study: College Freshmen Data) [16]. Data mining to predict the type of transaction on a cooperative loan with algorithm C4.5 [17]. Decision support system-based decision tree in the awarding of the scholarship case study: AMIK "BSI yogyakarta" [18]. Sentiment analysis forest fire news public opinion through comparisons of algorithms of support vector machine and k-nearest neighbor based particle swarm optimization [19]. Application of algorithm C4.5-based particle swarm optimization for ease of service results prediction donate tithes and program [20].

III. Methodology

In this study using the methodology CRISP-DM (Cross-Industry Standard for Data Mining).Stages of the Crisp-Dm is comprised of Business Understanding, Understanding, Data Preparation, Data Modeling, Evaluation, Deployment [21].

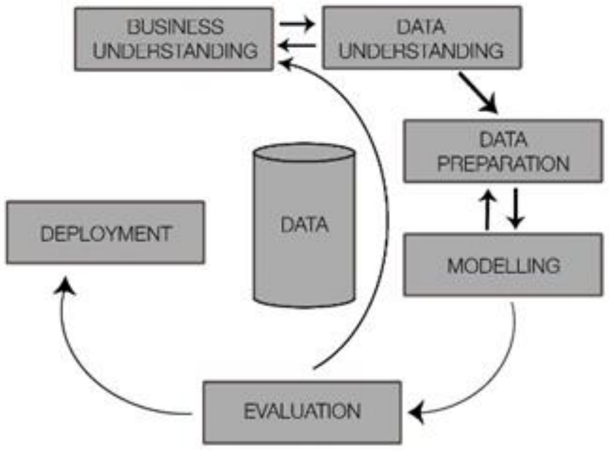

Fig. 1. CRISP-DM.

A. Business Understandiing

This chapter will discuss about business understanding to objects of research is MAN 4 Karawang students. The focus of the research object in the students of class XI, where they have a history of previous semester report cards value. The fluctuating value of report cards that are felt by the students into a research draws on this occasion. Therefore, researchers will conduct the dissemination of questionnaires to find out the cause and effect of the fluctuating value of the report cards.

B. *Data Understanding*

At this stage carried out data collection the value of students report cards from the staff of the curriculum, the value of IQ as well as students assisted with the dissemination of a questionnaire in order to obtain more accurate information. After that the data source is analyzed so that found the fluctuating value of the intersection of students MAN 4 Karawang.

TABLE I. DATA FROM THE STUDENTS OF CLASS XI

No	Majors	Total Class	Total Students
1	IPA(nature of science)	3	117
2	IPS(social science)	4	146
Subtotal			263

C. *Data Preparation*

At this stage of preparing the data already obtained. Then the data in the analysis of the fluctuating value of the report cards as many as 263 data. To obtain high-quality data, then do the preprocessing techniques. As for the techniques used in preprocessing [22] as follows: Data Cleaning, Data Integration, Data Reduction. After the preprocessing phase is complete then obtained the selected attribute data to serve as training and testing data.

TABLE II. ATTRIBUTES OF A LEARNING ACHIEVEMENT PREDICTIONS

No	Attribute	Value
1	NIS	Numerik
2	Gender	MALE & FEMALE
3	Majors	IPA (nature of science) & IPS(social science)
4	Organization	ACTIVE & NOT ACTIVE
5	IQ	ABOVE THE AVERAGE
		AVERAGE
6	Play Games Online	FREQUENT & RARE
7	Online Gaming Comunity	YES & NO
8	Play games online/week	1 TIMES A WEEK, 2 TIEMES A WEEK , EVERY DAY
9	The Duration of Playing Online Games	1 HOURS, 2-5 HOURS, >5 HOURS
10	Duration of Study	<1 HOURS, 2-3 HOURS, >5 HOURS
11	Place to Play Games Online	HOME, SCHOOL, CAFE
12	Ram Capacity	2 GB, 3-4 GB, >4 GB
13	Online Game	Mobile Legend, Clas of Clas, PUBG dan lain-lain.
14	The Achivements Of The Class	DECREASED & ESCALATE

In the table above there are 14 field that had passed the results of data cleaning, data integration, data reduction, and 178 record as research material in determining the prediction game online on the learning achievements of students in the MAN 4 Karawang.

D. *Modeling*

Stages of modeling based on model algorithm used in the study. Research on modeling for this time using 3 models: naïve bayes algorithm i.e., random forest, c 4.5. The third model of the algorithm is processed using tools rapid miner 7.3.

1) Naïve Bayes Algorithm: Bayes method this is a good method in machine learning based on data training, by using conditional probabilities as a requirement [23]. According to Bramer in the journal septiani Naive Bayes Classification 2017 is the classification of statistics that can be used to predict the probability of membership of a class [24].

Bayesian classification based on Bayes theorem, named after a mathematician who is also Minister of the United Kingdom Presbyterian, Thomas Bayes. Naïve bayes method has a rule that is used to calculate the probability of a class. Naïve bayes algorithm provides a method to combine the chance or opportunity advance with the terms likely to be a formula that can be used to calculate the odds of any chances of that happening. As for the General form of the theorem, bayes ' rule as follows:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

2) Random forest algorithm: Random Forest (RF) is an algorithm or the development of derivatives of a single decision tree. RF algorithms which are composed of some tree or a decision tree where each tree is done training data samples [25]. The methods of Random Forest (RF) is a method which can improve the accuracy of the result, because in the child node to evoke each node is done randomly. This method is used to construct a decision tree that consists of a root node, internal nodes, and leaf nodes by taking random data attributes and corresponding provisions in force. The root node is the node that is located at the top, or commonly referred to as the root of the decision tree. An internal node is the node branching node, where it had an output of at least two and there is only one input While the leaf node or terminal node is the last node only has one input and does not have any output. To rate the value of entropy and information gain value can use the following equations:

$$Entropy\ (Y) = -\sum_i p(c|Y) log2\ p(c|Y) \tag{2}$$

Where Y is the set of case and p (c | Y the Y value) is the proportion of class c.

Information Gain

$$(Y,a)= Entropy\ (Y) \sum_{v \in Value(a)} \frac{|Yv|}{|Ya|}\ Entropy\ (Yv) \tag{3}$$

Where values (a) represents all possible values in a set of cases. YV is a sub class of the Y class v relating to class a. Yes, is all the values that correspond to a.

The selection of attributes as nodes, both root (root) or internal nodes based on information gain is the highest of the attributes exist.

$$Split\ Information\ (S,A) = \sum_{i=1}^{c} \left(\frac{|Si|}{|S|}\right) log2 \left(\frac{|Si|}{|S|}\right) \quad (4)$$

Where the split information (S, A) is the value of the input variable entropy estimation of S that have class c and/Si///S/is the probability of class i attribute.

$$Gain\ Ratio\ (S,A) = \frac{Information\ Gain\ (S,A)}{Split\ information\ (S,A)} \quad (5)$$

3) C4.5 Algorithm: Algorithm of decision tree or algorithm is called C 4.5 is an algorithm that has a concept on the approach to devide-and-conquer for a classification process is an issue [18]. C 4.5 essentially in its rate decision tree can be done in several steps: setting up the data of the selected node, determine training (selected node is the smallest value of entropy search results), and the last step is create a decision tree using the rule that has been obtained [22].

The algorithm C 4.5 there are several stages in making the decision tree as follows [26] among other things: Prepare the training data, Find and calculate the Entropy before searched each Entropy class, Calculate the value of the gain and the average gain.

Calculate Entropy

$$H(X) = \sum_{j} -pj\ log_2\ (pj) \quad (6)$$

Calculate the value of the gain and the average gain

$$Gain\ Average = H(T) - Hsaving(T) \quad (7)$$

All the above algorithm is implemented using rapid miner version 7.3. as for the model that is represented as follows:

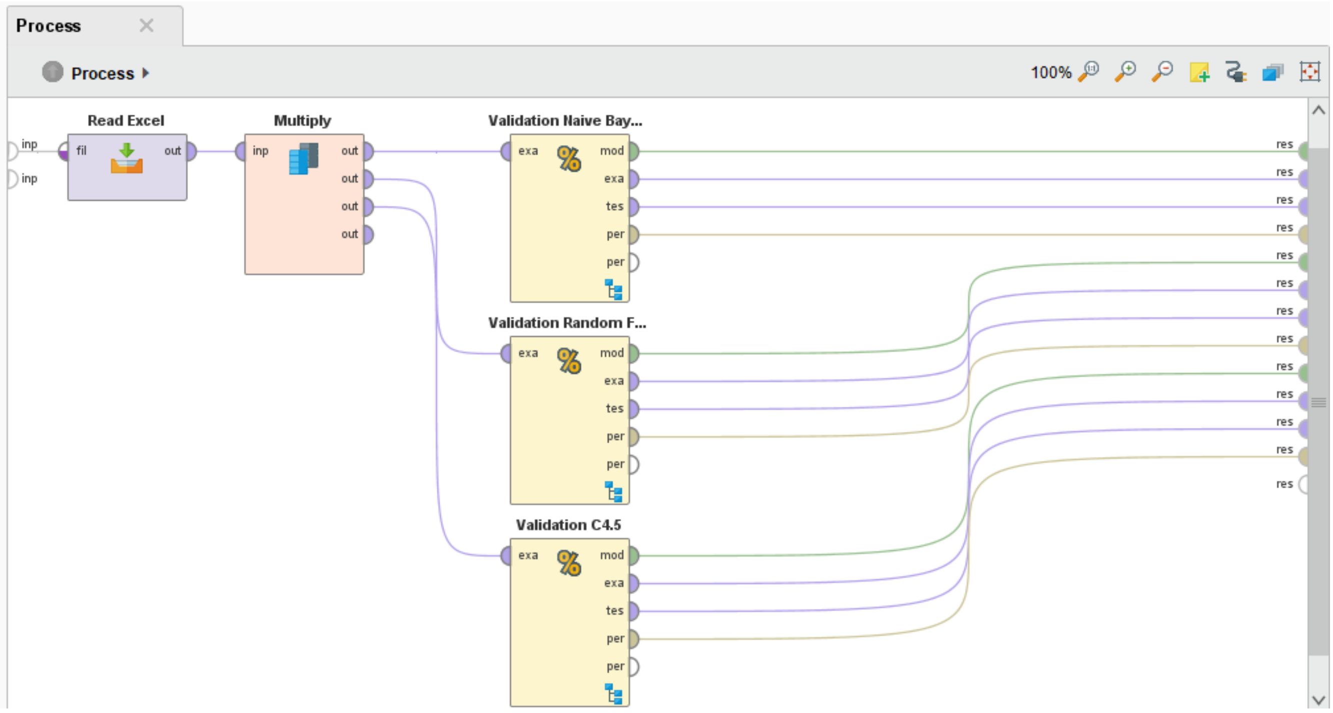

Fig. 2. Algorithm Modeling Naïve Bayes, Random Forest, C4.5.

Shows the process of testing the prediction game online against the learning achievements. Testing is done by the method of cross validation, standards for evaluation is 10-fold cross validation will thus have obtained accurate estimation results.

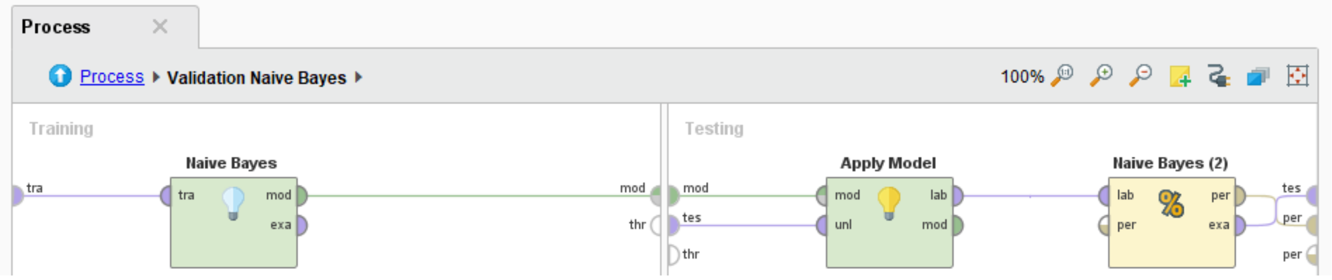

Fig. 3. Algorithm modeling advanced naïve bayes.

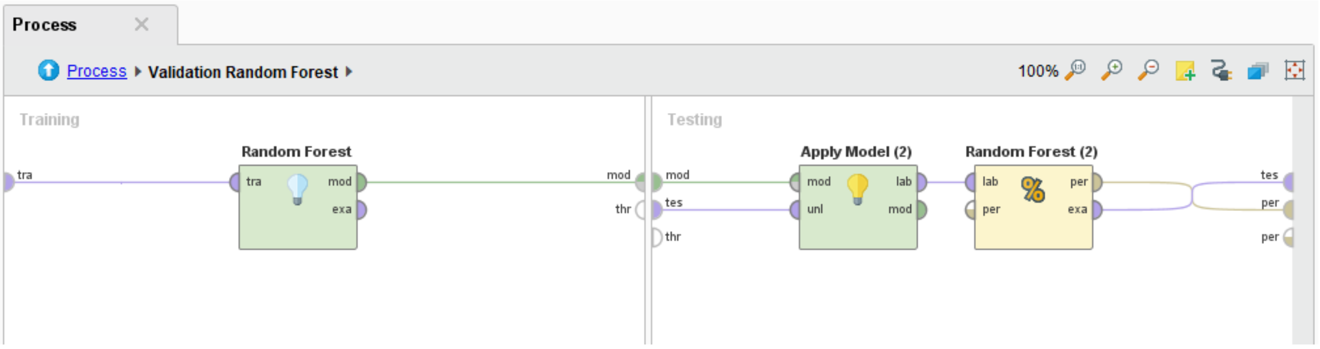

Fig. 4. Algorithm modeling advanced random forest.

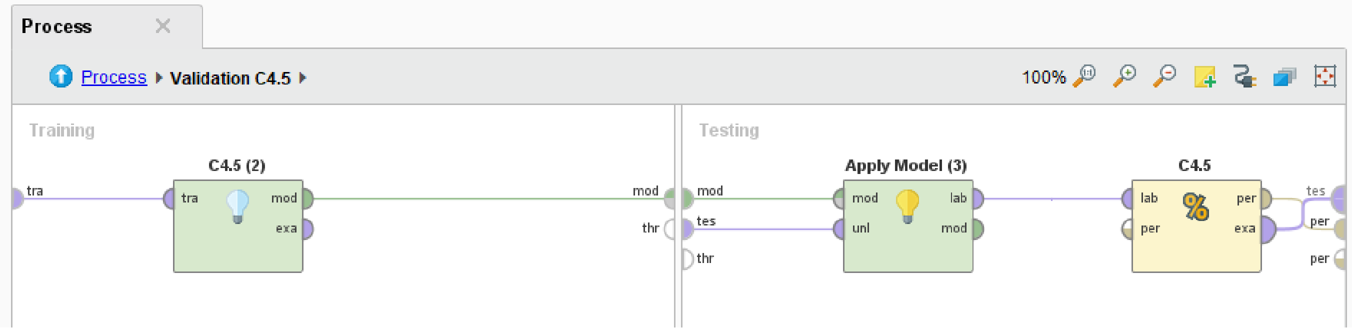

Fig. 5. Algorithm Modeling Advanced C4.5.

Figure 3,4, and 5 shows that the naive bayes algorithm, the random forest and performance testing conducted C4.5 algorithms. So will the accuracy value is obtained from the respective algorithms.

E. Evaluation

After naïve bayes algorithm, random forest, and C4.5 run using rapid miner 7.3 then obtained the results form the value of accuracy so it can know the confusion matrix of each algorithm. confusion matrix aims to find out if the number of cases is positive or otherwise case is negative [27].

TABLE III. CONFUSION MATRIX

Actual Class		**Predicted Class**	
		Positive	***Negative***
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

F. Deployment

Deployment is the stage for the implementation of the results of comparisons of 3 algorithms, then of the three algorithms that have the highest accuracy values can be used as a material development of prediction on online gaming for learning achievement predictions.

IV. RESULT

This chapter describes the results of the research implementation of naïve bayes algorithm, random forest, C4.5 on online gaming for learning achievements. The results of the research are as follows:

A. The Results of the Comparison Algorithm

The table below is the result of comparisons of 3 algorithms used in research, seen that there is a difference between the accuracy and value of the AUC (Area Under Classification).

TABLE IV. THE RESULTS OF THE COMPARISON

Method	**Naïve Bayes**		**Random Forest**		**C4.5**	
	Accuracy	***AUC***	***Accuracy***	***AUC***	***Accuracy***	***AUC***
Cross Validation	70.13%	0.775	64.51%	0.749	67.58%	0.691

B. The Results of the Design of the AUC

This section will look at the design of AUC (Area Under Clasification) of each algorithm.

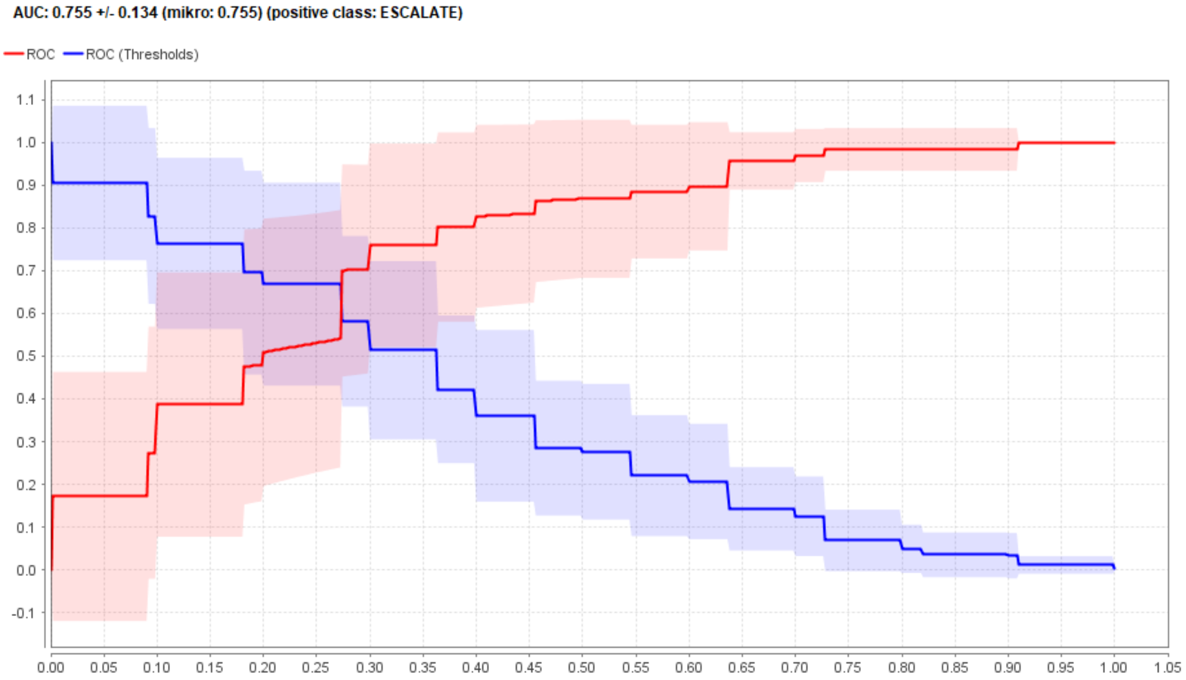

Fig. 6. ROC curves with Naïve Bayes Method.

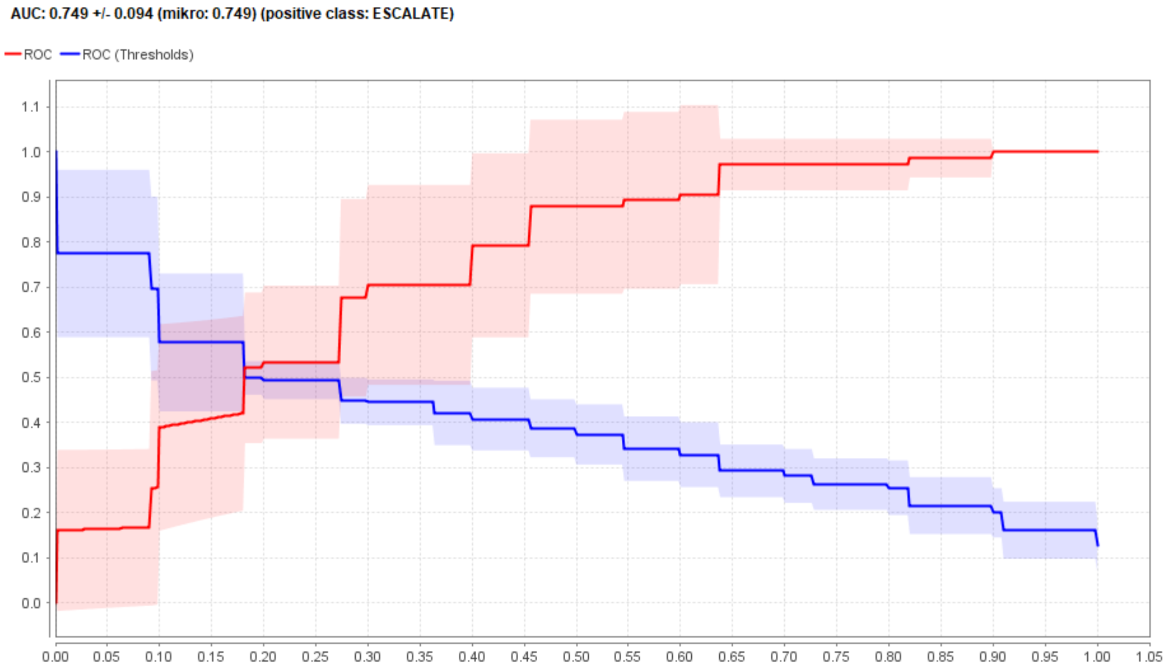

Fig. 7. ROC curves with Random Forest Method

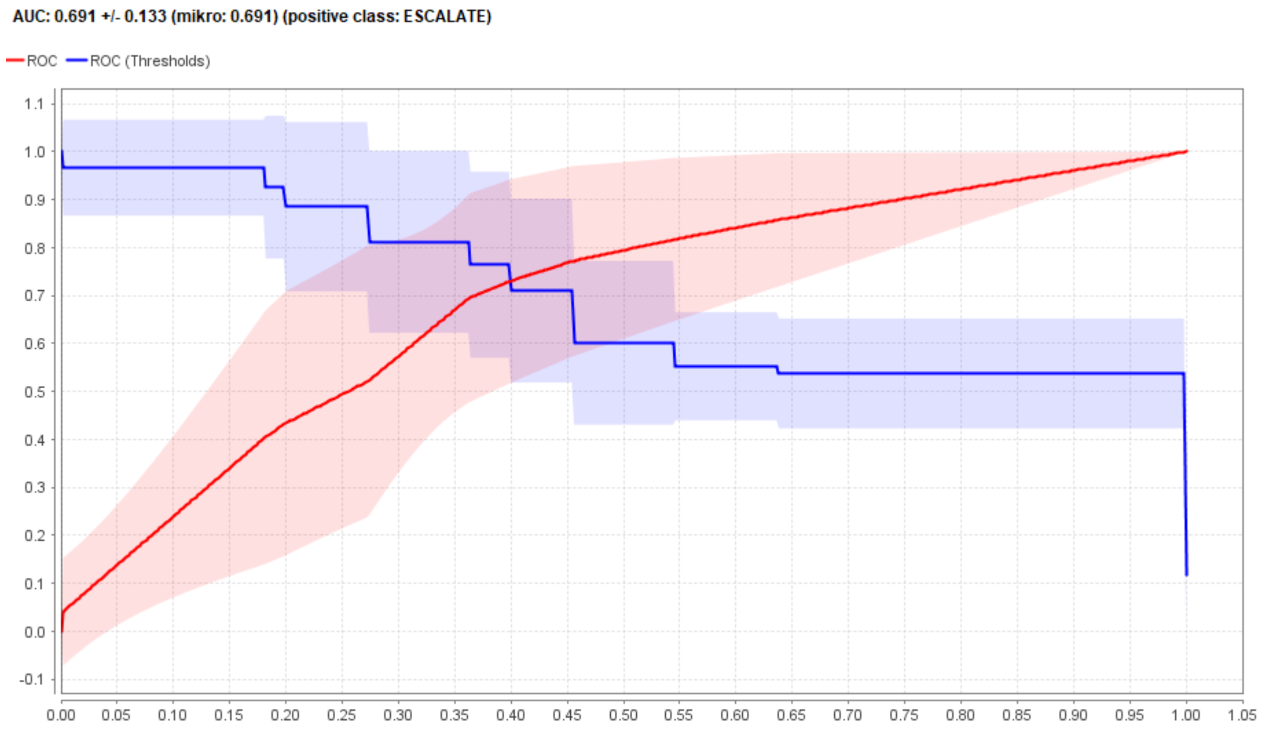

Fig. 8. ROC curves with C4.5 Method.

C. *The Results of the Decision Tree*

1) *Random forest algorithm:*

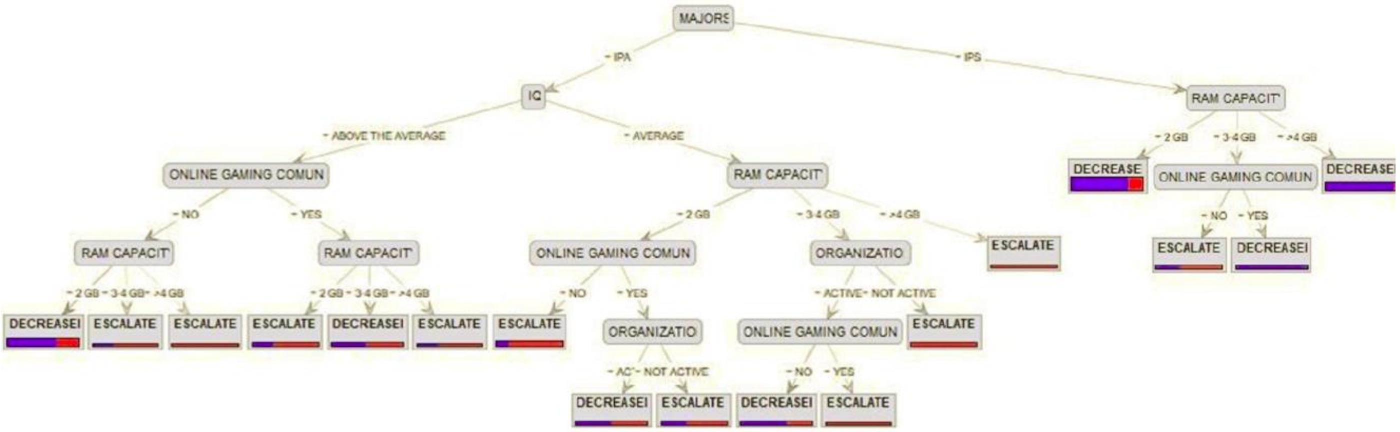

Fig. 9. The results of the random forest.

2) C4.5 algorithm:

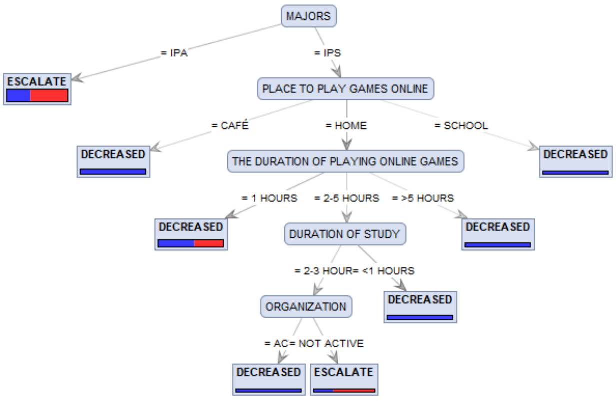

Fig. 10. The results of the C4.5.

The results of a decision tree can be used as a decision making in the future as the evaluation of an activity. From Figure 9 and 10 visible fundamental difference. The algorithm random forest, all attributes as nodes of the decision tree, while for algorithm C 4.5 not all attributes as nodes. This is the difference between the random forest algorithm with C4.5.

D. The Results of the Deployment Algorithms

Deployment on this research is with how to make a prediction accuracy of the algorithm chosen, i.e. naïve bayes algorithm. The deployment process is done through several stages including the following:

- After testing data is done using rapid miner 7.3, there are results in the form of a table of distribution that will be used for the prediction accuracy of the algorithm.
- The results of comparisons of subsequent distribution table again with the previous training data.
- After the comparison is complete then found a new comparison value.

TABLE V. DEPLOYMENT ACCURACY

Naïve Bayes algorithm Deployment accuracy	
Value prediction Accuracy Before deployment	70.13 %
Value prediction Accuracy After deployment	71.78%

From table 5. Seem that difference in the value of the accuracy of the previous values are compared with the values of accuracy, after deployment is 2.6%. The value of accuracy previously 70.13% + / - 6.38%, it means that the value has a value range accuracy accuracy 56.44 s/d 83.82%. Therefore, the process of deployment has the value of accuracy of the selected algorithm.

V. CONCLUSION

From these results it can be concluded that the naïve bayes algorithm has higher accuracy compared with the random forest algorithm and C4.5, making it look the difference in accuracy between the naïve bayes with random forest of 5.62%, whereas the difference in the between the naïve bayes with C4.5 of 2.55%. Naïve bayes algorithm thus can predict the learning achievements of students with better.

Naive Bayes algorithm can be used to predict the relationships students play online games against its value. Naive bayes algorithm implementation using Microsoft excel application in a form that can be used by teachers.

On this research has been getting the expected results, i.e. knowing the algorithm model of accuracy prediction for online games on student achievement. Naïve bayes algorithm proved that value the accuracy and value of 70.13% AUC of 0.775 so that it contains the clasification, fair for the random forest algorithm has an accuracy of 64.51% and AUC values of 0.749 so that it contains the fair clasification, while for algorithm C4.5 67.58% accuracy and value of the AUC of 0.691 so that it contains the poor clasification. From these results it can be concluded that the naïve bayes algorithm has higher accuracy compared with the random forest algorithm and C4.5, making it look the difference in accuracy between the naïve bayes with random forest of 5.62%, whereas the difference in the between the naïve bayes with C4.5 of 2.55%. Naïve bayes algorithm thus can predict the learning achievements of students with better.

of report cards, as well as the data of the questionnaire. The third had never been anyone doing such research.

REFERENCES

[1] R. Syahran, "Ketergantungan Game Online dan Penanganannya," J. Psikol. Pendidik. Konseling, vol. 1, pp. 84–92, 2015.

[2] T. Jap, S. Tiatri, E.S. Jaya, and M.S. Suteja, "The Development of Indonesian Online Game Addiction Questionnaire," PLoS One, vol. 8, no. 4, pp. 4–8, 2013.

[3] I. Beydha, "Game Online dan Prestasi Belajar," pp. 1–10, 2015.

[4] A. Latubessy, "Hubungan Antara Adiksi Game Terhadap Keaktifan Pembelajaran Anak Usia 9-11 Tahun," J. SIMETRIS, vol. 7, no. 2, pp. 687–692, 2016.

[5] R.A. Amanda, "Pengaruh Game Online terhadap Perubahan Perilaku Agresif Remaja di Samarinda," J. Ilmu Komun., vol. 4, no. 3, pp. 290–305, 2016.

[6] N. Husna, E. Normelani, and S. Adyatma, "Hubungan Bermain Games dengan Motivasi Belajar Siswa Sekolah Menengah Pertama (SMP) di Kecamatan Banjarmasin Barat," JPG, Jurnal Pendidik. Geogr., vol. 4, no. 3, pp. 1–14, 2017.

[7] S. Setiaji and S. Viirlia, "Hubungan Kecanduan Game Online Dan Keterampilan Sosial Pada Pemain Game Dewasa," J. Psikol. Psibernetika, vol. 9, no. 2, pp. 93–101, 2016.

[8] D. Rahmawati, D. Mulyana, S. Karlinah, and P. Hadisiwi, "The Cultural Charateristics Of Online Players In The Internet Cafes Of Jabodetabek, Indonesia," J. Theor. Appl. Inf. Technol., vol. 96, no. 7, pp. 1868–1883, 2018.

[9] M. P. Ni and A. Marheni, "Hubungan Kecanduan Game Online dengan Prestasi Belajar Siswa SMP Negeri 1 Kuta," J. Psikol. Udayana, vol. 2, no. 2, pp. 163–171, 2015.

[10] M. Nur, "Pengaruh Keaktifan Berorganisasi, Bimbingan Orang Tua, Kedisiplinan Belajar terhadap Prestasi Belajar Mahasiswa Pendidikan Ekonomi Universitas Kanjuruhan Malang," Jurnal, pp. 4–29, 2015.

[11] V. Effendy and Z.K. Baizal, “Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest,” 2014 2nd Int. Conf. Inf. Commun. Technol., pp. 325–330, 2014.

[12] B. Afeni, T. Aruleba, and I. Oloyede, “Hypertension Prediction System Using Naive Bayes Classifier,” J. Adv. Math. Comput. Sci., vol. 24, no. 2, pp. 1–11, 2017.

[13] L. Dong, X. Li, and G. Xie, “Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification,” Hindawi, vol. 2014, pp. 2–8, 2014.

[14] S. Salmu and A. Solichin, “Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta,” Pros. Semin. Nas. Multidisiplin Ilmu, 2017.

[15] N. Nasution, K. Djahara, and A. Zamsuri, “Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes (Studi Kasus : Fasilkom Unilak),” J. Teknol. Inf. Komun. Digit. Zo., vol. 6, no. 2, pp. 1–11, 2015.

[16] Syarli and A.A. Muin, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus : Data Mahasiswa Baru Perguruan Tinggi),” J. Ilm. Ilmu Komput., vol. 2, no. 1, pp. 1–5, 2016.

[17] H. Widayu, S. Darma, N. Silalahi, and Mesran, “Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5,” Issn 2548-8368, vol.1, June, p. 7, 2017.

[18] A. Andriani, “Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa Studi Kasus : Amik ‘ Bsi Yogyakarta, ” Semin. Nas. Teknol. Inf. dan Komun. 2013 (SENTIKA 2013), vol. 2013, no. Sentika, pp. 163–168, 2013.

[19] L.A. Utami, “Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization,” Pilar Nusa Mandiri, vol. 13, no. 1, pp. 103–112, 2017.

[20] N. Lutfiyana, “Penerapan Algoritma C4.5 Berbasis Particle Swarm Optmization Untuk Prediksi Hasil Layanan Kemudahan,” PILAR, vol. 14, no. 1, pp. 103–110, 2018.

[21] M.S. Brown, Data Mining For Dummies 1st Edition True PDF {PRG}.pdf. Jhon Willey & Sonc Inc, 2014.

[22] J. Han, M. Kamber, and P. Jian, Data Mining : Concepts and Techniques Third Edition. 2015.

[23] H. Basri and R. Eko Indrajit, “Implementasi Information Retrivals Untuk Meningkatkan Pemasaran Produk,” J. Pilar Nusa Mandiri, vol. 13, no. 2, 2017.

[24] W.D. Septiani, “Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis,” Pilar Nusa Mandiri, vol. 13, no. 1, pp. 76–84, 2017.

[25] K.A. Sambodo, M.I. Rahayu, and N. Indriasari, “Klasifikasi Hutan-Non Hutan Data Alos Palsar Menggunakan Metode Random Forest,” Pros. Semin. Nas. Penginderaan Jauh 2014, pp. 120–127, 2014.

[26] D.T. Larose and C.D. Larose, Data Mining and Predictive Analytics. John Wiley and Sons, Inc., Hoboken, New Jersey, 2015.

[27] E.E. Services, Data Science and Big Data Analytics. 14 August 2015, 2015.