COMPARISON OF DATA MINING CLASSIFICATION ALGORITHM FOR PREDICTING THE PERFORMANCE OF HIGH SCHOOL STUDENTS

Tiska Pattiasina[1]**; Didi Rosiyadi**[2]

[1, 2] Masters in Computer Science
STMIK Nusa Mandiri Jakarta, Indonesia
[1]elleoratiska07@gmail.com, [2] didi.rosiyadi@gmail.com

Abstract— *Data Mining is a series of processes to explore added value in the form of unknown information manually from the database. In the world of data mining education can be used to obtain information about student performance. In this study the researchers took research samples from class XI (eleven) students at SMAN 3 Ambon by classifying student performance based on thirteen attributes, namely: age, sex, school organization, extracurricular activities, pocket money, duration of study at home, duration of social media, online game duration, attendance, illness, permits, semester 1 and semester 2 grades. Using the KDD (Knowledge Discovery Database) method and classification algorithm that will be used, namely, decision tree, Naïve Bayes and K-Nearest Neighbor. And then do the test using k-fold cross validation.*

Essence — *In the world of data mining education can be utilized to obtain student performance information. This research presents the results of new decision making using data mining techniques. This study aims to compare the Decision Tree, Naive Bayes, and k-Nearets Neighbor (k-NN) algorithms to improve accuracy in student performance at Ambon State High School 3. And the research method used is the classification method by comparing two algorithms, Naive Bayes and K-nearest neighbor. In this research, the highest accuracy is obtained in the Decision Tree algorithm, which is 99.6047%.*

Keywords*: Data mining, classification, Decision Tree, Naive Bayes, KNN.*

INTRODUCTION

Data Mining is the extraction of important or interesting information or patterns from existing data in a large database (Siregar & Pusphabuana, 2017). And data mining is a mixture of statistics, artificial intelligence, and database research that is still developing (Gorunescu, 2011). Education Data Mining also referred to as EDM is defined as a field of scientific inquiry centered around developing methods to make discoveries in unique types of data that originate from educational settings, and use these methods to better understand students and the settings they learn (Peterson, Penelope L ; Baker, Eva ; McGaw, 2010)

In the world of education research using data mining has been carried out by several researchers to provide increased excellence in higher education, which creates human resources is an ongoing subject. Therefore, the prediction of high school level student performance is very important for continuing education, because the quality of the teaching process can be provided according to student needs. In this case the data and information collected to be able to maintain the quality of students.

Some research in the world of education has been carried out, namely the first research Prediction of Student Performance Using Decision Tree C 4.5 Algorithm, the study aims to predict student achievement based on parents' socioeconomic status, student discipline and student achievement using data mining methods with the Decision Tree, CHAID, Regression algorithm Multiple. Research subjects were students at SD Negeri 4 Trimuloyo. after recording data obtained 352record. Based on data analysis using a decision tree, data mining to predict student achievement based on the socioeconomic status of parents, student discipline and student achievement using data mining methods obtained the following results: economic variables are variables that determine the potential for student success. or not learn achievements in the future. This is evidenced by the variables that become the root node in the decision tree formed. Student achievement variable, is the second important variable in the success of student studies. This shows that aspects of students' knowledge or intelligence are very influential on the success of their learning. Conversely, even though students have less predictable knowledge, high willpower can still be achieved at least in the b or c category. The average C4.5 success algorithm in carrying out classification data reaches 99.43% in accuracy. This shows that this algorithm has a reliable performance in doing classification (Kuntoro & Sudarwanto, 2017)

The second study with the title using data mining to predict secondary school student performance, the beginning of this study discusses the level of education of the Portuguese population which has increased in the last decade but it turns out statistical results make Portugal remain at the tail end of Europe because of the high rate of student failure. In particular, the lack of success in the Mathematics and Potugic core classes is very serious. On the other hand, the field of Business Intelligence (BI) or / Data Mining (DM), which aims to extract high-level knowledge from raw data, offers interesting algorithmic automation that can help the education field. This study aims to approach student achievement in secondary education using the BI or / Data Mining technique. Current real-world data (eg student grades, demographics, social and school features) are collected using school reports and questionnaires. Two core classes (ie Mathematics and Portuguese) are modeled under binary / five-level classification and regression tasks. Also, four DM models (namely Decision Tree, Random Forest, Neural Network and Support Vector Machine) and three input choices (for example with and without previous values) are tested. The results show that good predictive accuracy can be achieved, provided the grades for the first and / or second school periods are available. Although student performance has been strongly influenced by past evaluations, explanatory analysis has shown that there are also other relevant features such as the amount of absenteeism, work and parental education, alcohol consumption (Cortez & Silva, 2008).

The third study, entitled the K-Nearest Neighbor (k-NN) Algorithm Model For Student Graduation Prediction, predicting the rise in the ability of students to complete studies on time is one element of university accreditation assessment This research was conducted to predict student graduation using data mining classifications with processing student data obtained from 1633 student databases, with attributes nim, name, age, faculty, semester 1 to IP semester 8 using the K-Nearest Neighbor algorithm by clustering data k = 1, k-2, k-3, k = 4, and k = 5 (Rohman, 2015).

In this study the researchers took a sample of research from class XI students of SMA Negeri 3 Ambon by classifying student learning outcomes based on age, organization and extracurricular activities followed by school students, pocket money earned by students from parents, duration or the amount of time students used to study, social media, play online games, attendance, permits and illness during 2 semesters, average semester 1 and 2. Using the KDD (Knowledge Discovery Database) method and classification algorithm to be used, namely Decision Tree, Naïve Bayes and K-Nearest Neighbor use weka tools. And then testing is done using k-fold cross validation, then evaluating and validating with a confusion matrix, precision, recall, ROC.

The study aims to compare the Decision Tree, Naive Bayes, and k-Nearets Neighbor (k-NN) algorithms to improve accuracy in student performance at Ambon State High School 3.

MATERIALS AND METHODS

A. Datasets

The data used in this research is real-world data collected from class XI report cards (eleven) with a total of 253 students and questionnaire results that have been filled out by students of SMA Negeri 3 Ambon. Data taken in the form of name, class, age, gender, organization at school, extracurricular, pocket money, duration of study, duration of social media, duration of online games, information on attendance, illness, permission, semester value 1, semester value 2 and label sample raw data can be seen in table 1. The data consists of two labels namely; Very Good (A) with a total of 81 and Good (B) with a total of 172.

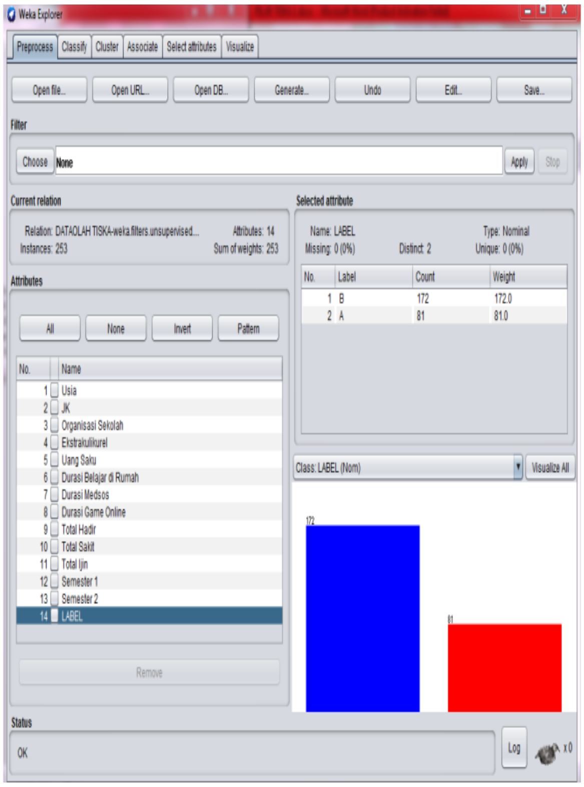
Table 1. Data SMA Negeri 3 Ambon

No	Name	Age	Gen	School organization	Extra-curricular	Pocket money	Duration of study at home	Social media duration	Duration of online games	Total present	Total sick	Tot Permi-ssion	Sem = 1	Sem =2	L
1	Dini Erna Karlina	15	2	2	1	15000	3	2	2	240	0	0	85	86	B
2	Shanty	14	2	2	2	35000	2	3	0	237	3	0	81	81	B
3	Muhammad Joseph	14	1	2	2	20000	1	4	1	240	0	0	81	81	B
4	Fatahilah Bayu	16	1	2	1	15000	1	1	1	240	0	0	85	86	B
5	Akmal A	15	1	2	1	10000	1	2	2	240	0	0	80	81	B
6	Taftaniza	16	2	1	2	12000	4	1	2	236	4	0	83	84	B
7	Giansun	15	2	2	1	20000	1	3	0	237	2	1	80	83	B
8	Mirna	16	2	2	2	10000	1	2	0	240	0	0	81	82	B

9	Ananda	15	1	1	1	15000	2	2	2	237	2	1	80	80	B
10	Ilhamni Zein	15	2	2	2	50000	3	2	0	240	0	0	91	92	A
11	M Rival	16	1	1	2	30000	1	1	1	240	0	0	81	81	B
12	Nissaeryn Z Umayyah	15	2	2	1	15000	2	2	0	240	0	0	91	91	A
13	Novita	15	2	2	2	25000	2	0	0	240	0	0	91	92	A
14	Giannisa R	16	2	2	2	20000	2	1	0	240	0	0	81	82	B
15	Safni Tuara	16	2	1	2	20000	2	1	0	240	0	0	81	82	B
16	Veni Armianti	15	2	2	2	20000	1	2	3	240	0	0	91	92	A
17	Juenda Anaci	15	2	2	2	15000	1	2	0	240	0	0	81	83	B
18	Aldo Sahetapy	15	1	2	1	15000	2	1	1	240	0	0	81	82	B
19	Roy A Souhoka	16	1	2	1	25000	1	2	2	240	0	0	83	84	B
20	Michelle Manane	15	2	1	1	35000	2	1	0	240	0	0	80	81	B
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253	Estrella L Mararessy	15	2	2	2	15000	2	2	0	240	0	0	70	76	B

Source : (Pattiasina & Rosiyadi, 2020)

Following are the attributes that are inputted on the Weka tools can be seen in Figure 1.



Filter: None

Relation: DATAOLAH TISKA-weka.filters.unsupervised... Attributes: 14
Instances: 253 Sum of weights: 253

Name: LABEL Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	B	172	172.0
2	A	81	81.0

No.		Name
1	□	Usia
2	□	JK
3	□	Organisasi Sekolah
4	□	Ekstrakulikurel
5	□	Uang Saku
6	□	Durasi Belajar di Rumah
7	□	Durasi Medsos
8	□	Durasi Game Online
9	□	Total Hadir
10	□	Total Sakit
11	□	Total Ijin
12	□	Semester 1
13	□	Semester 2
14	□	LABEL

Class: LABEL (Nom)

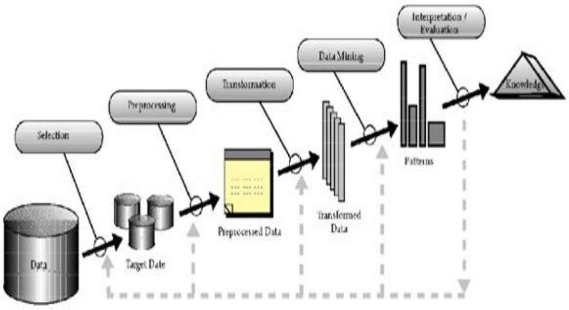
Status: OK

Source: (Pattiasina & Rosiyadi,2020)

Figure 1. Dataset in Weka Tools

B. Method

This study is to classify the learning outcomes of class XI (eleven) students with the research methodology used in this study is to use the Knowledge Discovery in Database (KDD) method consisting of 5 stages: (Sugianto, 2015):



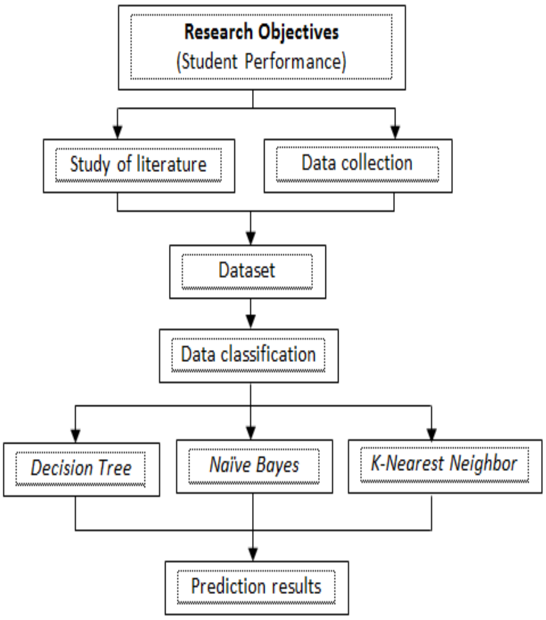
Source : (Sugianto, 2015)

Figure 2. Research Methods

1. Determine the target data which includes data selection and focus more on the subset of data.
2. Cleaning and integration, for data cleaning, which removes noise and inconsistent data, while Integration is combining data from a variety of different sources.
3. Selection and transformation, in data selection that is taking data in accordance with the task of analysis from the database, while Data transformation, which combines data into a form or model suitable for excavation through summary or aggregation operations.
4. Data mining is an important and primary process for extracting patterns from data with more recent methods.
5. Pattern evaluation, which identifies interesting patterns and represents knowledge based on interestingness measures.
6. Knowledge presentation, in the form of presentation of knowledge that is extracted and presented to users using visualization and

knowledge representation techniques (Sugianto, 2015).

Then the research steps that will be carried out can be seen in the picture below:



Source: (Pattiasina & Rosiyadi,2020)

Figure 3. Research steps

In the research steps above can be explained as follows:

a. In accordance with the objectives of the study, which is to implement and compare the Decision Tree, Naïve Bayes and K-Nearest Neighbor algorithms to improve accuracy in student performance at Ambon 3 High Schools.
b. Then do a search of data to the curriculum section of Ambon 3 High School as data to be processed and look for literature studies that are in accordance with research.
c. Data obtained from the curriculum is then processed using the classification method with the Decision Tree algorithm, naive bayes and k-nearest neighbor.
d. Data were tested using validation to get the best accuracy from the algorithm.
e. The results of the algorithm are compared in order to get the best algorithm results, which will be used as a reference in knowing student learning outcomes.

The results of the implementation of the data classification are evaluated and validated so that it can be known how accurate the prediction results will be. Evaluation and validation can be done using confusion matrix and receiver operating characteristics (ROC).

Confusion matrix is a method used to calculate accuracy in the concept of data mining. Evaluation with confusion matrix produces accuracy, precision, and recall values. Accuracy in classification is the presentation of the accuracy of data records that are classified correctly after testing the classification results. Precision or confidence is the proportion of positively predicted mattresses that is also positive true to the actual data. Recall or sensitivity is the proportion of positive cases that are actually correctly predicted correctly (Hand, 2007). This method uses a matrix table as in table 3.

Table 2. Confusion Matrix Model

Class Pred	***Actual***	
	+	**-**
+	*True positive*(A)	*False negative*(B)
-	*False positive*(C)	*True negative*(D)

Source: (Hand, 2007)

The ROC (Receiver Operating Characteristic) curve is another way to evaluate the accuracy of the classification visually (Gorunescu, 2011). The accuracy can be diagnosed as follows:
Accuracy 0.90 - 1.00 = Excellent classification
Accuracy 0.80 - 0.90 = Good classification
Accuracy 0.70 - 0.80 = Fair classification
Accuracy 0.60 - 0.70 = Poor classification
Accuracy 0.50 - 0.60 = Failure

RESULTS AND DISCUSSION

Classification is done to discuss the performance of high school students using weka with the decision tree algorithm, naive bayes, k-nearest neighbor. The classification test is shown by cross validation using 10 fold.

A. Decision Tree

Decision Tree can be seen from the probability that will affect student performance. In the Decision Tree the data is expressed in tabular form with attributes and records. Decision tree process flow is to change the form of table data into a model tree, change the tree into a rule and simplify it. The decision tree algorithm stages are; prepare training data, determine roots and trees, calculate the gain value. The results can be seen in table 3.

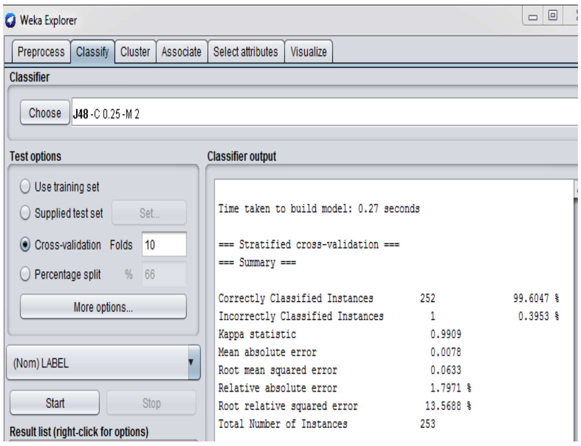
Table 3. Entrophy and gain values

The Knot		B	A	ENTROPY	GAIN
TOTAL	253	172	81	0.90455330	
Age					1.7544107
14 Year	7	6	1	0.59167277	
15 Year	120	85	35	0.87086446	
16 Year	114	73	41	0.94239154	
17 Year	10	8	2	0.72192809	

The Knot		B	A	ENTROPY	GAIN
18 Year	2	0	2	0	
total	253	172	81		
Gender					1.0682747
1. Man	102	68	34	0.91829583	
2. Woman	151	104	47	0.89462059	
total	253	172	81		
School organization					1.5041040
1 Yes	42	28	14	0.91829583	
2. No	211	144	67	0.90168129	
Total	253	172	81		
Extracurricular					1.3034146
1 Yes	72	50	22	0.88797632	
2. No	181	122	59	0.91075254	
Total	253	172	81		
Pocket money					1.1154092
5000 - 10.000	91	57	34	0.95341587	
11.000- 20.000	122	89	33	0.84216948	
21.000 up	40	26	14	0.93406805	
Total	253	172	81		
Duration of study at home					1.0789076
1 Hour	100	68	32	0.90438148	
2 Hour	115	82	33	0.86475726	
3 Hour	25	12	13	0.99884556	
4 Hour	13	10	3	0.77934987	
Total	253	172	81		
Social media duration					1.7335611
No	9	5	4	0.99107606	
1 Hour	89	61	28	0.89841977	
2 Hour	84	59	25	0.87836099	
3 Hour	39	24	15	0.96123665	
4 Hour	32	23	9	0.85714847	
Total	253	172	81		
Duration of game online					0.7057835
No	151	105	46	0.88688405	
1 Hour	48	33	15	0.89603823	
2 Hour	38	30	8	0.74248757	
3 Hour	9	3	6	0.91829583	
4 Hour	7	1	6	0.59167279	
Total	253	172	81		
Total present					1.77582
227 -231	3	2	1	0.91829583	
232-236	12	11	1	0.41381685	
237 - 241	238	159	79	0.91688890	
total	253	172	81		
Total sick					0.0066497
0 – 4	250	169	81	0.90867838	
5 – 9	3	3	0	0	
Total	253	172	81		
Total permission					0.0065187
0 - 6 Kali	252	172	80	0.901598235	
7 - 13 kali	1	0	1	0	
total	253	172	81		
Semester 1 grades					1.5505393
70 – 79	40	40	0	0	
80 – 89	187	132	55	0.873981048	
90 -100	26	0	26	0	
Total	253	172	81		
Semester 2 grades					1.5183595
70 – 79	13	13	0	0	
80 – 89	204	159	45	0.76124015	
90 -100	36	0	36	0	
Total	253	172	81		

Source: (Pattiasina & Rosiyadi,2020)

As can be seen in table 3 that the attribute with the highest gain value is the total present which is 1.77582, then the total present becomes the root node. The following accuracy results can be seen in Figure 4, using the Weka tools with J48 decision tree algorithm.



J48 -C 0.25 -M 2

● Cross-validation Folds 10

```
Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         252               99.6047 %
Incorrectly Classified Instances         1                0.3953 %
Kappa statistic                          0.9909
Mean absolute error                      0.0078
Root mean squared error                  0.0633
Relative absolute error                  1.7971 %
Root relative squared error             13.5688 %
Total Number of Instances              253
```

Source: (Pattiasina & Rosiyadi,2020)

Figure 4. Accuracy results on weka tools

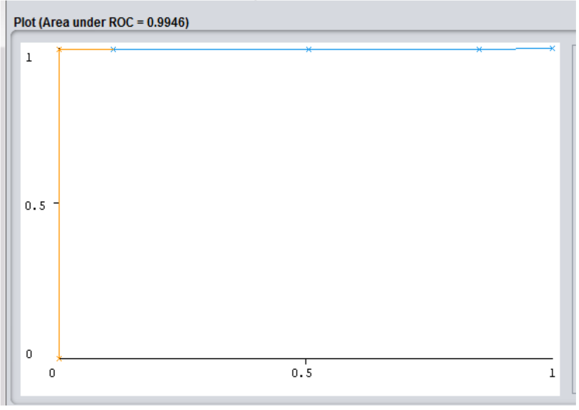
In figure 4 the accuracy in j48 is 99.6047, the configuration matrix can be seen in table 4.

Table 4. Configuration matrix results

	True B	*True A*	*Class Pecission*
Pred. B	171	1	100%
Pred. A	0	81	98.80%
Class recall	99.40%	100%	

Source: (Pattiasina & Rosiyadi,2020)

Following are the results of the ROC graph on the J48 decision tree algorithm in the Weka tools can be seen in Figure 5.



Source: (Pattiasina & Rosiyadi,2020)

Figure 5. ROC results from decision tree j48

The results of ROC processing for j48 decision tree algorithm is 0.9946 with a diagnosis level of Excellent classification.

B. Naive Bayes

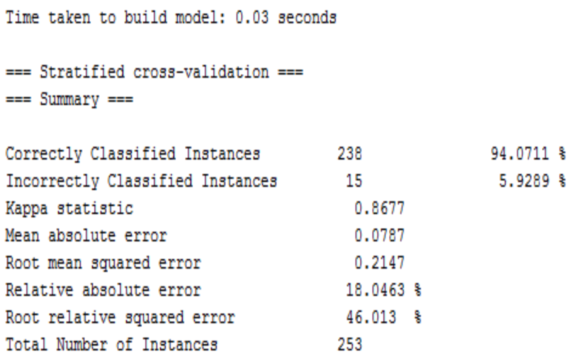
The following classification results using Naive Bayes can be seen in Table 5 using Weka tools with Naive Bayes algorithm.

Table 5. Results of the Naive Bayes classification

Attribute	Class	
	B	**A**
	(0.68)	**(0.32)**
Age		
Mean	15.4826	15.6173
std. dev.	0.6423	0.6771
weight sum	172	81
Precision	1	1
Gender		
Mean	1.6047	1.5802
std. dev.	0.4889	0.4935
weight sum	172	81
Precision	1	1
School organization		
Mean	1.8372	1.8272
std. dev.	0.3692	0.3781
weight sum	172	81
Precision	1	1
Extracurricular		
Mean	1.6977	1.7284
std. dev.	0.4718	0.4448
weight sum	172	81
Precision	1	1
Pocket money		
Mean	16882.69	16372.55
std. dev.	7686.855	7522.873
weight sum	172	81
precision	2647.059	2647.059
Duration of study at home		
mean	1.7907	1.8395
std. dev.	0.8086	0.8234
weight sum	172	81
precision	1	1
Social media duration		
mean	1.9942	1.963
std. dev.	1.0702	1.0823
weight sum	172	81
precision	1	1
Duration of game online		
mean	0.6163	0.9012
std. dev.	0.8716	1.2727
weight sum	172	81
precision	1	1
Total present		
mean	239.222	239.6027
std. dev.	1.7283	1.5401
weight sum	172	81
precision	1.1818	1.1818
Total sick		
mean	0.4622	0.0741
std. dev.	1.2459	0.4015
weight sum	172	81
precision	1.5	1.5
Total permission		
mean	0.189	0.2407
std. dev.	0.7694	1.5226
weight sum	172	81
precision	2.1667	2.1667
Semester 1 grades		
mean	80.5465	88.9037
std. dev.	3.1834	3.3689
weight sum	172	81
precision	1.2	1.2
Semester 2 grades		
mean	82.1618	89.7757
std. dev.	2.513	3.248
weight sum	172	81
precision	1.1667	1.1667

Source: (Pattiasina & Rosiyadi,2020)

The following accuracy results for the Naive Bayes algorithm can be seen in Figure 5.



```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         238               94.0711 %
Incorrectly Classified Instances        15                5.9289 %
Kappa statistic                          0.8677
Mean absolute error                      0.0787
Root mean squared error                  0.2147
Relative absolute error                 18.0463 %
Root relative squared error             46.013  %
Total Number of Instances              253
```

Source: (Pattiasina & Rosiyadi,2020)

Figure 6. Results of NB accuracy on weka tools

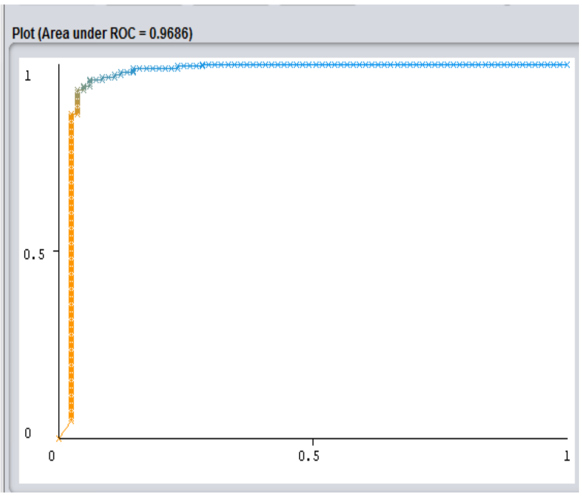
In Figure 6 the accuracy of Naive Bayes is 94.0711, the details of the configuration matrix can be seen in Table 6.

Table 6. Configuration matrix results

	True B	*True A*	*Class Pecission*
Pred. B	160	12	98.20%
Pred. A	3	78	86.70%
Class recall	93.00%	96.30%	

Source: (Pattiasina & Rosiyadi,2020)

Following are the results of the ROC graph on the Naive Bayes algorithm on the Weka tools can be seen in Figure 7.



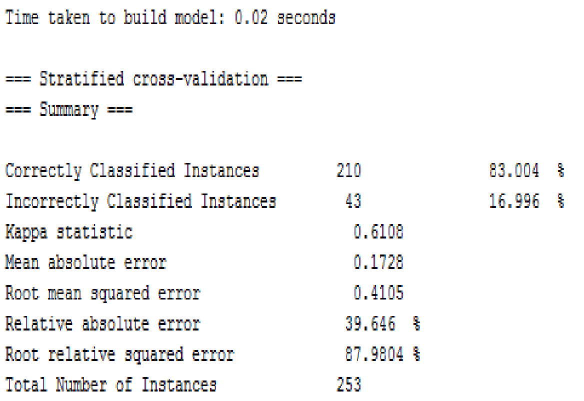
Source: (Pattiasina & Rosiyadi,2020)

Figure 7. ROC results from naive bayes

The results of ROC processing for naive bayes algorithm is 0.9686 with a diagnosis level of Excellent classification.

C. K-Nearest Neighbor

The following results can be seen in the accuracy of Figure 8, using weka tools with k-NN algorithm.



```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         210               83.004  %
Incorrectly Classified Instances        43               16.996  %
Kappa statistic                          0.6108
Mean absolute error                      0.1728
Root mean squared error                  0.4105
Relative absolute error                 39.646  %
Root relative squared error             87.9804 %
Total Number of Instances              253
```

Source: (Pattiasina & Rosiyadi,2020)

Figure 8. Accuracy of k-NN on weka tools

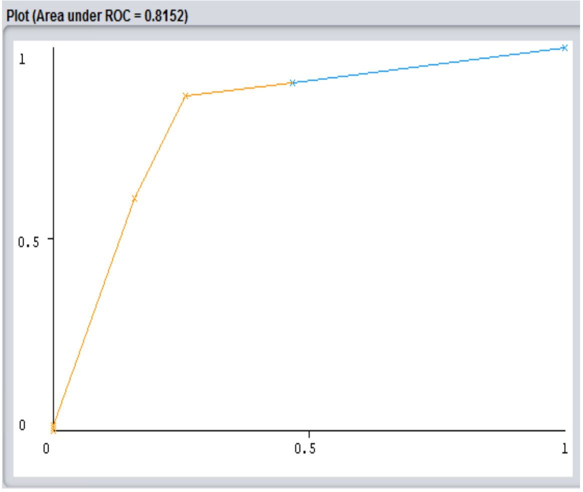
In Figure 8 the accuracy of the K-NN is 83.004, the detailed configuration matrix can be seen in table 8.

Table 8. Configuration matrix results

	True B	*True A*	*Class Pecision*
Pred. B	150	22	87.70%
Pred. A	21	60	73.20%
Class recall	87.20%	74.10%	

Source: (Pattiasina & Rosiyadi,2020)

Following are the results of the ROC graph on the K-NN algorithm on the Weka tools can be seen in Figure 9.



Source: (Pattiasina & Rosiyadi,2020)

Figure 9. ROC results from K-NN

The results of ROC processing for K-NN algorithm is 0.8152 with a diagnosis level of Good classification.

D. Evaluation of Testing Results

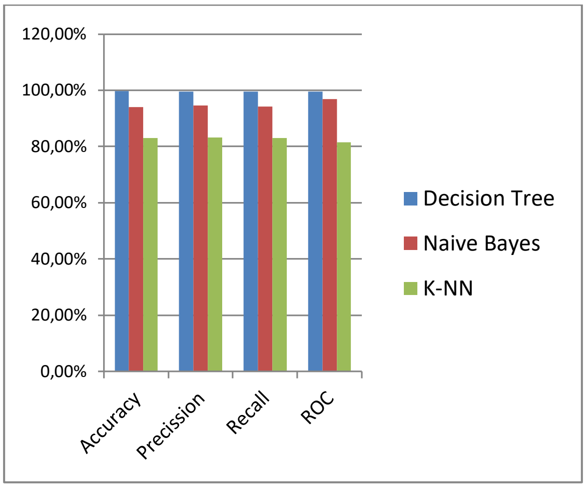
After testing the tools for three algorithms, namely decision tree, naive bayes and k-nearest neighbor, showed different levels of different results. In testing the decision tree algorithm J48 has a higher accuracy that is 99.6047%. Even for other criteria such as Precision and recall, it is still above the two algorithms, namely Naive Bayes and K-NN.

Table 9. Testing DT, NB and k-NN Algorithms

	Decision Tree	*Naive Bayes*	*K-NN*
Accuracy	99.6047%	94.0711%	83.004%
Precission	99.60%	94.50%	83.10%
Recall	99.60%	94.10%	83.00%
ROC	99.46%	96.86%	81.52%

Source: (Pattiasina & Rosiyadi,2020)

The following can be seen in Figure 10, a comparison chart of the prediction results of decision tree, naive bayes and K-NN algorithms.



Source: (Pattiasina & Rosiyadi,2020)

Figure 10. Graph of comparison of algorithm results.

Overall, the J48 decision tree algorithm is one type of algorithm that fits the researcher's data, so that it can obtain a high level of accuracy.

CONCLUSION

This research was taken from class XI students of SMA Negeri 3 Ambon with a dataset of 253 students. There were fourteen attributes used in the form of age, sex, organization involved in school, extracurricular activities, pocket money, duration of study, duration of social media, duration of playing online games, information on attendance, illness, permission, semester grades one and two. After that, use the Knowledge Discovery in Databes (KDD) data mining method. with the classification algorithm Decision tree J48, Naive Bayes and K-Nearest Neighbor. After the data were processed and labeled, 172 students were classified as good (B) and 81 students as very good (A). For accuracy based on the application Weka can be seen as follows: Accuracy using the Decision Tree Algorithm is 99.6047%. The accuracy using Naive Bayes algorithm is 94.0711% and K-Nearest Neighbor is 83.004% so the Decision Tree J48 algorithm has a higher accuracy value than the Naive Bayes algorithm and K-Nearest Neighbor..

REFERENCES

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, *2003*(2000), 5–12.

Gorunescu, F. (2011). No Title. In *Data Mining : Concepts, Models and Techniques. New York: Springer-Verlag.*

Hand, D. J. (2007). Principles of data mining. In *Drug Safety* (Vol. 30, Issue 7). https://doi.org/10.2165/00002018-200730070-00010

Kuntoro, R. K., & Sudarwanto, R. (2017). *Prediction Of Student Performance Using Decision Tree C 4 . 5 Algorithm*. 214–219.

Peterson, Penelope L ; Baker, Eva ; McGaw, B. (2010). A Survey on Feature Selection Methods For Imbalanced Datasets. *International Encyclopedia of Education*. https://www.scholars.northwestern.edu/en/publications/international-encyclopedia-of-education

Rohman, A. (2015). Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa. *Neo Teknika*, *1*(1), 1–9. https://doi.org/10.1017/CBO9781107415324.004

Siregar, M., & Pusphabuana, A. (2017). No Title. In *Data Mining: Pengolahan Data Menjadi Informasi dengan Rapidminer*. https://books.google.com/books?hl=en&lr=&id=rTlmDwAAQBAJ&oi=fnd&pg=PR7&dq=Data+Mining+adalah+serangkaian+proses+untuk+menggali+nilai+tambah+berupa+informasi+yang+selama+ini+tidak+diketahui+secara+manual+dari+suatu+basis+data.+Informasi+yang+dihasilkan+dip

Sugianto, C. A. (2015). Penerapan Teknik Data Mining Untuk Menentukan Hasil Seleksi Masuk Sman 1 Cibeber Untuk Siswa Baru Menggunakan Decision Tree. *Tedc*, *9*, 39–43.