Jurnal Mantik
Volume 4, Number 1, May 2020, pp. 441-451 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5

Lili Dwi Yulianto[1], *Agung Triayudi[2], Ira Diana Sholihati[3]

Sistem Informasi
Universitas Nasional, Universitas Nasional, Jl. Sawo Manila, RT.14 / RW.3, Ps. Sunday, Kec. Ps. Minggu, Kota Jakarta Selatan, Special Capital Region of Jakarta 12520

Email: [1]lilidwianto@gmail.com, *[2]agungtriayudi@civitas.unas.ac.id, [3]iradiana2803@gmail.com

ARTICLE INFO

Article history:
Received: 04/14/2020
Revised: 15/05/2020
Accepted: 25/01/2020

Keywords:
Educational Data Mining (EDM), Decision Tree C4.5, K-Nearest Neighbor

ABSTRACT

Data Mining is very useful and widely applied in various fields, one of which is in the field of education. Data mining can be applied to the field of Educational institutions and is also often referred to as Educational Data Mining (EDM). Colleges currently compete with each other to improve the quality of their education for the creation of competent and high-integrity human resources. The amount of data contained in Higher Education can be utilized for the need for useful information so that the data attributes can be known so that the data is analyzed to improve student performance and achievement. And also the results of the analysis are expected to be able to anticipate the problem of delays in the study period that is often experienced by students. In this study conducted using two algorithm models namely K-Nearest Neighbor and Decision Tree C4.5. The best accuracy value is the K-Nearest Neighbor algorithm model with an accuracy rate of 59.32%, whereas in the Decision Tree C4.5 model the accuracy rate is 54.80%, the application of EDM and is expected to be maximized and developed so that it can contribute and develop in education world especially in data mining

1. Introduction

Data Mining, commonly called KDD (Knowledge Discovery in Databases), is an effective branch of knowledge for data processing, where data is processed to have a very large capacity, data growth is very fast, and has different data forms and formats. the application of data mining has a big hand in handling various kinds of cases one of the applications is in the world of education, namely universities. [1]

Knowledge Discovery In Database (KDD) is a method for obtaining knowledge from existing databases. In the database there are tables that are interconnected / related. The results of knowledge gained in the process can be used as a knowledge base for decision making. The terms Knowledge Discovery in Database (KDD) and data mining are often used interchangeably to explain the process of extracting hidden information in a large database. Actually the two terms have different concepts, but are related to each other, and one of the stages in the whole KDD process is data mining. [2]

Data Mining is very useful and widely applied in various fields, such as marketing analysis, medicine, manufacturing engineering, education and others. Data mining can be applied to the field of educational institutions or institutions and is often referred to as Educational Data Mining (EDM), which is a development of methods in exploring the types of education data types that are unique in order to learn in understanding student performance and environmental settings in the place where students learn, As for one way in higher education that is by using educational data and digging knowledge in it so that it can be found about the main attributes that can affect the quality of student performance, the abundance of data in higher education can be used optimally in accordance with needs and can be processed into useful information so that can know the relationship between the data attributes in which can be analyzed and expected to have an output in the form of student performance related to the study period that can be categorized as appropriate or late in studying. Educational Data Mining (EDM) can be applied with many methods such as using Decision Tree techniques, Artificial Neural Networks, Naive Bayes Classifier, and others [3]

Research in the field of education has been carried out by (Ms.Tismy Devasia, Ms.Vinushree TP, and Mr.Vinayak Hegde, 2016) with predictive analysis of student performance using the Naive Bayes Classifier (NBC) Method while the results of the research are to classify 60 students and produce 4 attributes in it so

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

that concludes the Naive Bayes Classifier (NBC) method gives good accuracy results. This research is intended to provide facilities for students and lecturers to improve student performance at the beginning of a shift. [4]

Research conducted by (Mokhairi Makhtar, Hasnah Nawang, and Syadiah Nor Wan Shamsuddin, 2017) conducted a student performance analysis with a total data of 488 students with attributes of 7 subject values and a comparison of grouping algorithms namely Naive Bayes Classifier, Random Tree, Nearest Neighborhood (IB1), Multi Class Classifier, Conjunctive Ruleyang produce output in the form of the effect of the course scores on student performance and produce the Naive Bayes Classifier (NBC) as the algorithm with the best accuracy value compared to the other four algorithms. [5]

Subsequent research conducted by (Yoga Pristyanto, Irfan Pratama, and Anggit Ferdita Nugraha, 2018) conducted a test to classify data whose data models were unbalanced by comparing the methods of Synthetic Minority Over-Sampling Technique (SMOTE) and One Sided Selection (OSS)), in his research the SMOTE method has a high degree of accuracy compared to the OSS method that is equal to 85.505%. [6].

Maulana Aditya Rahman, Nurul Hidayat, Ahmad Afif Supianto, 2018 in his research applied the concept of data mining to classify the quality of clean water in PDAM Tirta Kencana in Jombang Regency and to compare two models namely K-Nearest Neighbor with Naïve Bayes resulting in the overall average value of the method K-Nearest Neighbor is 82.45% and the Naive Bayes method is 72.52%. While testing based on the amount of training data the K-Nearest Neighbor method has an average overall accuracy of 83.32% and the Naive Bayes method of 70.91%. Based on the accuracy obtained from all the K-Nearest Neighbor testing methods is the best method with a total average accuracy of 82.89%. [7]

While the results of research conducted by Galih, 2019 analyzed Student Performance with the Comparison of 2 Classification Models, namely Naive Bayes Classifier (NBC) and C4.5 algorithm. Based on testing on the 2 algorithm models using training data ratios can affect the results of each accuracy value, while the best accuracy value on the Naive Bayes Classifier (NBC) algorithm model is 86.83% with a training data ratio of 80%, whereas in the model C4.5 algorithm is 88.10% with a training data ratio of 90%. [8]

Based on previous research, the use of EDM is mostly done using algorithms including Naive Bayes Classifier (NBC), Decision Tree, One-Sided Selection (OSS), K-Nearest Neighbor, Decision Tree C4.5 and others. Resulting in the Decision Tree C4.5 and K-Nearest Neighbor algorithm has the best accuracy value. so it can be concluded to conduct research using the two algorithms namely Decision Tree C4.5 and K-Nearest Neighbor to compare which algorithm is the best so that it becomes the foundation and entry for the Faculty of Communication and Information Technology of the National University so that it can be applied in the academic field can help the teaching-learning process and also minimize the delay in student studies

2. Research Methods

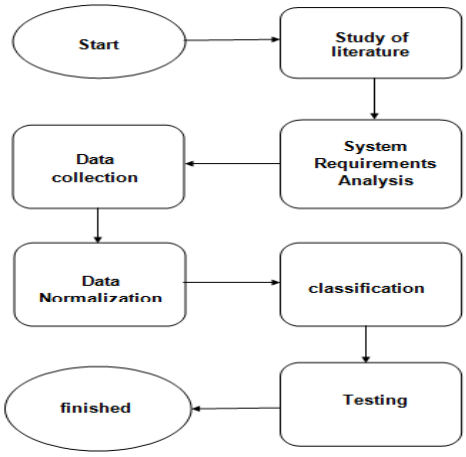


Fig 1. Research Flowchart

At the research stage shows the process carried out in this study. Several stages, starting from:

a. Literature Study
   At this stage the authors take 5 National journals and 10 International journals as references and references in conducting research.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

b. System Requirements Analysis
So that the research process can run optimally, there are several things that are needed, namely:
1) Hardware requirements
The hardware used is the AMD A6 3.9Hz processor, a minimum of 4GB RAM, a hard disk with a capacity of 50GB or more, VGA and a monitor with a minimum resolution of 800x600 pixels, Mouse, Keyboard and LAN Card.
2) Software requirements
Software needed is Microsoft Word 2007, Microsoft Exel 2007 for writing and recapitulation of data, python and rapid miner as a tool to support the data classification process.

c. Data Collection
The author uses several data collection techniques for research purposes predicting student performance. Some things done in data collection such as:
1) Questionnaire (Questionnaire)
Making a set of questions or written statements to respondents to answer, the questionnaire provided is semi-open meaning the choice of answers has been given by the author, but the object of research is still given the opportunity to answer according to their wishes.
2) ü Literature study
Gather references from the literature that can be taken into consideration to help analyze student performance using the classification method. From each of these literature, the author determines some existing data to be used in order to facilitate the retrieval of the right data with the attributes that will be used

d. Data Normalization
Dataset in the study must be simplified in advance so that it can be used in accordance with the proposed method, while in the initial processing of data are as follows.

Table 1.
Data Normalization

Attribute	Description	value
NIM	Student ID Number	ID
Name	Full Student Name	String
Religion	Student Religion	Islam, Kristen, Katolik, Protestan, Budha
Gender	Gender, Male or Female Students	Male or Female
The origin of the school	chool Category of Origin Students before continuing to college	High School, Vocational High School, Madrasah Aliyah, Home Schooling
domicile	Explain the Origin of Student Housing	Bandung, Jakarta, Bogor, Ciawi, Ciamis, Bekasi, Sukabumi, Tasikmalaya, dll
Distance of Student Residence to Campus	Distance of Student Residence to Campus is <10km or> 10km.	< 10km or >10km.
how many siblings	Total Siblings in the family	1,2,3,4 >4
Parents' job	Employment status of parents or guardians of students	Petani, PNS, Polri, TNI, Swasta, Wiraswasta, Karyawan
IPK	Indeks Prestasi Komulatif	0,00 to 4,00
Vehicle	Number of vehicles owned by student families	Motorcycle, Car
Scholarship	Information on whether students received scholarships or not	Yes or not
Study time	Special Study Time for one day	1,2,3,4 >4
Graduated Status	Passed 7.5 Semester, 8 Semester and Not Yet Graduated	Exactly, Too Late, Not Yet Graduated

In the table above is a database of students at the National University to be selected and reduced data, the data has different attributes and records that need to be adjusted to the needs of research. For the data elimination stage, some attributes will be deleted data that is considered not very influential on the process of classifying data, namely in the NIM column, name, and Religion. and data elimination will also be done if there is one field that is not filled in by students or many fields are empty so that the field must be deleted. After the initial data processing (preprocessing) is produced valid data with a total of 588 data.

e. Classification

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The proposed model will use 2 classification techniques and make comparisons on both Decision Tree C4.5 and K-Nearest Neighbor to produce the highest accuracy value using the PHP programming language and additional data mining tools namely Rapid Miner Studio 9.2,000.

Based on previous research, the use of EDM is mostly done using algorithms including Naive Bayes Classifier (NBC), Decision Tree, One-Sided Selection (OSS), K-Nearest Neighbor, C4.5 algorithm and others. So that the Decision Tree C4.5 and K-Nearest Neighbor have the best auration value. So it was decided to make a comparison with the two algorithms and choose the best results.

f. K-Nearest Neighbor Algorithm

The working principle of the K-Nearest Neighbor (KNN) algorithm is to find the closest distance to the nearest neighbor in the training data with the data to be tested. The technique of grouping new data is by calculating the distance of new data to the nearest data / neighbor. The K-Nearest Neighbor algorithm is instead based learning, where training data is stored so that the classification for new records that have not been classified can be found by comparing the most similarities in training data (Kustiyahningsih, 2013).

$$euc = \sqrt{\sum_{i=1}^{n}(x_{2i} - x_{1i})^{2}}$$

Information:

- ✓ X2 = training data
- ✓ X1 = test data
- ✓ i = data variable
- ✓ n = data dimension

In the training phase, this algorithm only stores feature vectors and classifies training sample data. In the classification phase, the same features are calculated for testing data (whose classification is unknown). The distance of this new vector to all the training sample vectors is calculated and the closest number of k is taken. The new points are predicted to be included in the most classifications of these points.

g. Decision Tree C4.5

The decision tree models the top-down sample classification, which starts from the root node (root) by keeping a distance from the results of the internal node test, until the leaf node is reached by the assigned label class. The most important advantage of the Decision Tree or Decision Tree is that knowledge can be separated or extraction process carried out and can be represented in the form of classification rules, if-then branching (Yu, Huang, Hu, & Cai, 2010)

Decision trees use a hierarchical structure for supervised learning. The process of the decision tree starts from the root node to the leaf node which is done recursively. Where each branching states a condition that must be met and at each end of the tree states the class of data. The process in the decision tree is to change the shape of the data (table) into a tree model (tree) and then change the tree model into rules.

The steps that must be taken to form a decision tree with the C4.5 algorithm are as follows:

1) Prepare Training Data Samples
   Training Data is historical data that has been previously processed and has been formed into several classes.
2) Forming Tree Roots
   In making the root of the data tree obtained from the attributes that have been processed, which is done by calculating the gain values that exist in each attribute. if the calculation results obtain the highest gain value it will be placed in the first root position
   Before determining the gain value, first determine the value of entropy which is a probability distribution on the information theory used by the Decision Tree (C4.5)
   in determining the level of similarity (homogeneity) of the class distribution of a dataset. If a dataset has a high value of entropy, the more homogeneous the class distribution. To calculate the entropy value the formula is used:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log2pi$$

Information:

- ✓ S = dataset case

- ✓ n = the number of partitions S
- ✓ pi= proportion of Si to S

After that to calculate the Gain value with the equation:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Information :

- ✓ S= himpunan (dataset) case
- ✓ A= features
- ✓ n= the number of partitions attribute A
- ✓ |Si|= proportion of Si to S
- ✓ |S|= the number of cases in S

h. Testing and Evaluation

a) Cross Validation

Cross Validation is one technique to assess / validate the accuracy of a model that is built based on a specific dataset. Validation is also a standardized test carried out to predict error rates.

b) Confusion Matrix

Confusion matrix is a method that is usually used to calculate accuracy in the concept of data mining. This formula performs calculations with 4 outputs, namely: recall, precision, accuracy and error rate. Evaluation of the classification model is based on testing to estimate the right and wrong objects (Wu, 2009)

Table 2.
Prediction Classes

Classification	*Predicted Class*	
	Prediction = Yes	Prediction = No
Actual = Yes	a (*true positive* - TP)	b (*false negative* - FN)
Actual = No	c (*false positive* - FP)	d (*true negative* - TN)

The formula used in calculating the confusion matrix model equations to produce accuracy, recall and precision values:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall\ (positive) = \frac{TP}{TP+FP}$$

$$Precision\ (positive) = \frac{TP}{TP+FN}$$

$$Recall\ (negative) = \frac{TN}{FN+TN}$$

$$Precision\ (negative) = \frac{TN}{FN+TP}$$

3. Results And Discussion

The results of the data collection process carried out by means of the questionnaire (questionnaire) and literature study, then obtained a student dataset, in the form of training data of 588 data,

Table 3.
Student training data

No	**Name**	**GPA**	**Study time**	**Scholarship**	**Semester**
1	Timoteus Loy	3,41	1	Not	Right
2	Januarius	3,41	1	Not	Right
2	Josua Dean	3,2	2	Not	Right
3	Gilang Ramadhan	3,2	2	Not	Right
5	Indah Silvia	3,76	2	Not	Right
6	Frederica Ristaruli Felicia	3,87	2	Not	Right
7	Hodijatul Afifah	3,51	4	Not	Right

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

No	Name	GPA	Study time	Scholarship	Semester
8	Khaofah Hasanah	3,61	4	Not	Right
9	Muhammad Fauzi	3,2	5	Not	Right
10	Gatasa Ruswandy	3,76	5	Not	Right
11	Muhamamad Damar	3,76	5	Not	Right
…	…	…	…	…	…
588	Riska Afiska	2,89	9	Yes	Late

A. Decision Tree c4.5

The results of the student training data, then performed calculations using the C4.5 Algorithm method, in order to get the type of classification Based on the training data, the results obtained total outlook attributes, namely:

Table 4
Node 1 (root)

Atribut OutLook Total (root)		
Jumlah Data	[S]	586
Jumlah Data 7,5 Semester (Tepat)	[S1]	269
Jumlah Data 8 Semester (Terlambat)	[S2]	317
Entrophy Total		0.995154725

The student training data contained discrete data or continuous data, then the grouping process is required to make split values on student training data, starting from sorting the smallest value to the largest value and then calculating the value of entropy, gain, split info and gain ratio based on the humadity attribute.

Table 5
IPK humadity attributes

Humidity	Amount of data	7,5 |S1|	8 |S2|	Entrophy	Gain	Split Info	Gain Ratio
					0.994777788	0.032788881	30.33893867
<=2,89	2	0	2	0			
>2,89	586	269	317	0			
					1.98503834	0.04619244	42.9732297
<=2,91	3	0	3	0			
>2,91	585	269	316	0.995338811			
					1.983525477	0.058758157	33.75744875
<=2,98	4	0	4	0			
>2,98	584	269	315	0.995519933			
					1.98048897	0.082143051	24.11024351
<=2,99	6	0	6	0			
>2,99	582	269	313	0.995873153			
...	...	...	...	...	...	...	...
					0.001466565	0	0
<=3,99	582	263	319	0.993311223			
>3,99	0	0	0	0			

From the IPK table above it is known that the attribute that has the highest gain is found at the threshold value of IPK <= 2.91,> 2.91, that is 1.98503834.

Table 6.
Attributes humadity study time

Humidity	Amount of data	7,57 |S1|	8 |S2|	Entrophy	Gain	Split Info	Gain Ratio
					1.983777014	0.04619244	42.9459238
<=1	3	3	0	0			
>1	585	266	319	0.994071017			
					1.91517705	0.238222664	8.039440993
<=2	23	6	17	0.828055725			
>2	565	252	313	0.991575305			
					1.894819664	2.83500152	6.683663652
<=4	29	21	8	0.849751137			
>4	559	248	311	0.990818257			
					1.84446297	0.377494363	4.88606758
<=5	43	28	15	0.933025295			
>5	545	241	304	0.990339403			
					1.749872705	0.526617066	3.322856054
<=6	70	38	32	0.994693795			
>6	518	231	287	0.99155285			
					1.527254732	0.780257655	1.957372314
<=7	136	65	71	0.998595537			

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Humidity	Amount of data	7,57 |S1|	8 |S2|	Entrophy	Gain	Split Info	Gain Ratio
>7	452	204	248	0.99315362			
					1.009878719	0.999699543	1.010182236
<=8	288	138	150	0.998747298			
>8	300	131	169	0.988395231			
					0.063774986	0.261306657	0.24406185
<=9	562	266	296	0.997943537			
>9	26	3	23	0.51594693			
					0.035870016	0.14372617	0.249571921
<=10	576	266	310	0.995786645			
>10	12	3	9	0.811278125			
					0.016595857	0.070699727	0.234737214
<=11	583	267	316	0.994898334			
>11	5	2	3	0.970950595			
					0.006820588	0.032788813	0.208015694
<=12	586	268	318	0.994742037			
>12	2	1	1	1			

From the IPK table above it is known that the attribute that has the highest gain is found in the learning time threshold <= 1,> 2 which is 1.91517705. Thus the two attributes will be entered into a calculation table to make a decision tree.

Table 7.
Calculation Results

Atribut	Value	Jumlah Kasus	7.5	8	Entropy	Gain	Split Info	Gain Ratio
Total		588	269	319	0.994777788			
Outlook						1.151882812	0.985228136	1.169153387
	Ya	20	17	2	0.485446076			
	Tidak	568	252	316	0.990822371			
Beasiswa						1.937815867	0.236131537	9.500870411
	<=1	3	3	0	0			
	>1	585	266	319	0.994071017			
Waktu Belajar						1.983777014	0.04619244	42.9459238
	<=2,91	3	0	3	0			
	>2,91	585	269	316	0.995338811			
IPK						1.98503834	0.04619244	42.9732297

From the table above it can be seen that the attribute that has the highest gain ratio is the GPA attribute with 42.9732297. thus the IPK attribute can be used as a root node. there are two values in the node, namely <= 2.91 and> 2.91, the attribute <= 2.91 in the collectibility of classifying cases is 8 Semesters.

Next to continue making branches in the decision tree look for nodes in the study time and scholarship attributes by filtering based on the GPA attribute value> 2.91, so that the learning time attribute is obtained as the second branch. there are two values in the node, namely <= 1 and> 1, the attribute <= 1 in the collectibility of classifying cases is 7.5 Semesters.

Next to continue making branches in the decision tree look for nodes in the scholarship attribute by filtering based on the value of the study time attribute> 1, so that the scholarship attribute is obtained as the last branch. there are two values in the node, namely yes and no, the attribute yes in classifying cases is 7.5 Semesters while the attribute not in classifying cases is 8 Semesters. From the results of these calculations can be drawn a temporary decision tree that looks like the following picture 3:

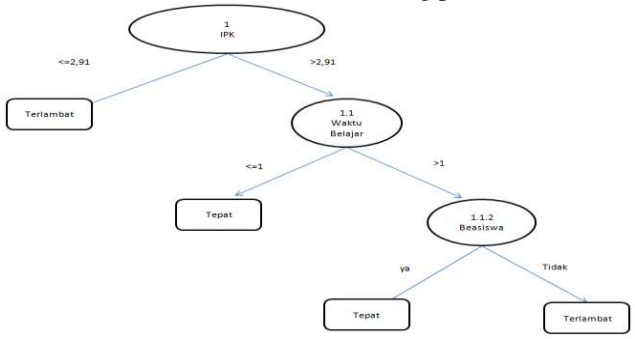


Fig 2. Decision Tree

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

B. K-Nearest Neighbor

Data Dissemination of students on the GPA and Learning Time,

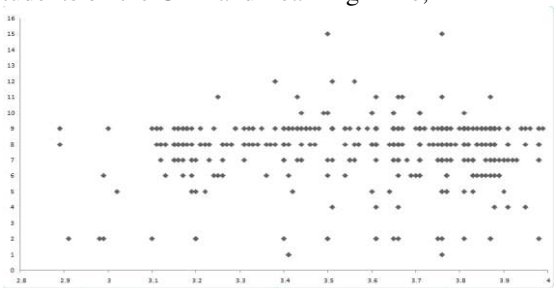


Fig 3.Distribution of Students

By calculating the Euclidean distance square (query instance) of each object to the given training data, it can be seen in table 3.6, while the k value used in this case is k = 3

Table 8.
Euclidean Distance

No	Rank	Euclidian Distance	Beasiswa	Tahun
1	575	5.005756686	Tidak	Tepat
2	575	5.005756686	Tidak	Tepat
3	581	5.034689663	Tidak	Terlambat
4	540	4.000112498	Tidak	Terlambat
5	540	4.000112498	Tidak	Terlambat
6	540	4.000112498	Tidak	Terlambat
7	546	4.006607043	Tidak	Terlambat
8	550	4.013589416	Tidak	Tepat
9	550	4.013589416	Tidak	Terlambat
10	556	4.024127235	Tidak	Terlambat
11	558	4.02869706	Tidak	Terlambat
12	567	4.041831268	Tidak	Terlambat
13	568	4.043278373	Tidak	Tepat
14	569	4.050876448	Tidak	Tepat
15	571	4.060788101	Tidak	Terlambat
16	571	4.060788101	Tidak	Tepat
43	40	1.001798383	Tidak	Tepat
44	41	1.004041832	Tidak	Tepat
45	42	1.009752445	Tidak	Tepat
46	44	1.026109156	Tidak	Tepat
47	44	1.026109156	Tidak	Tepat
...	...	...	...	...
588	587	9.006047968	Tidak	Tepat

Normality Test Table at K = 3

No	Rank	Euclidian Distance	Beasiswa	Tahun
1	1	0	Tidak	Tepat
2	2	0.02	Tidak	Tepat
3	3	0.04	Tidak	Terlambat

C. Testing Accuracy and Comparison

The following accuracy testing will be performed using the RapidMiner 9.6 application between the Decision Tree c4.5 method and the K-Nearest Neighbor method. The calculation will be done with 25 test data that has been tested and 588 datasets to determine how much the level of accuracy and will be compared with the aim of whether the method used is greater accuracy with other methods or not.

1) Testing the accuracy of the Decision Tree c4.5 method with Rapid Miner

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

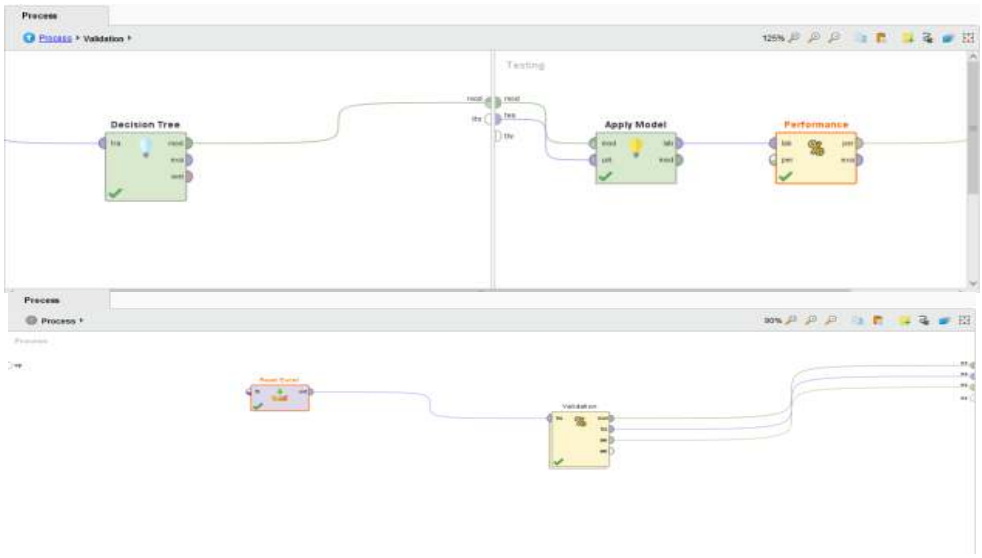


Fig 4. Design of the Decision Tree model c4.5

In Figur 3.3, the first test is carried out with the Decision Tree c4.5 model in which there are dataset and test data, where the test data will be processed according to the existing dataset with the Applicable Model and then conducted training and testing on the Performance test to find out the accuracy of the method .

2) Testing the accuracy of the K-Nearest Neighbor method with Rapid Miner

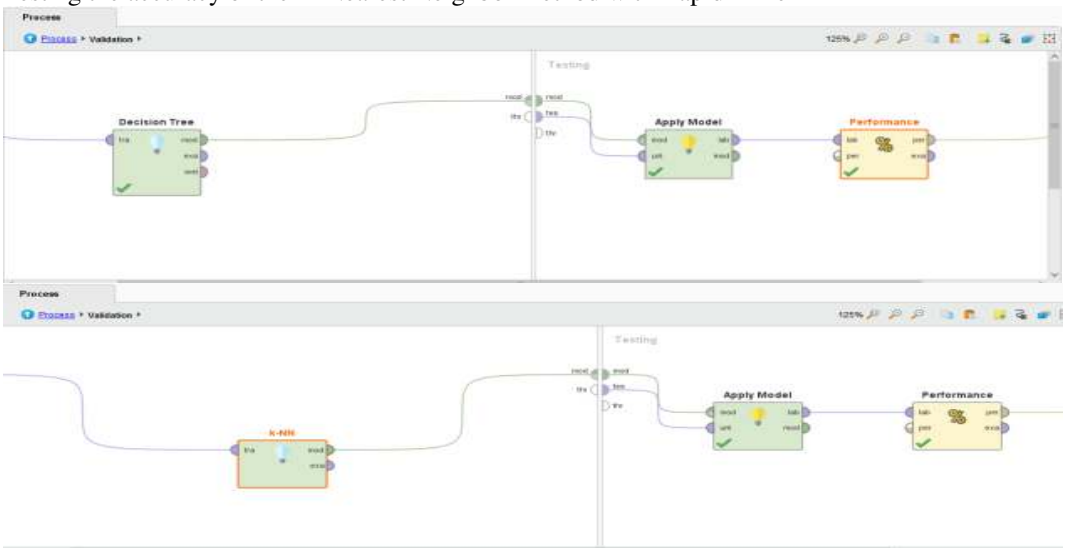


Fig 5. K-Nearest Neighbor model design

In Figur 5, the second test is done with the K-Nearest Neighbor model. in which there are datasets and test data, where the test data will be processed according to the existing dataset with the Applicable Model then training and testing are performed on the Performace test to find out the accuracy of the method.

Table 9.
Comparison of Two Algorithm Models

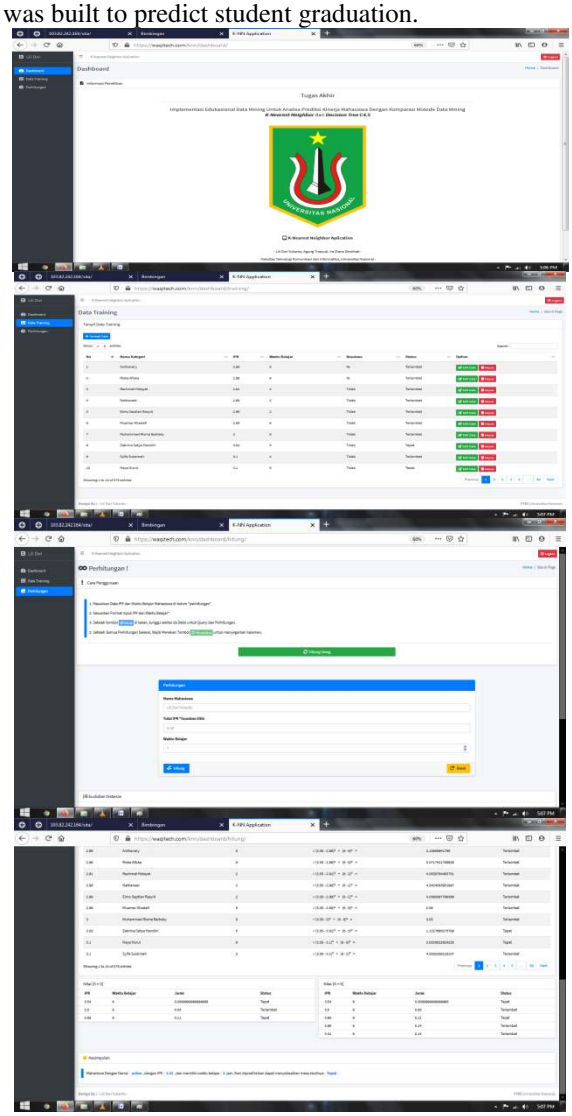
Metode	Akurasi %	Kappa
Decision Tree c4.5	54,80%	0,013
K-Nearest Neighbor	59,32%	0,202

In table 9 comparison by looking at the classification results between the 2 methods that have been calculated it can be said that the K-Nearest Neighbor method works better than Decision Tree c4.5 of the training data tested with an Accuracy level of 59.32%.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

D. Implementation

The application that has been designed is an application that uses the Hypertext Preprocessor (PHP) programming language that was built to predict student graduation.



4. Conclusion

Based on the results of the discussion, calculation, testing and comparison of the two Decision Tree c4.5 comparisons with the K-Nearest Neighbor method, the following conclusions can be drawn.

a. Based on testing on the two algorithm models using ratio training data can affect the results of each accuracy value, while the best accuracy value on the K-Nearest Neighbor algorithm model with an accuracy rate of 59.32% while the Decision Tree c4 algorithm model .5 has an accuracy rate of 54.80%
b. It can be concluded that the best accuracy value from the comparison of the two algorithm models is obtained by the K-Nearest Neighbor algorithm model with an accuracy rate of 59.32%. Great Hope This Educational Data Mining concept can be maximized and developed so that it can contribute and progress in the world of education
c. The application that will be developed is in the form of an Intelligent System to predict the graduation of student studies using the K-Nearest Neighbor algorithm model

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

5. Reference

[1] Rashi Bansal, Akansha Mishra, Dr. Shailendra Narayana Singh, 2017 "Mining Of Educational Data For Analysing Students' Overall Performance", International Conference on Cloud Computing, Data Science & Engineering, Januari 2017.

[2] Yuli Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5", Jurnal Edik Informatika, Volume 2, No. 1, Juni 2017.

[3] Khokhoni Innocentia Mpho Ramaphosa, Tranos Zuva, Raoul Kwuimi, " Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools", international Conference on Information and Communications Technology, Agustus, 2018.

[4] Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde, " Prediction of Students Performance using Educational Data Mining", International Conference on Data Mining and Advanced Computing, Maret 2016.

[5] Mokhairi Makhtar, Hasnah Nawang, Syadiah Nor Wan Shamsuddin, " Analysis On Students Performance Using Naïve Bayes Classifier", Journal of Theoretical and Applied Information Technology, Vol.95. No.16, Agustus 2017.

[6] Yoga Pristyanto, Irfan Pratama, Anggit Ferdita Nugraha, " Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification", International Conference on Information and Communications Technology, Maret 2018.

[7] Maulana Aditya Rahman, Nurul Hidayat, Ahmad Afif Supianto, " Komparasi Metode Data Mining K-Nearest Neighbor Dengan Naïve Bayes Untuk Klasifikasi Kualitas Air Bersih (Studi Kasus PDAM Tirta Kencana Kabupaten Jombang)", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 12, Desember 2018.

[8] Galih, " Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswas dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar", Jurnal Teknologi Sistem Informasi dan Aplikasi, Vol. 2, No. 1, Januari 2019

[9] Sari Dewi, "Komparasi 5 metode Algoritma Klasifikasi Data Mining pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan", Jurnal Techno Nusa Mandiri, Vol. XIII, No. 1 Maret 2016.

[10] Chitra Jalota, Rashmi Agrawal, " Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, Februari 2019.

[11] Langgeng Listiyoko, Rosalia Wati, Achmad Fahrudin, " Klasifikasi Siswa Untuk Meningkatkan Nilai Rata-Ratakelas Menggunakan Metode Data Mining", SeminarNasionalTeknologi dan Rekayasa, 2017.

[12] Senna Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan" Universitas Indraprasta PGRI, 2018.

[13] Daniel David, Sani M. Isa, " Penerapan Educational Data Mining Untuk Memprediksi Hasil Belajar Siswa SMAK Ora et Labora", Department of Computer ScienceBina Nusantara (BINUS) University, VOL. XII No, September 2019.

[14] Diky Firdaus, " Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer", Universitas Mercu Buana, Volume 6, No 2, Tahun 2017.

[15] Sheena Angra, Sachin Ahuja, "Implementation of Data Mining Algorithms on Student's Data using Rapid Miner," International Conference on Cloud Computing, Data Science & Engineering,Maret, 2017.

[16] Anoopkumar M, Dr. A. M. J. Md. Zubair Rahman, " A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration", International Conference on Cloud Computing, Data Science & Engineering, Maret, 2016.

[17] Edona Doko, Lejla Abazi Bexheti, "A Systematic Mapping Study of Educational Technologies based on Educational Data Mining and Learning Analytics, Mediterranean Conference On Embedded Computing, Juni, 2018.

[18] Nurul Hidayat, Retantyo Wardoyo, Azhari SN, " Educational Data Mining (EDM) as a Model for Students' Evaluation in Learning Environment", Third International Conference on Informatics and Computing (ICIC), Oktober,2018.

[19] Triayudi, Agung, and Iskandar Fitri. "A new agglomerative hierarchical clustering to model student activity in online learning." Telkomnika 17.3 (2019): 1226-1235.

[20] Prameswari, Eka Ayunda, Agung Triayudi, and Ira Diana Sholihati. "Web-based E-diagnostic for Digestive System Disorders in Humans using the Demster Shafer Method." International Journal of Computer Applications 975: 8887

[21] Triayudi, Agung, and Iskandar Fitri. "Comparison of parameter-free agglomerative hierarchical clustering methods." ICIC Express Letters 12.10 (2018): 973-980.