

Identification of Student Academic Performance using the KNN Algorithm

Aldi Nugroho¹, Osvaldo Richie Riady², Alexander Calvin³, Derwin Suhartono⁴

^{1,2,3,4} Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480

aldi.nugroho002@binus.ac.id; osvaldo.riady@binus.ac.id;
alexander.calvin@binus.ac.id; dsuhartono@binus.edu

Abstract – Students are an important asset in the world of education also an institution and therefore also need to pay attention to students' graduation rates on time. The ups and downs of the percentage of students' abilities in classroom learning is one important element for assessing university accreditation. Therefore, it is necessary to monitor and evaluate teaching and learning activities using the KNN Algorithm classification. By processing student complaints data and seeing the results of previous learning can obtain important things for higher education needs. In predicting graduation rates based on complaints, this study uses the K-Nearest Neighbor classification algorithm by grouping data $k = 1$, $k = 2$, $k = 3$ with the smallest value possible. In experiments using the KNN method the results were clearly visible and showed quite good accuracy. From the experiment it was concluded that if there were fewer complaints from one student it could minimize the level of student non-graduates at the university and ultimately produce good accreditation.

Keywords: KNN; Classification algorithm; Academic quality.

I. INTRODUCTION

An important aspect that determines the quality of a tertiary institution is how the quality of students is produced. The quality of students is determined by the magnitude of graduation rates and grades in each course as the basis for their calculations.

Classification is one of the topics in machine learning. In its development the classification is very much used in needs. the fact of research ever conducted by (Lau, E. T., 2019) in his research used machine learning techniques that can identify or predict student performance results in an exam. In the end the results of the study are very useful in academic interests and can help students in the learning

process.

However, in many universities there is still an imbalance in the level of graduation in certain majors, which can be seen precisely in the subjects. From there we saw a number of data lessons which always produce a large number of students who conduct short semesters in each semester, therefore we are interested in doing this research. With this research we will take data from students randomly which data is about student graduation.

In this study, the data taken is data of students who did not graduate and the reasons. From there the data will be processed with the aim of measuring the academic development of students at the university, so that it can balance the quality of the process of delivering material that is received equally by each student according to his abilities.

Research studied by (Al-Shehri, H., 2017) using the questioners method with the K-NN & Vector Machine algorithm. Of the 395 datasets it gives advantages if tested with many combinations of results not much different in comparison which has a weakness of the regression problem is very difficult to compare with the classification problem but can still get good results.

Research conducted by (Strecht, P., 2015). The students were analyzed using the SVM, CART, NB, RF, AB, Baseline, KNN algorithms. Of the 5779 datasets giving advantages to this test the performance between the stable algorithm runs which shows good results which are also in the journals made (Al-Shehri, H., 2017) explained the performance between the algorithm running stable indicates good results, but has the disadvantage that some algorithms are more sensitive to parameters than others. Therefore, it is not guaranteed that they run with the best configuration.

Research previously researched by (Nagesh, A. S., 2017) using the examination method with the KNN algorithm, Support Vector Machine, Bayesian Belief Network. Of the 25 datasets providing strengths, we can

improve the performance of weak students by guiding and conducting individual sessions, so that passing tests that have weaknesses when processing data can be long because of large datasets but explained by (Al-Shehri, H., 2017) using the same type of method namely KNN has the advantage that if tested with many combinations the results are not much different than those that have weaknesses of the regression problem is very difficult to compare with the classification problem but can still get good result.

Research conducted by (Hastarimasuci, R., 2019) using the recapitulation method with Data mining algorithm, KNN, Naïve Bayes Classifier, Confusion Matrix. 356 datasets provide advantages by comparing the calculations obtained from all methods can find good results, which has the disadvantage of not doing the comparison, the results obtained are less. Same as (Strecht, P., 2015) using various classifications as well as 5779 datasets giving an advantage for this test of performance between stable running algorithms that show good results.

Previous studies were investigated by (Alferi, S. S., and Maghari, A. Y., 2018) using the method of analyzing students with the KNN algorithm. Of the 14 datasets giving advantages to the efficiency of the algorithm results by increasing classification accuracy, which has the disadvantage of the results obtained at different distances at the end, but the results performed (Hastarimasuci, R., 2019) using 356 datasets provide advantages by comparing the calculations obtained from all methods to find good results, and therefore it may be more appropriate to use the classification method if the dataset used is large.

Research conducted by (Kumari, P., 2018) using the method of analyzing students with algorithms KNN, SVM, ID3. Of the 500 datasets giving results with the advantages of increasing yield using the ensemble method with four traditional classifiers (ID3, NB, KNN, SVM), with the lack of methods only done well in the case of NB and SVM. as discussed (Alferi, S. S., and Maghari, A. Y., 2018) the results of the use of the KNN method will be seen from here if the use of large datasets, the expected results will be more.

Subsequent research with researchers by (Amra, I. A. A., and Maghari, A. Y., 2017) using the method of analyzing students with the KNN algorithm, Naïve Bayes Algorithm. Out of the 500 datasets, the resulting classification model yields relatively accurate predictions, so that ministries can achieve progress based on attributes, with the lack of results from calculations likely to be several percent different as in the previous case, from the results given (Kumari, P., 2018) have a similar method but when compared with previous studies that have been discussed can be seen the resulting classification model produces relatively accurate predictions.

Previous studies examined by (Samuel, M. G., 2016) using the method of analyzing students with KNN, SVM, ID3 algorithms. From 300 datasets, the prediction

advantages based on SVM classification are more accurate compared to KNN-based predictions, with the lack of other methods to compare if KNN results are less accurate. But from several studies conducted one of them (Al-Shehri, H., 2017) using the same type of method namely KNN has a good profit calculation results.

Research with (Putpuek, N., 2018) using the method of analyzing students with the CRISP-DM, KNN, Naïve Bayes algorithm. Of the 2281 datasets with the advantages of Naïve Bayes providing the best predictive value among other methods, it has a disadvantage seen from the overall accuracy of the data that is still not good because of the data mining model used. but from several studies that exist and as an example of (Samuel, M. G., 2016) has the opposite results, because of the results of the research obtained the advantages obtained are good accuracy.

Research conducted by (Ab Razak, W. M. W., 2019) using the questionnaires method with the Correlation and regression analysis algorithm. Of the 500 datasets having the advantage of producing accurate data, the shortcomings of the results need to be processed first. Can be seen if coupled with the method we use in research that is the KNN method, it may be that the results obtained will have more advantages.

Previous research that was researched by (Maganga, J. H., 2016) using the interview method with the interview algorithm. From 33 datasets resulting in excess researchers can find out more about the reasons of the respondents, having shortcomings takes a long time. Nearly similar research was carried out (Ab Razak, W. M. W., 2019) but with the large number of different datasets taken that have significant differences in results that can be concluded with the number of datasets, the results obtained will be better.

Research conducted by (Kassarnig, V., 2018) using the Dedicated Smartphones method with the ANOVA F-test algorithm. From 538 datasets it produces excess processing data retrieval is easy, has drawbacks. By using the same method (Ab Razak, M. W., 2019) a similar result was obtained in the same study as well as having the advantage of accuracy that was obtained.

Previous research conducted by (Mushtaq, I., and Khan, S. N., 2012) using the questionnaire method with Descriptive algorithm, correlation, and regression analysis. Of the 155 datasets with the advantages of accurate results, having weaknesses requires a lot of factors. With the method used and also 155 datasets that were tested it gets good accuracy, but compared to the research we did with the dataset also also strengthened by previous studies getting satisfactory results when compared with regression analysts.

Research studied by (Gbollie, C., and Keamu, H. P., 2017) using the Cross-sectional quantitative research design method with the One-way repeated-measures ANOVA algorithm. Of the 323 datasets yielding an advantage of accurate results because it uses statistics, it has the disadvantage of needing respondents from different regions for more accurate results. With the same purpose but with different methods, it might be a good research to be done or

tested to distinguish also from the research carried out, but the difference with the KNN method used is more difficult to apply.

Research conducted by (Dev, M., 2016) using the Expost-facto method with the General mental ability test algorithm, Multiphasic Interest Inventory, Home Environment Inventory. Of the 110 datasets producing excess types of tests that can greatly improve the accuracy of the results, the lack of respondents is less flexible in answering. In its application, there are deficiencies in this method, compared to KNN. This method has the disadvantage of being less flexible when compared to KNN which has good flexibility.

Research with (Sumbawati, M. S., and Anistiyasari, Y., 2018) using the questionnaire method with 4-stages quantitative research algorithm. Of the 60 datasets tested yielded simple, easy to analyze advantages, it lacked detailed results. With quantitative use compared to the KNN method which is easy to apply, this method is also easy to benefit, but from the results found in this study, the results are less detailed.

Previous research was researched by (Singh, S. P., Malik, S., and Singh, P., 2016) using the Questionnaire method with regression analysis algorithm. Of the 200 datasets tested yielding the advantages of accurate results, having weaknesses for accurate results, requires many respondents. The same as (Mushtaq, I., and Khan, S. N., 2012) with the method used to obtain benefits but when used the KNN method or compared with it to get satisfactory results.

Research studied by (Cavilla, D., 2017) using the questionnaire method with Qualitative and quantitative Analysis algorithms. Of the 242 datasets it produces advantages that are easy to read, have weaknesses that cannot be implemented into large datasets. There are also many studies using this method similar to (Singh, S. P., Malik, S., and Singh, P., 2016) also a dataset whose comparison is not too far to produce similar results but the comparison with KNN is still better in its implementation.

Research conducted by (Mesarić, J., and Šebalj, D., 2016) using the Decision Tree method with J48, RandomForest, RandomTree, REPTree algorithms. Of the 665 datasets yielding advantages, it can determine which algorithm is suitable for determining the sample category lack of data processing will be long if the factors are many. From the method we see from this study, there are differences in the use of the method which has its own advantages but seen from the results of the research has the disadvantages of processing the data compared to research (Kumari, P., 2018) which makes a comparison of methods but has good results, it can be said other methods are best used if the comparison is done first.

Research with (Crisp, G., Taggart, A., and Nora, A., 2016) using the Literature review method. With this research produced advantages. To discuss many factors, this method is more effective because the consideration of seeing the results obtained can be combined with other journals and can be explained more broadly, having deficiencies The

factors taken cannot be too specific. With this research we can conclude if we use the method of literature review compared with research that uses the method of at least one method still has the advantage of research that has one method. Can be seen the results of this study have a side factor that is the lack of specifics.

The study was researched by (Khan, A., 2017) using the survey method with regression algorithm. From 418 datasets yielding the advantages of how one factor influences clear data changes and how large the changes are, it has a deficiency that is difficult to classify if there are many factors. Such as (Singh, S. P., Malik, S., and Singh, P., 2016) but what distinguishes the previous studies using several methods, but has similar results that is difficult to classify the data.

Research that has been studied by (Costa, A., dan Faria, L., 2018) using the meta analysis method with The Pearson's correlation coefficient (r) algorithm. From 412022 the dataset produced excess categories from the results that can be clearly distinguished from the dividing line, has the disadvantage of not being able to account for multidimensional relations. In this study, different methods were obtained from the several research methods that have been discussed and it is possible that this research method is good for use as a comparison, but judging from the results obtained, it has the disadvantage of being unable to explain multidimensional relationships. But that does not rule out the possibility that this method will be used as a comparison.

Research conducted by (Mueen, A., Zafar, B., and Manzoor, U., 2016) using data mining methods with the Naive Bayes algorithm, KNN, Decision tree. Of the 60 datasets yielding the advantages of naive bayes grouping is more accurate and specific precision, has the disadvantage of KNN not being applied to the order of the most influential attributes of the results. But if we compare it with research (Ab Razak, W. M. W., 2019) with more than 500 datasets tested, we have the advantage of producing higher data accuracy, perhaps this is constrained from the results of research that only uses a few datasets.

Research studied by (Seibert, G. S., 2017) using the Survey and report method with statistical analysis algorithm. From 550 datasets, there is an advantage of grouping the value of each factor with clear results, having the lack of a positive and negative factor standard. From the research conducted (Cavilla, D., 2017) compared to this study which is seen from the number of datasets tested and also the use of the same method, and has less good results when compared to the use of the KNN method.

Research that has been studied by (Thomas, C. L., 2017) using the add course credit method with Descriptive analysis and Hierarchical regression analysis algorithms. From 534 datasets, there is a descriptive advantage of producing clear qualitative data, hierarchical regression to get data that clearly influences based on the given stage, has a lack of unclear sequence of results from the two analyzes. With similar methods used (Singh, S. P., Malik, S., and Singh, P., 2016) that have a deficiency in accuracy when compared to the methods we use KNN it is better to use the

KNN method.

Research that has been examined by (Akessa, G. M., and Dhufera, A. G., 2015) using a survey method with information regression algorithm. Of the 294 datasets yielding excess factors and the sample taken is very complete, it has no drawbacks due to internal differences in majors too generalized. Previous research has also been carried out by some researchers such as (Thomas, C. L., 2017) that must be recognized is the lack of accuracy obtained when compared with other methods, this method is very less to be recommended.

Research conducted by Angelica Moè. In 2015 using the path model method with regression algorithm. Of the 218 datasets yielding the excess results obtained are described in the form of paths so that clearly the relation affects the results, has the disadvantage of no grouping which clearly only sees positive and negative results (Moè, A., 2015).

The related research has been conducted by Abdul Rohman in 2015 which is about a prediction of students' graduation using K-Nearest Neighbor Algorithm model. The issues that are assessed are evaluation of performance that will help students, lecturers, administrators and policymakers. The method used is K-Nearest Neighbor (K-NN). Results obtained from 1633 research samples using K-Nearest Neighbor (K-NN) obtained accuracy value is 82.25% and AUC value is 0,500, with the data cluster $k = 2$ accuracy is 79.45% and AUC value is 0,826, with the data cluster $k = 3$ accuracy is 83.95% and the AUC value is 0,853, with the data cluster $k = 4$ accuracy is 82.62% and AUC value is 0,874, with the data cluster $k = 5$ accuracy is 85.15% and AUC value is 0,888 (Rohman, A., 2015).

Furthermore, the research conducted by Mutiara, Budiman, and Farmadi in 2015 about K-Optimal implementation on K-Nearest Neighbor Algorithm with the intention of increasing accuracy percentage to predict Computer Science student graduation until forth semester. With the aim of knowing the K-Optimal value and its accuracy level on the K-NN algorithm to predict timely graduation of students based on IP. The method used is K-NN algorithm. Of the 110 sample data generates the K-Optimal value on the K-NN algorithm for timely graduation prediction of students based on IP up to semester 4 is $k = 5$. The K-Fold Cross Validation process obtained a level of accuracy for $k = 5$ in the K-NN algorithm for the timely prediction of students based on IP up to semester 4 is 80.00% (Banjarsari, M. A., 2015).

In addition, there was also a study conducted by Mustakim, Oktaviani in 2016 with the title Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. With the aim of calculating the K-NN algorithm implemented to an Early Warning System (EWS). The method used is K-NN and data mining. From 165 samples, calculate the K-NN algorithm applied in predicting the predicate. Students are able to obtain an accuracy of 82% (Mustakim, M., and Oktaviani, G., 2016).

Previous research conducted by Annisa in 2015 with different method in this case by using CRISP-DM to predict student graduation rate. With the aim of studying

the prediction of the final results of graduating students based on aspects of assessment in lectures. The method used is CRISP-DM and tried 118 samples on the decision tree pattern based on sample data is known to be the most influential assessment aspect first, in graduating student studies taking UTS courses, whereas in the rule pattern applied, the influential value is UAS this is known based on the calculation of the final value applied by the university, namely: $\text{END VALUE} = 50\% \text{ final test} + 30\% \text{ mid test} + 20\% \text{ ASSIGNMENT}$ (Fadillah, A. P., 2015).

Research conducted by Izzah in 2016 using hybrid fuzzy system to find out cluster of student that predicted not to graduate. Having the aim to find out the final graduation in the middle of the semester, the lecturer will be able to pay special attention to those clusters. With the Hybrid fuzzy inference system method and 106 sample data, 5 rules can be generated which are then run in the FIS. Classifier evaluations are calculated using measures of accuracy, sensitivity, and specificity. Of the three measures obtained an accuracy of 94.33%, a sensitivity of 96.55% and a specificity of 84.21% (Izzah, A., and Widyastuti, R., 2016).

Previous research conducted by Kusumadewi in 2004 to analyze connection of factors between lecturer and student. With the aim of determining how much the qualitative factor of student assessment of the performance of IT department lecturers, influences the relationship between lecturer attendance and student final grades, using fuzzy quantification theory I. The method used is Fuzzy quantification I, using 78 samples that produce more than 10 attendance times, then each fuzzy group will provide a positive correlation with the percentage of student passing scores \geq 'B' reaching more than 45% (Kusumadewi, S., 2004).

In a study conducted by Jessica in 2017 about effect of cluster grouping of classes. The problem being tested is to prove whether clustering based on student performance results in better performance. By using the Survey method, Summative assignment, and interview using 47 samples that produce clustering groups make students perform better than heterogeneous grouping (Christman, J. L., 2017).

In addition to the research carried out above, there is also the same research conducted by (Slavin, R. E., 1987) doing many kind of grouping. Research aimed at Shows class grouping based on skills, knowledge, development, learning rate of students using the 0-2 year learning method with a sample of 23,962 getting conclusions Cluster grouping makes students performed better rather heterogeneous grouping. With the grouping method has good results but the results of this method have not been compared with other methods as a comparison.

Furthermore, research conducted by (Rees, D. I., 2000) to identify the effect of ability grouping based on student performance. The method used is one year of learning based on assessment grouping. With a sample of 16,142 concluded a Formal or non-formal track doesn't appear to be impacting student abilities or achievements but the environment that has a similar composition that improves their abilities. The similarity of methods used by

(Slavin, R. E., 1987) with similar results is good. But this grouping method is still not good enough to be used or must be tested against other methods.

There was research by (Bunkar, K., 2012) about graduate students' performance prediction using data mining technique in specific decision tree. The aim to show that decision tree method has a better accuracy than any other methods. The result show that the precision is between 0.6.0-9. With FAIL Class in 0.845 and PASS Class in 0.684 shows that this method is perform a bit inconsistent rather than K-Nearest Neighbor method. Maybe the use of this method will be better if later compared to the method used by (Amra, I. A. A., and Maghari, A. Y, 2017) which has a comparison of accuracy with good results.

Other Research conducted by (Bekele, R., and Menzel, W., 2005) about Bayesian Networks approach to students' performance on their study. The result classified into three categories, below satisfactory, satisfactory, and above satisfactory with about 64% of records said classified correctly. This shows that K-NN method has better accuracy to classify data. As well as the method used (Amra, I. A. A., and Maghari, A. Y, 2017) and also other researchers from the use of the Bayesian method do have good results and also its accuracy but still has disadvantages compared to KNN.

There was research conducted by (Shahiri, A. M., and Husain, W., 2015) about review on data mining method to predict students' performance. This research uses five methods of data mining namely Decision Tree, Neural Network, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine; this is a further experiment to test what is the most accurate method of data mining. This review shows that the highest prediction accuracy is Neural Network by 98% followed by Decision Tree by 91%, next is K-Nearest Neighbor and Support Vector Machine method by 83%, and last is Naïve Bayes method by 76%. With that result obviously that the method supposed to be used is Neural Network with the highest percentage or Decision Tree which is higher than K-Nearest Neighbor but there are some factor involved lead to K-Nearest Neighbor more fitted to predict student performance. All related to psychometric factor that usually a qualitative data. By including psychometric factor, Neural Network accuracy decreased into only 69% because this method can't handle qualitative data instead of quantitative data. Then Decision tree method accuracy also decreased into 65% because psychometric value. But in the other hand, K-NN method accuracy didn't affected after psychometric value included.

There was research conducted by (Affendey, L. S., 2010) about ranking of factors influencing student performance by using Data Mining. The method used in data mining is Waikato Environment for Knowledge Analysis (WEKA). The result shows that from all of the method, AODE get the highest accuracy with 95.29% than the other method. With research conducted (Hastarimasuci, R., 2019) with the same method but by comparing it with several other methods, it was found that the lack of this method is the results obtained are less.

There was research conducted by (Carter, A. S., 2015) about prediction of computer science student performance based on their programming behavior. The method used is a programming state model by focus on problem solving ability fixing syntax error. Result of the research shows that there are large difference when the real result is about 18% - 36% error rate while the study only 3% - 12%. With this method problem solving will be good to be used also for additional comparisons from KNN and also others to find out the accuracy results that will be obtained.

II. METHODS

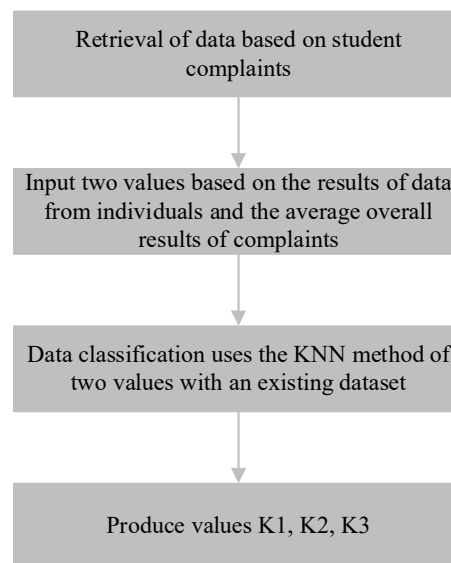


Figure 1 Student data processing flow

The method we use is the same KNN algorithm as (Rohman, A., 2015). From previous studies (Banjarsari, M. A., 2016), (Mustakim, M., and Oktaviani, G., 2016), it can be seen from figure 1. We practice the data by making one case. Divide data into two as input that will be compared with existing complaint data from 6. By comparing inputs with complaints, from there will be used as coordinates to provide location and easy to see images.

All experiments have been carried out with the KNN algorithm by getting the closest data value in the absence of complaints or complaints as much as 0 with the aim of minimizing students who get a short semester. Using experimental methods to see the best results, we also do it with several experiments with different data values.

III. RESULTS AND DISCUSSION

In collecting our dataset we collected data from students who took the short semester and did not take the short semester. To complete the data classification process that we will use by using the Google form we get approximately 35 data for us to measure. From this data we take the names of any courses that have a short semester as well as their complaints about why they can take a short semester. The valuable data will then be labeled.

We will make the complaint data as the main reference or benchmark.

The data will be tested using the proposed method (Banjarsari, M. A., 2016). The results obtained can be used to create inputs that will be classified. In this paper we use euclidean distance. The results will be chosen based on the closest data, as we want with the smaller the value, the better the results obtained.

The classification of data that we use will compare the results of the data we enter and produce a comparison value in accordance with the complaint data that we have previously determined.

Table 1 Student complaint data results

Complaint	1 Person	Average
Lecturer teaching is not good	0	18
Lack of learning material	0	12
Lack of exercise	0	21
Absent	0	7
Exam question are different from the material	0	8
Etc.	0	1
Input	0	3

In testing data from table 1. we will test data with the model (x, y) and calculate it using Euclidean distance. We fill in the data by giving number 1 for the answer yes, and giving number 0 for the answer no, for each person. In table 1. we see the number of complaints in 1 person, as many as 0 data will be made (x), and take the 3 largest average value of all complaints made (y) then we will get an input value (0, 3).

Table 2 Student complaint test data

Result	Amount yes	Amount not
Short semester	6	6
Short semester	5	5
Short semester	4	4
Short semester	3	3
Not short semester	2	2
Not short semester	1	1
Not short semester	0	0

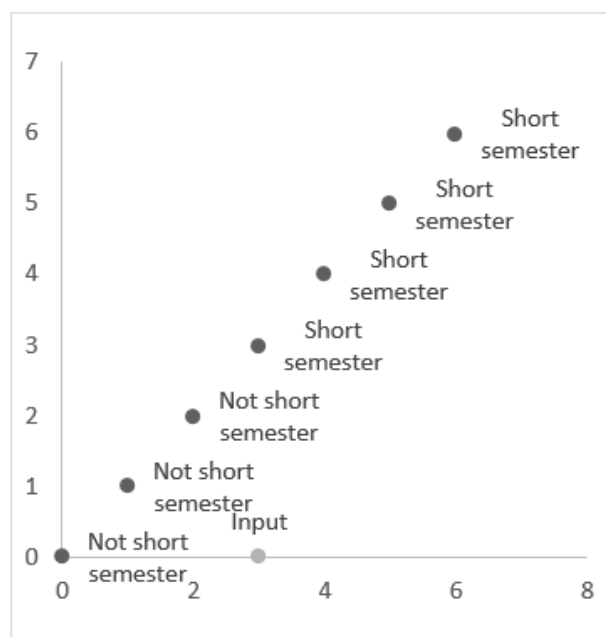


Figure 2 Visualization of test results

From the input values obtained from table 1. obtained (0,3), it will be classified (x2, y2) according to table 2. which is a 1:1 ratio between the number of yes as (x2) and the amount not like (y2) which obtained from 6 complaints and inversely proportional. in Figure 2. can be seen visualization of data that has been compared.

Table 3 Results of data calculation

Rank	Euclidean Distance	Label
7	6,70820	Short semester
6	5,38516	Short semester
5	4,12311	Short semester
3	3,00000	Short semester
1	2,23607	Not short semester
1	2,23607	Not short semester
3	3,00000	Not short semester

Table 4 K1, K2, and K3 data classification results

K	Label
1	Not short semester
2	#N/A
3	Short semester

From the classification of table 1., it can be seen from table 3. which uses euclidean distance calculation using the formula to produce the K value that has been ranked in the closest value. The results of # N / A in table 4. will be the second closest ranking value or the results of # N / A in table 4. will be obtained if the results of the euclidean distance are the same and will be averaged.

IV. CONCLUSION

We show the closest results from each of the data compared to the 6 data complaints. The results of the data can be seen in K and Label showing the value of data based on the results of rank. We see from the results that come out of the input (0,3) of the first experiment, the value obtained is not short semester.

Based on experimental data:

- That if the value of 1 person and the average of table 1. is getting bigger, it will increase the students to get a short semester, conversely if the value of 1 person and average of table 1. get smaller then it will decrease or the possibility of no students getting a short semester.
- By minimizing complaints in lectures can minimize the level of students getting a short semester.

REFERENCES

- Ab Razak, W. M. W., Baharom, S. A. S., Abdullah, Z., Hamdan, H., Abd Aziz, N. U., & Anuar, A. I. M. (2019). Academic Performance of University Students: A Case in a Higher Learning Institution. *KnE Social Sciences*, 1294-1304.
- Affendey, L. S., Paris, I. H. M., Mustapha, N., Sulaiman, M. N., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, 9(4), 832-837.
- Akessa, G. M., & Dhufera, A. G. (2015). Factors That Influence Students Academic Performance: A Case of Rift Valley University, Jimma, Ethiopia. *Journal of Education and Practice*, 6(22), 55-63.
- Alfere, S. S., & Maghari, A. Y. (2018). Prediction of Student's Performance Using Modified KNN Classifiers. *Prediction of Student's Performance Using Modified KNN Classifiers*.
- Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017, April). Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)* (pp. 1-4). IEEE.
- Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.
- Banjarsari, M. A., Budiman, I., & Farmadi, A. (2016). Penerapan K-Optimal Pada Algoritma Knn Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan Ip Sampai Dengan Semester 4. *Klik-Kumpulan Jurnal Ilmu Komputer*, 2(2), 159-173.
- Bekele, R., & Menzel, W. (2005). A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. *algorithms*, 22(23), 24.
- Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012, September). Data mining: Prediction for performance improvement of graduate students using classification. In *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)* (pp. 1-5). IEEE.
- Carter, A. S., Hundhausen, C. D., & Adesope, O. (2015, August). The normalized programming state model: Predicting student performance in computing courses based on programming behavior. In *Proceedings of the eleventh annual International Conference on International Computing Education Research* (pp. 141-150).
- Cavilla, D. (2017). The effects of student reflection on academic performance and motivation. *SAGE Open*, 7(3), 2158244017733790.
- Christman, J. L. (2017). Performance-based cluster grouping in ninth grade honors physics.
- Costa, A., & Faria, L. (2018). Implicit theories of intelligence and academic achievement: A meta-analytic review. *Frontiers in psychology*, 9, 829.
- Crisp, G., Taggart, A., & Nora, A. (2015). Undergraduate Latina/o students: A systematic review of research identifying factors contributing to academic success outcomes. *Review of Educational Research*, 85(2), 249-274.
- Dev, M. (2016). Factors Affecting the Academic Achievement: A Study of Elementary School Students of NCR Delhi, India. *Journal of Education and Practice*, 7(4), 70-74.
- Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). *Jurnal Teknik Informatika dan Sistem Informasi*, 1(3).
- Gbollie, C., & Keamu, H. P. (2017). Student academic performance: The role of motivation, strategies, and perceived factors hindering Liberian junior and senior high school students learning. *Education Research International*, 2017.
- Hastarimasuci, R., Papilo, P., & Nazir, A. (2019, November). Variable Selection to Determine Majors of Student using K-Nearest Neighbor and Naïve Bayes Classifier Algorithm. In *Journal of Physics: Conference Series* (Vol. 1363, No. 1, p. 012057). IOP Publishing.
- Izzah, A., & Widyastuti, R. (2016). Prediksi Kelulusan Mata Kuliah Menggunakan Hybrid Fuzzy Inference System. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 2(2), 60-67.
- Kassarnig, V., Mones, E., Bjerre-Nielsen, A., Sapiezynski, P., Lassen, D. D., & Lehmann, S. (2018). Aca-

- ademic performance and behavioral patterns. *EPJ Data Science*, 7(1), 10.
- Khan, A., Khan, S., Zia-Ul-Islam, S., & Khan, M. (2017). Communication Skills of a Teacher and Its Role in the Development of the Students' Academic Success. *Journal of Education and Practice*, 8(1), 18-21.
- Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.
- Kusumadewi, S. (2004). Fuzzy quantification theory I untuk analisis hubungan antara penilaian kinerja dosen oleh mahasiswa, kehadiran dosen, dan nilai kelulusan mahasiswa. *Media Informatika*, 2(1).
- Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, 1(9), 982.
- Maganga, J. H. (2016). Factors Affecting Students' Academic Performance: A Case of Public Secondary Schools in Ilala District, Dar es Salaam (Doctoral dissertation, The Open University of Tanzania).
- Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croatian Operational Research Review*, 7(2), 367-388.
- Moè, A. (2015). Perceived control mediates the relations between depressive symptoms and academic achievement in adolescence. *The Spanish journal of psychology*, 18.
- Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11), 36.
- Mushtaq, I., & Khan, S. N. (2012). Factors Affecting Students' Academic Performance. *Global journal of management and business research*, 12(9).
- Mustakim, M., & Oktaviani, G. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. *Jurnal Sains dan Teknologi Industri*, 13(2), 195-202.
- Nagesh, A. S., Satyamurty, C. V., & Akhila, K. (2017). Predicting Student Performance using KNN Classification in Bigdata Environment. *CVR Journal of Science and Technology*, 13, 83-87.
- Putpuek, N., Rojanaprasert, N., Atcharyachanvanich, K., & Thamrongthanyawong, T. (2018, June). Comparative study of prediction models for final GPA score: a case study of Rajabhat Rajanagarindra University. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)* (pp. 9297). IEEE.
- Rees, D. I., Brewer, D. J., & Argys, L. M. (2000). How should we measure the effect of ability grouping on student performance?. *Economics of Education Review*, 19(1), 17-20.
- Rohman, A. (2015). Model Algoritma K-Nearest Neighbor (K-Nn) Untuk Prediksi Kelulusan Mahasiswa. *Neo Teknika*, 1(1).
- Samuel, M. G. COMPARATIVE ANALYSIS OF KNN AND SVM CLASSIFIERS FOR STUDENTS' ACADEMIC PERFORMANCE PREDICTION. for National Safety & Security, 149.
- Seibert, G. S., Bauer, K. N., May, R. W., & Fincham, F. D. (2017). Emotion regulation and academic underperformance: the role of school burnout. *Learning and Individual Differences*, 60, 1-9.
- Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.
- Singh, S. P., Malik, S., & Singh, P. (2016). Research paper factors affecting academic performance of students. *Indian Journal of Research*, 5(4), 176-178.
- Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of educational research*, 57(3), 293-336.
- Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.
- Sumbawati, M. S., & Anistiyasari, Y. (2018, January). The impact of research-based learning on student's academic performance and motivation. In *IOP Conference Series: Materials Science and Engineering* (Vol. 296, No. 1, p. 012043). IOP Publishing.
- Thomas, C. L., Cassady, J. C., & Heller, M. L. (2017). The influence of emotional intelligence, cognitive test anxiety, and coping strategies on undergraduate academic performance. *Learning and Individual Differences*, 55, 40-48.