IMPLEMENTATION OF DECISION TREE AND K-NN CLASSIFICATION OF INTEREST IN CONTINUING STUDENT SCHOOL

Daniati Uki Eka Saputri[1]; Fitra Septia Nugraha[2] ;Taopik Hidayat[3]; Abdul Latif[4]; Ade Suryadi[5]; Achmad Baroqah Pohan[6]

[1,2,3] Computer Science
[1,2,3] STMIK Nusa Mandiri Jakarta, Indonesia
[1,2,3]www.nusamandiri.ac.id
[1]daniati.due@nusamandiri.ac.id, [2]fitra.fig@nusamandiri.ac.id, [3]taopik.toi@nusamandiri.ac.id

[4]Program Studi Teknik Informatika, [5,6]Program Studi Sistem Informasi
[4,5,6] Universitas Bina Sarana Informatika, Jakarta, Indonesia
[4,5,6]www.bsi.ac.id
[4]abdul.bll@bsi.ac.id, [5]ade.axd@bsi.ac.id, [6]achmad.abq@bsi.ac.id

Abstract— Education is important to prepare quality Human Resources (HR) because quality human resources is an important factor for the nation and state development. Therefore, it is expected that every citizen has the right to get high educational opportunities from the 12-year compulsory education level. This study aims to implement the Decision Tree and K-NN algorithm in the classification of student interest in continuing school. This study proposes combining the Decision Tree and K-NN algorithm methods to improve accuracy with the Gain Ratio, Information Gain and Gini Index approaches for the measurement process. The test results show that the use of the Decision Tree algorithm produces an accuracy value of 97.30% while using the K-NN algorithm produces an accuracy of 89.60%. While the proposed method by combining the Decision Tree and K-NN algorithms produces an accuracy value of 98.07%. The results of evaluation measurements using the Area Under Curve (AUC) on the Decision Tree algorithm are 0.992 and the AUC on K-NN is 0.958 and on the combination of the Decision Tree and K-NN algorithms of 0.979. These results indicate that the proposed algorithm is very significant towards increasing accuracy in the classification of the interests of high school students continuing school

Keywords: Accuracy, Classification, Interest, Decision Tree, K-NN

Abstrak— *Pendidikan merupakan hal yang penting untuk menyiapkan Sumber Daya Manusia (SDM) yang berkualitas. Tingkat pendidikan semakin tinggi maka semakin besar pula kesempatan untuk memperoleh pekerjaan yang lebih baik. namun hal itu tidak sesuai yang diharapkan, Minat setiap siswa untuk melanjutkan sekolah kejenjang yang lebih tinggi masih rendah. Data memperlihatkan Angka Partisipasi Sekolah (APS) berdasarkan karakteristik Demografi dan kelompok umur, bahwa di umur 7-12 tahun dengan jenjang Pendidikan dasar mencapai 99,14%, namun di usia 19-24 tahun menunjukkan penurunan hingga 30% yang minat untuk melanjutkan sekolah. Penelitian ini bertujuan untuk mengklasifikasikan minat siswa lanjut sekolah ke jenjang lebih tinggi. Penelitian ini mengusulkan metode klasifikasi menggunakan algoritma Decision Tree dan K-NN dengan penambahan fitur Multiply untuk meningkatkan akurasi dengan pendekatan Gain Ratio, Iinformation Gain dan Gini Index untuk proses pengukurannya. Hasil pengujian menunjukkan bahwa penggunaan algoritma Decision Tree dengan pendekatan Information Gain menghasilkan nilai akurasi sebesar 97,30%, sedangkan dengan menggunakan algoritma K-NN menghasilkan akurasi sebesar 89,60%. Sementara metode yang diusulkan dengan menggabungkan algoritma Decision Tree dan K-NN menghasilkan nilai akurasi sebesar 98,07%. Hasil pengukuran evaluasi menggunakan Area Under Curve (AUC) pada algoritma Decision Tree sebesar 0,992 dan AUC pada K-NN sebesar 0,958, serta pada gabungan algoritma Decision Tree dan K-NN sebesar 0,979. Hasil ini menunjukkan bahwa algoritma yang diusulkan sangat signifikan terhadap peningkatan akurasi dalam klasifikasi minat siswa SMA lanjut sekolah*

Kata Kunci*: Akurasi, Klasifikasi, Minat, Decision Tree, K-NN.*

INTRODUCTION

Education is now considered as important for preparing quality human resources because quality human resources are an important factor for the nation and state development. Bearing this in mind, development in the field of Education is

explained in the Sustainable Development Goals (SDGs) especially in the 4th Goal which is ensuring the quality of inclusive and equitable education, and promoting lifelong learning opportunities for all (BPS:2017). In addition, it is explained in Government Regulation No. 47 of 2008 concerning compulsory education, Therefore, it is expected that every citizen is entitled to a higher educational opportunity from the 12-year compulsory education level (Peraturan pemerintah:2008).

Human Capital Theory considers Education is an investment for everyone and is closely related to the opportunity to get better jobs for those who have a higher level of Education (Prasojo et al: 2018) However, this is sometimes not as expected, everyone's interest to continue their education to a higher level is different. Everyone's interest in continuing their education to a higher level is different because there are those who have high interest, moderate interest and even no interest at all to continue schooling (Arifin & Ratnasari:2017). The difference is inseparable from the factors that affect student interest, including the desires, ideals and motivation, other factors namely the family environment, school and community environment (Aji & Suyitno: 2016). Students who have high enthusiasm and motivation to learn to have better hopes (Muhammad, 2017) of success and always have the desire to develop their potential by continuing education to a higher level. (Aji & Suyitno: 2016).

Data from the Central Statistics Agency shows the School Participation Rate (APS) based on the characteristics of Demographics and age groups, that at the age of 7-12 years with a basic education level reached 99.14%, but at the age of 19-24 years both in urban and rural areas showed a decrease up to 30% are interested in continuing school (BPS:2017).

Data mining or also called pattern recognition is an algorithm for managing data to find patterns that are still hidden so as to produce new knowledge of the data being processed (Sadewo et al:2018). At present many researchers have developed statistical methodologies for classification or searching for patterns based on mathematical techniques and statistics in processing and exploring a number of existing datasets.

Based on the above problems, the classification of advanced students' interests is carried out to a higher level with the data mining algorithm. The classification of the data obtained using the classification algorithm will produce an accuracy value to analyze the data of the interest of advanced school students towards the attributes that influence it. The classification algorithm used for the data of student interest in advanced school is the Decision Tree algorithm and K-NN.

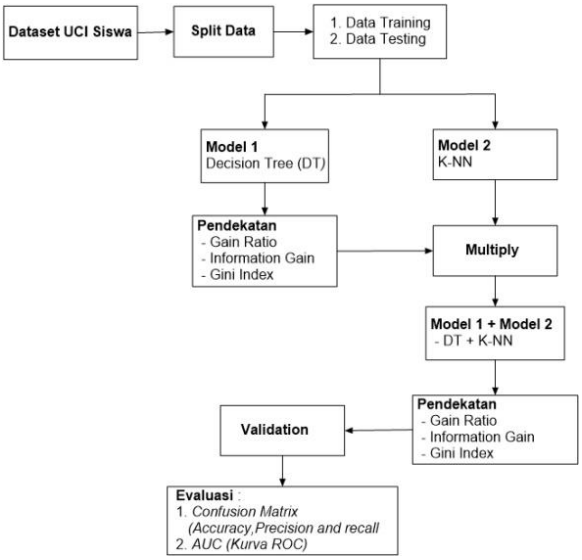
Research on the classification of data has been done by several researchers before. Some related studies include research by (Nugroho & Wibowo:2017) in 2017, this study discusses the determination of attributes that influence the graduation classification of students of the Faculty of Computer Science UNAKI Semarang using the classification method with the Naïve Bayes algorithm. The data used are data from an academic information system at the University of AKI in which there is some information such as NIM, name, majors, age, gender, origin, marital status, employment, scholarships, and others. This research obtained the accuracy of Naïve Bayes of 90.95% after combined with the forward selection feature obtained an accuracy of 97.14% with an AUC value of 0.981%.

In research (Dervisevic et al:2019) entitled Application of K-NN and Decision Tree Classification Algorithms in the Prediction of Education Success From the Edu720 Platform, using a comparison of the K-NN and Decision Tree algorithms to predict the educational success of the Edu720 platform yields an accuracy of each that is K-NN of 63, 92% and Decision Tree at 60.38%. After an approach with the splitting approach then random selection reduced, resulting in better accuracy which is increased to 76.39% for K-NN and Decision Tree algorithm by 72.27%. Based on previous research with data classification methods using the Naïve Bayes algorithm, Decision Tree and K-NN to find out the accuracy of the data classification results, which turned out to produce improved accuracy better after several additional approaches to the algorithm being tested. However, from some of these studies, there are still some shortcomings that need testing and comparison of algorithms and other features to gain wider knowledge. Therefore, this study aims to classify the interests of advanced school students to a higher level. This study proposes a classification method using the Decision Tree and K-NN algorithms with the addition of the Multiply feature to improve accuracy with the Gain Ratio, Gain Information, and Gini Index approach for the measurement process. So that it is expected to get better accuracy values

MATERIALS AND METHODS

The research methodology was conducted on data from UCI dataset students in 2 schools with 33 variables using the method of combining the Decision Tree classification algorithm and also K-NN. In the Decision Tree, algorithm process measurements are taken using the Gain Ratio, Gain

Information and Gini Index approaches. Then in the next process, a merging technique is carried out with the two algorithms by adding the multiply connection feature as a branching for integration into the student dataset. Following this is the process of the proposed research methodology.



Source : (Saputri et al., 2020)

Figure 1 Research Methodology

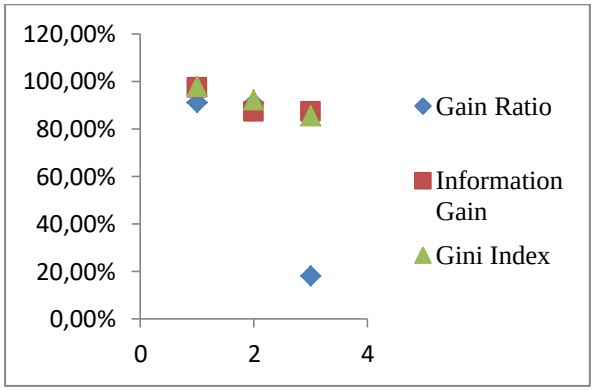
Based on Figure 1, there are two methodological processes, namely the process of calculating each classification algorithm and then the process of calculating the merging algorithm by adding multiply connection features as branching for integration into student datasets. Both processes are carried out to determine the effect of the performance of the algorithm to identify students' interest in continuing education to a higher level. The research process divides the data into training data by 80% and testing data by 20%. The algorithm performance results are measured based on accuracy, precision, recall and AUC values. AUC value is used to see the size formed and provide a single numerical metric so that it can be seen the performance comparison of the algorithm being processed. AUC values range from 0 to 1, which when approaching a value of 1 indicates better classification results. The following are the results of the decision tree for each classification algorithm. Following are the decision tree results from the Decision Tree and K-NN algorithms shown in table 1, and table 2.

Tabel 1. Hasil Pohon Keputusan *Decision Tree*

Decision Tree	*Accuracy*	*Precision*	*Recall*
Gain Ratio	91,14%	90,91%	18,18%
Information Gain	97,30%	87,27%	87,27%
Gini Index	97,69%	92,16%	85,45%

Source : (Saputri et al., 2020)

The Decision Tree as shown in table 1 shows that the study was very good. The value of the Gain Ratio approach is 91.14% in Accuracy, 90.91% in Precision and 18.18% in Recall. For the Information Gain approach 97.30% in Accuracy, 87.27% in Precision and 87.27% in Recall. Then approach with Gini Index 97.69% in Accuracy, 92.16% in Precision and 85.45% in Recall.



Source : (Saputri et al., 2020)

Figure 2 Scatter Diagram of Decision Tree Results Tree

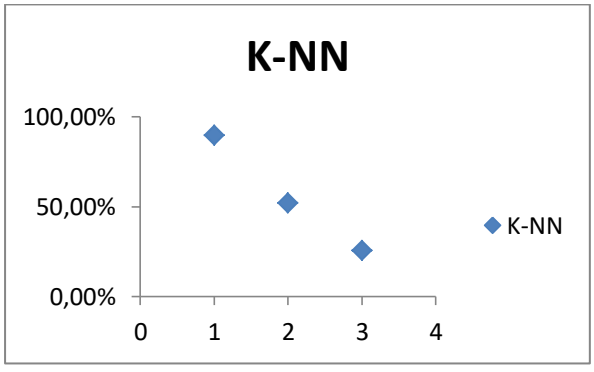
Figure 2 is a scatter diagram that shows that the results of the approach with Information Gain are the highest accuracy value compared to the other two approaches. Then the highest precision value is indicated by the Gini Index. The highest recall value is indicated by the Information Gain approach.

Table 2. Results of the K-NN Decision Tree

Algoritma	*Accuracy*	*Precision*	*Recall*
K-NN	89,60%	51,85%	25,45%

Source : (Saputri et al., 2020)

In table 2, the above shows the results of the KNN algorithm calculation. The resulting value is 89.60% in Accuracy, 51.85% in Precision, and 25.45% in Recall.



Source : (Saputri et al., 2020)

Figure 3 Scatter Diagram of K-NN Decision Tree Results

In Figure 3, shows a scatter diagram which is the display of Accuracy, Precision and Recall values from the K-NN algorithm calculate. The resulting accuracy value of 89.60% is a good study.

RESULTS AND DISCUSSION

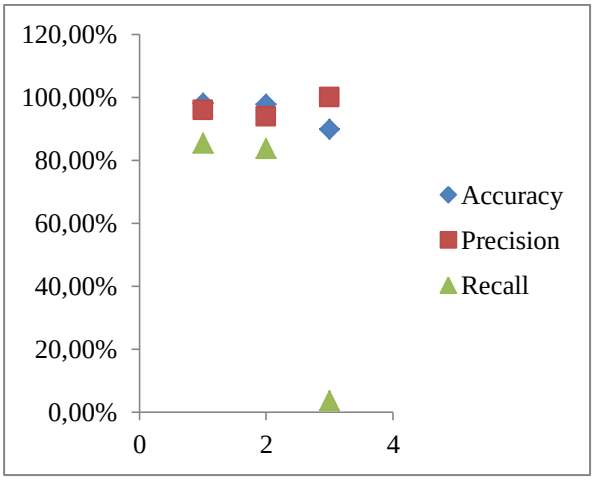
The test results are carried out to determine the accuracy of the Decision Tree algorithm by combining another algorithm, the K-NN algorithm for student datasets as many as 649. The experimental results show that the merging of the Decision Tree algorithm with K-NN shows increased accuracy. The following table 3 results of the comparison of algorithms with merging.

Table 3 Results of Comparison of Decision Tree Algorithms with K-NN

Decision Tree	*Accuracy*	*Precision*	*Recall*
DT(Information Gain)+K-NN	98,07%	95,92%	85,45%
DT(Gini Index)+K-NN	97,69%	93,88%	83,64%
DT(Gain Ratio)+K-NN	89,79%	100,00%	3,64%

Source : (Saputri et al., 2020)

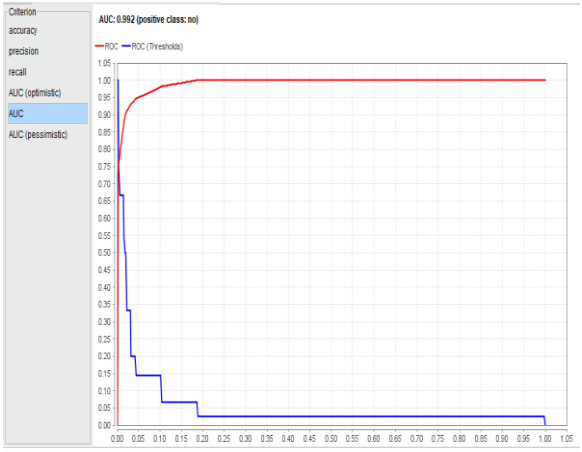
In table 3, above, shows the results of the comparison of the Information Gain, Gini Index and Gain Ratio approaches. Decision Tree value with the Information Gain approach yields 98.07% in Accuracy, 95.92% in Precision and 85.45% in Recall. Decision Tree value using the Gini Index approach produces 97.69% in Accuracy, 93.88% in Precision and 83.64% in Recall. Decision Tree value with the Gain Ratio approach produces 89.79% in Accuracy, 100.00% in Precision and 3.64% in Recall.



Source : (Saputri et al., 2020)

Figure 4 Graph of Test Results Comparison of Decision Tree and K-NN Algorithms

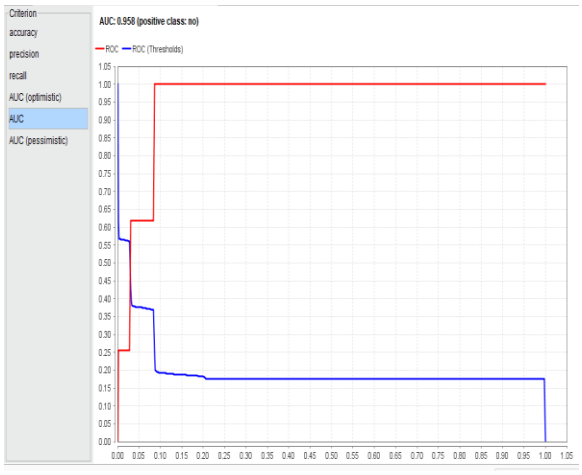
In Figure 4, a scatter diagram shows that the Decision Tree with the Information Gain approach combined with the K-NN algorithm shows the best level of accuracy reaches a value of 98.07% compared with 2 other approaches. The research is very good research. Evaluation with the ROC Curve shows the ROC graph with AUC Value for Decision Tree with Information Gain approach of 0.992, for AUC with K-NN algorithm produces a value of 0.958. After merging, the value obtained is DT (Information Gain) + K-NN of 0.979.



Source : (Saputri et al., 2020)

Figure 5. ROC Decision Tree Curve with Information Gain

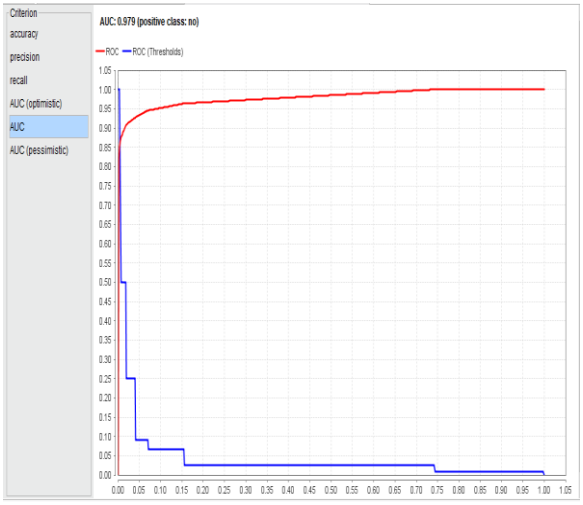
Figure 5, shows the results of the AUC Curve Decision Tree method with the Information Gain approach that is equal to 0.992 which means the classification test is very good.



Source : (Saputri et al., 2020)

Figure 6. K-NN ROC curve

In Figure 6, shows the results of the AUC Curve K-NN method that is equal to 0.958 which means the classification test is very good including.



Source : (Saputri et al., 2020)

Figure 7. DT (Information Gain) + K-NN ROC curve

In Figure 7. Shows the AUC Curve Decision Tree method with the Information Gain approach combined with the K-NN algorithm that is equal to 0.979 has decreased from before, but the classification test is still included as a very good classification.

CONCLUSION

From the research conducted on the student data above, it can be concluded that research using the same dataset for processing the Decision Tree algorithm by combining the K-NN algorithm can increase the value of accuracy better in classifying students' interest in continuing school. Decision Tree algorithm test results produce an accuracy value of 97.30%, for testing the K-NN algorithm of 89.60% then using the proposed method with the technique of merging the Decision Tree algorithm by adding multiply connection features increased to 98.07%. Meanwhile, the results of evaluations conducted with the ROC Curve on the Decision Tree algorithm test resulted in an AUC value of 0.992 and the K-NN algorithm resulted in an AUC value of 0.958. then using the proposed method by combining the Decision Tree algorithm and the K-NN algorithm to 0.979. The results of these studies produce higher accuracy than previous studies.

REFERENCE

Arifin, A. A., & Ratnasari, S. (2017). Hubungan Minat Melanjutkan Pendidikan ke Perguruan Tinggi dengan Motivasi Belajar Siswa. *JURKAM: Jurnal Konseling Andi Matappa*, *1*(1), 77–82. https://journal.stkip-andi-matappa.ac.id/index.php/jurkam/article/view/9

Badan Pusat Statistik. (2017). *Potret Pendidikan Indonesia Statistik Pendidikan*. Badan Pusat Statistik.

Dervisevic, O., Zunic, E., Donko, D., & Buza, E. (2019). Application of KNN and Decision Tree Classification Algorithms in the Prediction of Education Success from the Edu720 Platform. *2019 4th International Conference on Smart and Sustainable Technologies, SpliTech 2019*. https://doi.org/10.23919/SpliTech.2019.8783102

Peraturan Pemerintah Republik Indonesia Nomor 47 Tahun 2008 Tentang Wajib Belajar, Pub. L. No. 47, 10 (2008). https://peraturan.bpk.go.id/Home/Details/4861/pp-no-47-tahun-2008

Muhammad, M. (2017). PENGARUH MOTIVASI DALAM PEMBELAJARAN. *Lantanida Journal*, *4*(2), 87–97. https://jurnal.ar-raniry.ac.id/index.php/lantanida/article/view/1881

Nugroho, M. F., & Wibowo, S. (2017). Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes. *Jurnal Informatika Upgris*, *3*(1), 63–70. https://doi.org/10.26877/jiu.v3i1.1669

Prasojo, L. D., Mukminin, A., & Mobmudoh, F. N. (2018). *Manajemen Human Capital dalam Pendidikan*.

Sadewo, M. G., Windarto, A. P., & Wanto, A. (2018). Penerapan Algoritma Clustering Dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/ Mitigasi Bencana Alam Menurut Provinsi Dengan K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *2*(No.1 Oktober 2018), 311–319. https://doi.org/10.30865/komik.v2i1.943

Saputri, D. U. E., Nugraha, F. septia, Hidayat, T., Latif, A., Suryadi, A., & Pohan, A. B. (2020). *Final Report of Independent Research: Implementation of Decision Tree and K-NN in the Classification of Student Interest Level Continuing School*.

Suyitno, F. A. (2016). Faktor-Faktor Yang Mempengaruhi Minat Siswa Untuk Melanjutkan Studi SMK Jurusan TKR di SMP N 34 Purworejo. *Jurnal Autotech*, *8*(2), 113–

118.
http://ejournal.umpwr.ac.id/index.php/autotext/article/view/3112