Prediksi Kelulusan Mahasiswa Dalam Memilih Program Magister Menggunakan Algoritma *K-NN*

Aldo Sabathos Mananta[1], Green Arther Sandag[2]
e-mail: [1]s21610010@student.unklab.ac.id, [2]greensandag@unklab.ac.id
[1]Program Studi Sistem Informasi, [2]Program Studi Informatika, Universitas Klabat, Airmadidi

Abstrak

Kinerja akademik siswa didasarkan pada berbagai faktor seperti variabel pribadi, sosial, psikologis, dan lingkungan lainnya. Banyaknya jumlah data yang dimiliki oleh pihak universitas mengenai mahasiswa lulusan mereka dapat membantu dalam proses pengambilan keputusan ke jenjang pendidikan yang lebih tinggi. Data akademik mahasiswa dapat diolah untuk membantu pengambilan keputusan dalam penentuan masuk ke jenjang perguruan tinggi selanjutnya. Untuk melakukan proses pengolahan data tersebut di butuhkan metode data mining yaitu klasifikasi. *Aspek yang dilihat yaitu dari sisi accuracy, precision dan recall. Software yang digunakan untuk mengetahui kinerja dari algoritma tersebut adalah Rapid Miner Studio versi 9.2. Hasil pengujian menunjukkan bahwa algoritma K-NN dengan menggunakan independent test memiliki accuracy terbaik yaitu sebesar 96.25%, precision 98.08%, dan recall 70.00%. Sedangkan pada test yang menggunakan cross-validation, algoritma K-NN juga memiliki accuracy terbaik yaitu sebesar 91.88%, precision 81.29, dan recall 61.15%.*

Kata kunci *: Prediksi, Kelulusan Mahasiswa, K-NN*

1. Pendahuluan

Seiring dengan perkembangan zaman, perkembangan teknologi informasi telah berkembang dengan sangat pesat. Banyak data berupa informasi dari berbagai bidang yang dapat dihasilkan dari teknologi informasi yang canggih, baik itu dari bidang ekonomi, industri, ilmu dan teknologi, serta dari berbagai bidang lainnya. Dalam dunia pendidikan teknologi informasi dapat menghasilkan data yang berlimpah berupa data dari masing-masing siswa maupun individu lainnya. Dalam dunia pendidikan terutama pada perguruan tinggi, penerapan teknologi informasi dalam membantu institusi pendidikan untuk melakukan pengolahan data sangatlah penting dikarenakan data mahasiswa tiap tahun semakin bertambah, dapat menghasilkan informasi yang berlimpah setiap tahunnya berupa jumlah kelulusan, profil, dan hasil akademik mahasiswa selama menempuh kegiatan akademik pada perguruan tinggi tersebut [1].

Dengan adanya teknologi informasi, data yang berlimpah dapat diolah agar berguna bagi pihak universitas [2]. Pentingnya pengolahan data mahasiswa bagi pihak universitas untuk mengetahui informasi penting berupa pengetahuan yang baru, seperti informasi yang terkait mengenai profil dan informasi akademik dari mahasiswa tersebut. Salah satu standar mutu dari pendidikan tinggi adalah berdasarkan pada organisasi kemahasiswaan yang artinya perbandingan mahasiswa dan dosen. Harapan untuk pendidikan tinggi adalah mampu menghasilkan alumni dengan kualitas tinggi. Salah satu kriteria untuk kualitas tinggi dapat dijelaskan dengan status kelulusan siswa di sebuah institusi [3].

Kinerja akademik siswa didasarkan pada berbagai faktor seperti variabel pribadi, sosial, psikologis, dan lingkungan lainnya. Penggunaan data mining merupakan metode yang sangat berguna untuk mencapai tujuan ini. Teknik data mining digunakan untuk pengoperasian dalam data yang besar, bertujuan untuk menemukan pola dan hubungan yang tersembunyi dalam membantu pengambilan keputusan [4]. Banyaknya jumlah data yang dimiliki oleh pihak universitas mengenai mahasiswa lulusan mereka dapat membantu dalam proses pengambilan keputusan ke jenjang pendidikan yang lebih tinggi. Data akademik mahasiswa dapat

Table 1. Dataset

Attributes	Details	Value
Serial No	Nomor serial dari setiap rows dalam dataset.	*Integer*
GRE Score	*GRE (Graduate Record Examination)* adalah nilai ujian standar yang seringkali diperlukan untuk masuk ke program pascasarjana secara global.	*Integer*
TOEFL Score	*TOEFL (Test of English as a Foreign Language)* merupakan nilai test keahlian yang digunakan untuk mengukur kemampuan berbahasa inggris seseorang.	*Integer*
University Rating	Tingkatan atau peringkat dari sebuah universitas.	*Integer*
SOP	*SOP (Statement of Purpose)* adalah jenis esai yang ditulis ketika melamar ke program tertentu dalam hal ini program pascasarjana di universitas tertentu.	*Float*
LOR	LOR (Letter of Recommendation) juga merupakan jenis kertas lain yang ditulis saat melamar ke universitas tertentu.	*Float*
CGPA	CGPA (Cumulative Grade Points Average) adalah nilai mahasiswa yang diperoleh di semua mata pelajaran.	*Float*
Research	Merupakan penelitian untuk mencari tahu berapa banyak lulusan yang memiliki pengalaman penelitian.	*Integer*
Chance of Admit	Merupakan kemungkinan atau peluang mahasiswa diterima di suatu universitas.	*Float*

diolah untuk membantu pengambilan keputusan dalam penentuan masuk ke jenjang perguruan tinggi selanjutnya. Untuk melakukan proses pengolahan data tersebut di butuhkan metode data mining yaitu klasifikasi. Klasifikasi merupakan metode yang di gunakan untuk membangun model prediksi. Sebagian besar peneliti menggunakan klasifikasi untuk memprediksi kinerja siswa [5].

Dalam penelitian kali ini algoritma *K-NN* digunakan sebagai acuan untuk menentukan tingkat peluang kelulusan mahasiswa dalam menentukan universitas yang dia pilih. Dasar penentuan mahasiswa tersebut untuk lulus adalah GRE Scores, TOEFL Scores, University Rating, Statement of Purpose, Letter of Recommendation Strength, Undergraduate CGPA, Research Experience, dan Chance of Admit [6].

Penelitian sebelumya oleh kamagi yaitu implementasi data mining dengan algoritma c4.5 untuk memprediksi tingkat kelulusan mahasiswa mendapatkan hasil akurasi dari prediksi kelulusan sebesar 87,5% [7]. Kemudian pada penelitian oleh hastuti yang analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif yang menggunakan perbandingan algoritma logistic regression, decision tree, naïve bayes, dan neural network mendapatkan hasil bahwa algoritma decision tree memiliki akurasi terbaik yaitu sebesar 95,29% [8]. Penelitian terkait lainnya juga oleh ryan yang berjudul prediksi kelulusan mahasiswa berdasarkan kinerja akademik menggunakan pendekatan data mining pada program studi sistem informasi fakultas ilmu computer universitas brawijaya [9].

Kami menggunakan algoritma *k-nearest neighbor (K-NN)* karena algoritma ini memiliki tingkat akurasi yang tinggi. Tujuan dari penelitian ini adalah untuk mengetahui peluang lulus atau tidaknya mahasiswa dalam menentukan universitas pilihannya. Manfaat dari penelitian ini adalah untuk memberikan informasi kepada mahasiswa dalam menentukan universitas yang akan dia pilih.

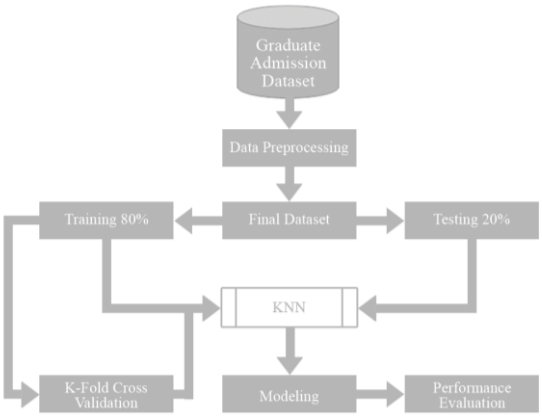
2. Metode Penelitian

Data yang digunakan pada penelitian ini adalah *Graduate Admission Dataset* yang dapat di akses dari Kaggle[10]. *Dataset* ini memiliki 400 *rows*, dan 9 *attributes* yang dijelaskan di Tabel 1.

Design Penelitian seperti pada Gambar 1 memperlihatkan proses *Prediksi Peluang Kelulusan Mahasiswa Dalam Memilih Universitas.* Proses pertama adalah mengambil *Graduate Admission Dataset* yang diambil dari *Kaggle*, dilanjutkan dengan *data*

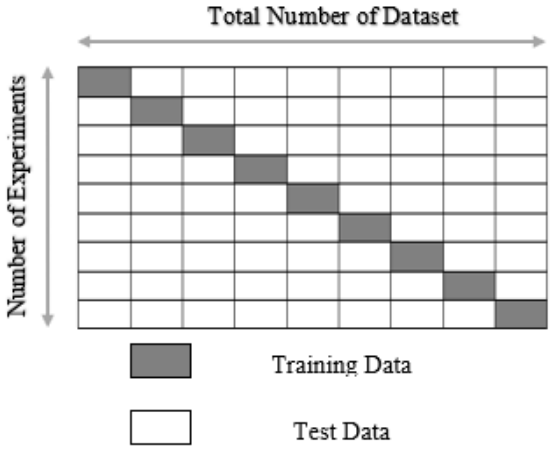
processing yang akan mengolah data untuk mendapatkan *final dataset*. Kemudian *final dataset* dibagi menjadi 80% *training* dan 20% *testing*. Penelitian ini menggunakan *K-Fold Cross Validation* sebelum melakukan *modelling*. Selanjutnya data di proses menggunakan algoritma *k-nearest neighbor (K-NN)*, yang kemudian akan dibuat modelnya dan dievaluasi.

Data preprocessing dibagi menjadi 2 bagian, yaitu *data cleaning* dan *data reduction*. *Data cleaning* adalah proses pembersihan data incomplete pada attribute di dataset untuk membuat data menjadi lebih konsisten. Sedangkan, *data reduction* adalah proses untuk menghapus data pada *attribute* yang kurang dominan sehingga data bisa dikurangi, namun tetap menghasilkan data yang akurat. Jadi data target dan independent akan dipisahkan sebagai X dan Y, lalu data akan dibagi menjadi train dan set tes dengan ukuran set tes 20%. Kemudian pemindaian fitur akan dilakukan pada fitur independent untuk melakukan standarisasi.



Gambar 1. Arsitektur untuk Prediksi Peluang Kelulusan Mahasiswa Dalam Memilih Universitas

Setelah data telah dibagi menjadi 80% *data training* dan 20% *data testing*, maka akan dilakukan *K-Fold Cross Validation* pada *data training*. *Cross Validation* adalah teknik untuk mengevaluasi model dengan cara mempartisi sampel asli ke dalam *training set* untuk melatih model, dan *test set* untuk mengevaluasi model. Dalam *K-Fold Cross Validation*, sampel asli secara acak dipartisi dalam *k equal size subsample*. Dari *subsample* k, satu *subsample* akan digunakan sebagai *testing data* dan sisanya akan menjadi *training data*. Proses *cross validation* akan diulang sebanyak k kali (kelipatan), dengan masing – masing dari *subsample* k digunakan sekali sebagai *validation data* [11]. Pada Gambar 2 menunjukkan proses *K-Fold Cross Validation*, data dibagi menjadi 9 partisi dan akan diuji sebanyak 9 kali sebelum dibuat modelnya.



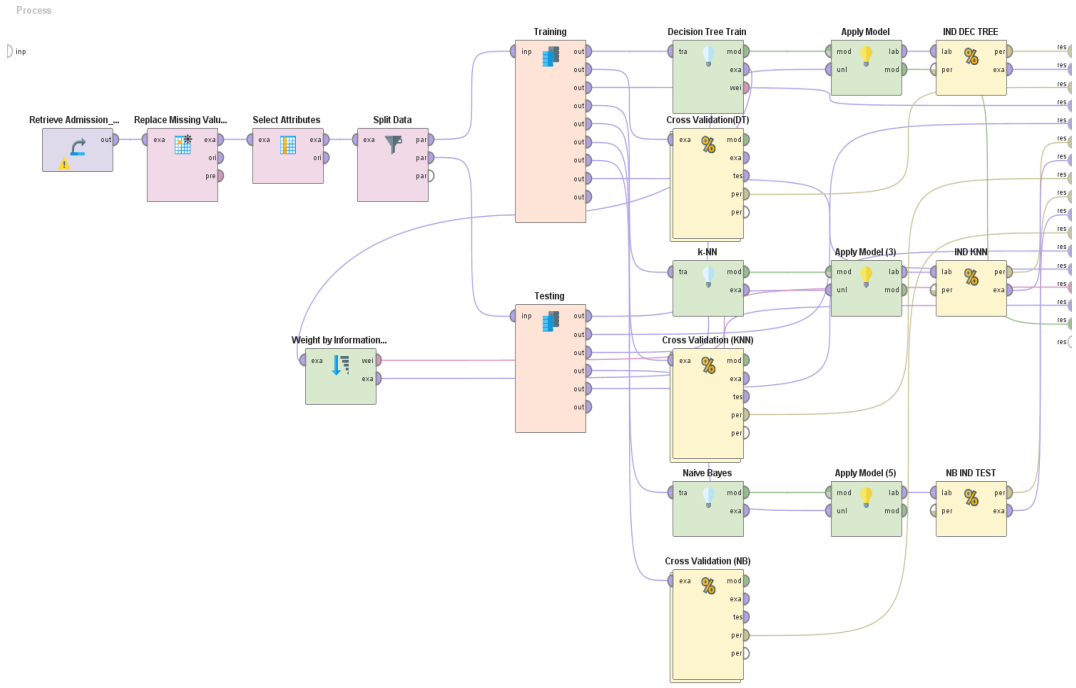
Gambar 2. *K-Fold Cross Validation*

Setelah proses *K-Fold Cross Validation* selanjutnya data akan diproses dengan menggunakan algoritma *K-NN*. Algoritma *K-NN* adalah metode untuk klasifikasi objek berdasarkan *training examples* yang terdekat. *K-NN* adalah *instance-based learning* atau disebut juga *lazy learning*, karena fungsinya hanya didekati secara local dan semua perhitungan ditunda hingga proses klasifikasi. *K-NN* adalah teknik klasifikasi yang paling sederhana ketika tidak ada pengetahuan tentang distribusi data [12]. Jarak dalam *K-NN* dihitung menggunakan *Euclidean Distance* dengan rumus :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}, \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Keterangan:
d : Jarak i : Jumlah data
n : Banyaknya data
x : Titik awal y: Titik akhir

Setelah pembuatan model maka langkah selanjutnya adalah melakukan evaluasi dengan *performance evaluation*. *Performance evaluation* berguna untuk menguji performa dari *classifier*. *Recall*, *precision*, dan *accuracy*. *Recall* adalah kumpulan data positif yang diklasifikasikan dengan benar sebagai data positif. *Precision* adalah kumpulan data yang diklasifikasikan sebagai positif yang benar – benar positif. *Accuracy* adalah ketepatan klasifikasi data [13].



Gambar 3. *Process K-Nearest Neighbor*

Berikut ini adalah rumus *recall*, *precision*, dan *accuracy* dalam *performance evaluation*:

Recall = *(TP / (TP+FN))*,…….(2)
Precision = *(TP / (TP+FP)*,……....(3)
Accuracy=*(TP+TN)/ (TP+TN+FP+FN)*,…….(4)

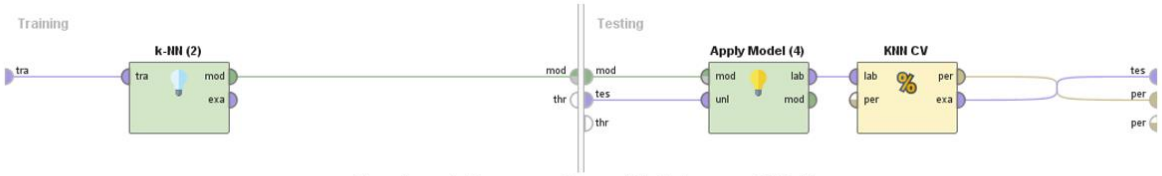
Keterangan :
TP : Nilai true positive TN : Nilai true negative
P : Jumlah data positive FP : Nilai false positive
N : Jumlah data negative FN : Nilai false negative

Metode *feature important* memegang peran penting dalam memilih *attribute* yang siginifikan, melalui penghapusan *attribute* yang tidak relevan, dan oleh karena itu dapat digunakan untuk identifikasi *attribute* yang berpengaruh [14]. Penulis menggunakan metode *information gain ratio* untuk menentukan berapa besar pengaruh suatu *attribute* dalam *dataset*. Machine *learning information gain* dapat digunakan untuk membuat peringkat dari *attributes*.

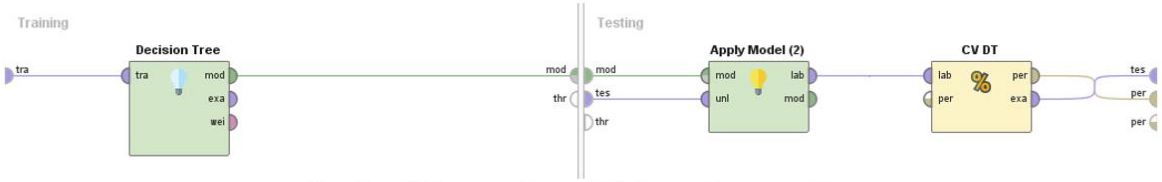
Attributes yang memiliki *information gain* yang tinggi harus diberi peringkat lebih tinggi daripada *attributes* yang lain karena lebih berpengaruh dalam mengklasifikasikan data[15]. Berikut ini adalah rumus dalam *information gain*:

$$IG(A) = H(S) - \sum \frac{Si}{s} H(Si), \ldots\ldots\ldots(5)$$

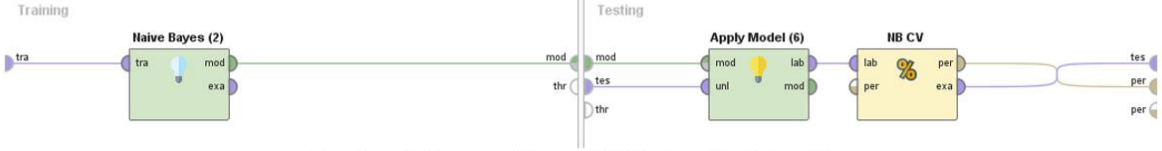
Keterangan:
H(S) : Entropi dari *dataset*
H(Si) : Entropi dari i *subset* yang dihasilkan oleh partisi S
A : Atribut dalam dataset



Gambar 4 *Process Cross Validation K-NN*



Gambar 5 *Process Cross Validation Decision Tree*



Gambar 6 *Process Cross Validation Decision Tree*

3. Hasil dan Pembahasan

Pada bagian ini penulis melakukan analisa terhadap *graduate admission prediction* menggunakan algoritma *k-nearest neighbor (K-NN)*. Gambar 3, 4, 5, dan 6 adalah use case proses pembuatan model dengan menggunakan algoritma KNN pada software Rapidminer.

Berdasarkan Gambar 7 dapat disimpulkan bahwa attribute *CGPA* adalah attribute yang paling berpengaruh dalam memprediksi *graduate admission* dengan hasil weight 0.154. Sedangkan, attribute lain memiliki hasil yaitu *Research 0.034, SOP 0.048, University Rating 0.052, LOR 0.075, TOEFL Score 0.103, dan GRE Score 0.097.*

Table 2 menunjukkan hasil akurasi performance evaluation menggunakan K-Fold Cross Validation. Dari hasil yang telah didapat dengan menggunakan algoritma *k-nearest neighbor (K-NN)*, didapati *accuracy* sebesar 91.88%, *precision* 81.29%, dan *recall* 61.15%.

Tabel 2 *Hasil Perbandingan Algoritma Dengan Menggunakan K-Fold Cross Validation*

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	Error (%)
K-NN	91.88%	81.29%	61.15%	8.13%
Decision Tree	91.56%	75.79%	66.95%	8.44%
Naïve Bayes	85.00%	67.00%	82.76%	15.0%

Tabel 3 menunjukkan hasil akurasi performance evaluation independent test. Dengan menggunakan algoritma *k-nearest neighbor (K-NN)*, didapati *accuracy* sebesar 96.25%, *precision* 98.08%, dan *recall* 70.00%.

Tabel 3 *Hasil Perbandingan Algoritma Dengan Menggunakan Independent Test*

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	Error (%)
K-NN	96.25%	98.08%	70.00%	3.75%
Decision Tree	92.50%	70.06%	77.33%	7.50%
Naïve Bayes	82.50%	63.16%	90.67%	17.50%

Tabel 4 menunjukan hasil perbandingan menggunakan algoritma *K-NN* dengan membandingkan *K-5, K-10, K-15, K-20,* dan *K-100.* Algoritma *K-NN Cross Validation* menggunakan *K-20* merupakan *k* optimal karena memiliki nilai *accuracy* yang tinggi dibandingkan dengan yang lain.

Tabel 4 *Hasil Perbandingan Algoritma K-NN Menggunakan Cross Validation*

K level	Accuracy (%)	Precision (%)	Recall (%)	Error (%)
K-5	90.94%	73.01%	63.62%	9.06%
K-10	91.56%	77.19%	62.47%	8.44%
K-15	91.56%	79.47%	59.48%	8.44%
K-20	91.88%	81.29%	61.15%	8.13%
K-100	90.62%	45.31%	50.00%	9.38%

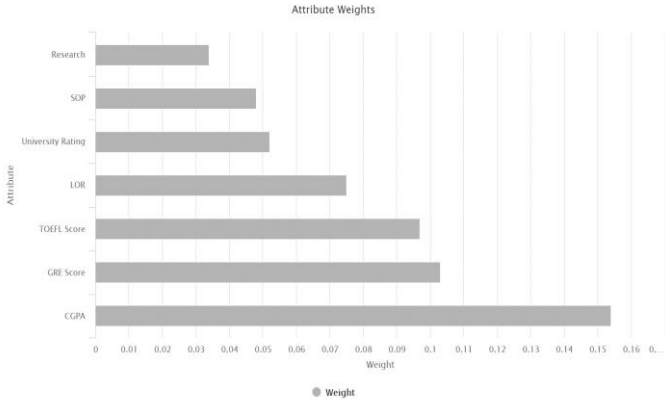
Tabel 5 menunjukan hasil perbandingan menggunakan algoritma *K-NN* dengan membandingkan *K-5, K-10, K-15, K-20,* dan *K-100*. Algoritma *K-NN Independent Test* menggunakan *K-5, K-10, K-15,* dan *K-20* merupakan *k* optimal karena memiliki nilai *accuracy* yang tinggi dibandingkan dengan *K-100.*

Tabel 5 *Hasil Perbandingan Algoritma K-NN Menggunakan Independent Test*

K level	Accuracy (%)	Precision (%)	Recall (%)	Error (%)
K-5	96.25%	98.08%	70.00%	3.75%
K-10	96.25%	98.08%	70.00%	3.75%
K-15	96.25%	98.08%	70.00%	3.75%
K-20	96.25%	98.08%	70.00%	3.75%
K-100	93.75%	46.88%	50.00%	6.25%

4. Kesimpulan

Dari ketiga algoritma yang telah dievaluasi yaitu *K-Nearest Neighbor, Decision Tree, dan Naïve Bayes,* untuk memprediksi *graduate admission* dapat disimpulkan bahwa algoritma *K-NN* memiliki hasil yang paling baik diantara algoritma yang lain dengan *accuracy* sebesar 96.25%, *precision* 98.08%, dan *recall* 70.00 untuk *independent test* sedangkan untuk hasil test yang menggunakan *cross-validation* memiliki *accuracy* sebesar 91.88%, *precision* 81.29%, dan *recall* 61.15%. Untuk kedepannya diharapkan model ini dapat berguna untuk pembuatan aplikasi prediksi *graduate admission,* dan untuk penelitian selanjutnya diharapkan peneliti dapat menggunakan metode dan algoritma lain agar dapat memaksimalkan *performance* dan mengurangi nilai error dalam pemodeling.



Gambar 7 Hasil *Feature Importance*

5. Daftar Pustaka

[1] Nugroho, Y.S., "Data Mining Menggunakan Algoritma Naïve Bayes," *UDiNus Repository,* pp. 1-11, 2014.

[2] Ong, J.O., "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," *Jurnal Ilmiah Teknik Industri,* vol. 12, no. 1, pp. 10-13, 2013.

[3] Kesumawati, A., "Implementation Naïve Bayes Algorithm for Student Classification Based on Graduation Status," *International Journal of Applied Business and Information Systems,* vol. 1, no. 2, pp. 6-12, 2017.

[4] Bhardwaj, B. K. dan Pal, S. , "Data Mining: A prediction for performance improvement using classification," *International Journal of Computer Science and Information Security,* vol. 9, no. 4, pp. 1-5, 2011.

[5] Mokhtar, M., Nawang, H. dan Shamsuddin, S. N. W., "Analysis On Students Permormance Using Naive Bayes Classifier," *Journal of Theoretical and Applied Information Technology,* vol. 95, no. 16, pp. 3993-4000, 2017.

[6] Acharya, M. S., "Kaggle," [Online]. Available: https://www.kaggle.com/mohansacharya/graduate-admissions. [Diakses 24 April 2020].

[7] Kamagi, D. H. dan Hansun, S., "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *ULTIMATICS,* vol. VI, no. 1, pp. 15-20, 2014.

[8] Hastuti, K., "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiwa Non Aktif," *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012,* pp. 241-249, 2012.

[9] Pambudi, R.D dan Supianto, A. A., "Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Brawijaya," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 3, no. 3, pp. 2194-2200, 2019.

[10] Acharya M. S., "Kaggle," [Online]. Available: https://www.kaggle.com/mohansacharya/graduate-admissions. [Diakses 24 April 2020].

[11] "OpenML," [Online]. Available: https://www.openml.org/a/estimation-procedures/1. [Accessed 20 April 2020].

[12] Imandoust, S. B. and Bolandraftar, M., "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *S B Imandoust et al. Int. Journal of Engineering Research and Applications,* vol. 3, no. 5, 2013.

[13] Bramer, M., "Principles of data mining," *Springer,* 2007.

[14] Liem, A. T, Sandag, G. A. Hwang, I. S and Nikoukar, A., "Delay analysis of dynamic bandwidth allocation for triple-play-services in EPON," 2017.

[15] Sui B., "Information Gain Feature Selection Based On Feature Interactions," 2013.