



Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Topik Skripsi Mahasiswa di Fakultas Ilmu Komputer

Application of the K-Nearest Neighbor Algorithm for Classification of Student Thesis Topics at the Faculty of Computer Science

Aditya Elanda Gumanti¹, Taslim^{*2}, Susi Handayani³, Dafwen Toresa⁴

^{1,2,3}Fakultas Ilmu Komputer, Universitas Lancang Kuning, Indonesia

Corresponden E-Mail: taslim@unilak.ac.id

Abstrak

Dalam suatu proses perkuliahan mahasiswa yang sudah menempuh pendidikan cukup lama dan akan menyelesaikannya, maka harus melalui tahapan akhir yang harus dikerjakan. Salah satunya membuat skripsi merupakan tahapan yang harus dilewati semua mahasiswa jika ingin lulus dari universitas. Dalam pengambilan topik skripsi, Fakultas Ilmu Komputer Universitas Lancang Kuning memberikan beberapa pilihan topik atau peminatan yang dapat dipilih oleh mahasiswa. Pemilihan topik atau peminatan tersebut akan lebih baik jika tidak hanya sesuai dengan minat tapi juga sesuai dengan kemampuan masing-masing mahasiswa. K-Nearest Neighbor merupakan salah satu metode yang digunakan dalam pengklasifikasian. Prinsip kerja K-Nearest Neighbor (KNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K-Nearest (neighbor) terdekatnya dalam data pelatihan. Penggunaan K-Nearest Neighbor bertujuan untuk mengklasifikasikan objek baru berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek baru tersebut. Oleh karena itu knn sangat cocok untuk prediksi topik skripsi mahasiswa menggunakan klasifikasi data. Kesimpulan yg didapat dari penelitian ini Optimasi nilai k menggunakan k-fold cross validation menghasilkan tingkat akurasi yaitu 56,67% dengan nilai k-fold cross validation = 2 dan nilai K-5. Dari hasil klasifikasi menggunakan Algoritma K-Nearest Neighbor hasilnya sebanyak 73 mahasiswa berminat mengambil topik skripsi rekayasa perangkat lunak (RPL), 48 mahasiswa berminat mengambil topik skripsi kecerdasan buatan (AI) dan 0 atau tidak ada mahasiswa yang berminat mengambil topik skripsi jaringan.

Keyword – Data Mining, skripsi, klasifikasi, kNN, optimasi

Abstract

In a lecture process, students who have taken education long enough and will complete it, must go through the final stages that must be done. One of them is writing a thesis, which is a stage that all students must pass if they want to graduate from university. In taking the thesis topic, the Faculty of Computer Science, Lancang Kuning University provides several choices of topics or specializations that students can choose from. The choice of topic or specialization will be better if it is not only in accordance with the interests but also in accordance with the abilities of each student. K-Nearest Neighbor is one of the methods used in classification. The working principle of K-Nearest Neighbor (KNN) is to find the shortest distance between the data to be evaluated and the closest K-Nearest (neighbor) in the training data. The use of K-Nearest Neighbor aims to classify new objects based on learning data that is closest to the new object. Therefore, KNN is very suitable for predicting student thesis topics using data classification. The conclusion obtained from this research is that the optimization of the k value using k-fold cross validation produces an accuracy rate of 56.67% with the k-fold cross validation value = 2 and the K-5 value. From the results of the classification using the K-Nearest Neighbor Algorithm, the results are 73 students are interested in taking software engineering thesis (RPL) topics, 48 students are interested in taking artificial intelligence (AI) thesis topics and 0 or no students are interested in network thesis topics.

Keywords - Data Mining, thesis, clasification, kNN, optimization

1. Pendahuluan

Tugas akhir atau skripsi merupakan salah satu syarat yang harus dilaksanakan oleh mahasiswa di sebuah perguruan tinggi untuk dapat lulus menjadi seorang sarjana. Dalam pengambilan topik skripsi, Fakultas

Ilmu Komputer Universitas Lancang Kuning memberikan beberapa pilihan topik atau peminatan yang dapat dipilih oleh mahasiswa. Pemilihan topik atau peminatan tersebut akan lebih baik jika tidak hanya sesuai dengan minat tapi juga sesuai dengan kemampuan masing-masing mahasiswa.

Data mining merupakan sekumpulan teknik untuk menemukan pengetahuan yang sebelumnya tidak diketahui dalam basis data yang besar[1]. Data mining tidak hanya dapat digunakan dalam menemukan pengetahuan atau fenomena baru, melainkan juga untuk meningkatkan pemahaman kita mengenai apa yang kita ketahui. Saat ini Data Mining (DM) telah menarik banyak perhatian di bidang analisis data yang dapat digunakan untuk mengekstrak pengetahuan yang berharga dan bermakna dari data[2]. K-Nearest Neighbor merupakan salah satu metode dalam data mining yang digunakan dalam pengklasifikasian[3]. Klasifikasi merupakan salah satu hal yang penting untuk memprediksi kelas dari sebuah variabel baru[4]. Prinsip kerja K-Nearest Neighbor (kNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K-Nearest (neighbor) terdekatnya dalam data pelatihan[5]. Penggunaan K-Nearest Neighbor bertujuan untuk mengklasifikasikan objek baru berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek baru tersebut[6].

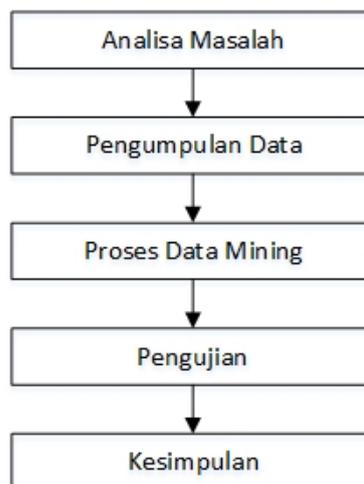
Beberapa penelitian terkait penggunaan kNN untuk prediksi kinerja akademik sudah dilakukan oleh para peneliti. Tyas melakukan klasifikasi prestasi akademik siswa dengan kNN dan kNN dengan Gain Ratio. Hasil penelitian dengan kNN menghasilkan tingkat akurasi sebesar 74.068, sedangkan kNN dengan Gain Ratio menghasilkan tingkat akurasi sebesar 75.105[7]. Nugroho pada percobaan menggunakan metode KNN menghasilkan tingkat akurasi yang cukup baik[8]. Dari hasil eksperimen terhadap 12 mata pelajaran pada pelajaran mengetik yang dilakukan oleh Tanner dimana penelitian ini menggunakan 15000 data siswa, hasilnya menunjukkan bahwa kNN dapat memprediksi kinerja siswa secara akurat. Tes awal pada keterampilan dapat menjadi prediktor kuat untuk skor akhir juga dalam kursus berbasis keterampilan lainnya. Selanjutnya metode ini akan diimplementasikan sebagai fitur peringatan dini bagi guru kursus mengetik, sehingga mereka dapat dengan cepat memusatkan perhatian mereka kepada siswa yang paling membutuhkan bantuan[9].

Pada penelitian ini akan dilakukan proses klasifikasi peminatan topik skripsi mahasiswa berdasarkan data akademik mahasiswa, yaitu dari hasil studi selama proses perkuliahan dari semester satu sampai semester tujuh. Hampir seluruh matakuliah yang diselenggarakan memiliki korelasi dengan topik atau peminatan yang dapat dipilih. Oleh karena itu, dapat dilakukan suatu analisis pada data akademik mahasiswa yang hasilnya dapat membantu untuk menentukan topik skripsi yang sesuai dengan minat dan kemampuan. Dengan adanya penelitian ini yang bisa menentukan topik skripsi sesuai minat dan kemampuan mahasiswa, hasil dari penelitian akan berguna untuk mahasiswa dan juga bagi program studi (prodi) di di fakultas ilmu komputer, hal ini diharapkan juga sebagai salah satu pertimbangan prodi agar bisa mempersiapkan dosen pembimbing yang sesuai dengan topik skripsi yang diambil oleh mahasiswa. Dan dengan topik dan peminatan yang sesuai.

Pada penelitian ini Data Training yang digunakan adalah data transkrip nilai mahasiswa Fakultas Ilmu Komputer Universitas Lancang Kuning Angkatan 2015 dan 2016 sedangkan data yang akan diklasifikasi adalah data transkrip nilai mahasiswa Angkatan 2018.

2. Metode Penelitian

Pada penelitian ini tahapan yang dilakukan yaitu dimulai dari analisa masalah sampai dengan tahapan penerikan kesimpulan dari hasil penelitian. Detail tahapan penelitian dapat dilihat pada gambar 1 berikut.



Gambar 1 Tahapan Tahapan Penelitian

1. Analisa Masalah

Analisa masalah dilakukan untuk memahami masalah yang telah ditentukan batasan dan ruang lingkungannya. Dengan dilakukan analisa masalah tersebut, Maka diharapkan masalah dapat dipahami dengan baik. Adapun tahapan-tahapan analisis yang digunakan yaitu mengidentifikasi permasalahan yang terjadi, memahami permasalahan dengan cara melakukan pengumpulan data yang dibutuhkan, mencari kriteria yang digunakan untuk proses klasifikasi dan mengumpulkan informasi tentang kebutuhan yang diperlukan dalam penelitian.

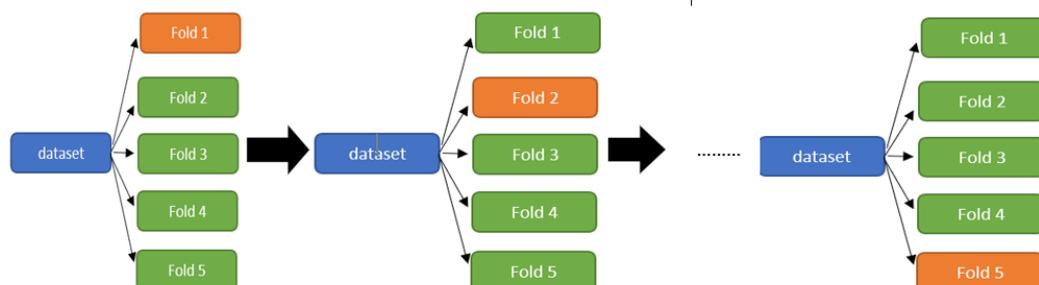
2. Pengumpulan Data

Pada tahapan ini untuk memperoleh data dan informasi yang akurat yang dapat menunjang dalam proses penelitian. Berikut adalah metode pengumpulan data yaitu sebagai berikut:

- Studi Literatur, yaitu pengumpulan data dan informasi dengan cara mengumpulkan literatur, jurnal dan buku yang berhubungan dengan topik penelitian.
- Observasi, adalah aktivitas ataupun pengamatan langsung terhadap tempat penelitian dengan maksud mendapatkan serta bisa mengetahui permasalahan yang ada dalam tempat penelitian.
- Wawancara, yaitu proses tanya jawab dengan pihak yang bersangkutan di Fakultas Ilmu Komputer Universitas Lancang Kuning untuk pengumpulan data yang dibutuhkan. Dalam penelitian ini peneliti akan melakukan tanya jawab dengan Kepala Program Studi Teknik Informatika, Kepala Program Studi Sistem Informasi dan Pihak Pihak yang bersangkutan lainnya.

3. Proses Data Mining

Pada tahapan ini peneliti menyeleksi data data yang diperlukan seperti data training dan data yang akan diklasifikasi, selanjutnya menerapkan algoritma kNN untuk mendapatkan prediksi tentang topik skripsi mahasiswa yang sesuai dengan minat dan kemampuan mahasiswa berdasarkan data data akademik mahasiswa. Akurasi hasil prediksi dioptimalkan dengan algoritma *k fold cross validation*[4]. *K-Fold cross validation* bekerja dengan cara membagi kumpulan data menjadi sejumlah K bagian di mana setiap lipatan digunakan sebagai kumpulan pengujian di beberapa titik. Pada iterasi pertama, lipatan pertama digunakan untuk menguji model dan sisanya digunakan untuk melatih model. Pada iterasi kedua dan seterusnya fold digunakan sebagai testing set sedangkan sisanya sebagai training set. Proses ini diulang sampai setiap lipatan telah digunakan sebagai set pengujian. Berikut gambaran dari proses k-fold cross validation (Gambar 2).



Gambar 2. K-fold cross validation

4. Pengujian

Pada tahapan ini proses pengujian, dimana proses ini menggunakan aplikasi rapidminer. Proses pengujian ini dilakukan untuk menguatkan hasil dari proses data mining yang dihitung secara manual ditahapan sebelumnya.

5. Kesimpulan

Pada tahap ini penulis dapat mengambil kesimpulan dari hasil penelitian yang telah dilakukan.

3. Hasil Dan Pembahasan

3.1. Data Training

Untuk data training digunakan data mahasiswa Angkatan 2015 dan 2016 yang sudah lulus dan mempunyai topik skripsi. Data training yang digunakan di penelitian ini berjumlah 150 data. Adapun data training yang digunakan dapat dilihat pada tabel dibawah ini:

Tabel 1 Data Training

No	NIM	Nama	Jaringan	AI	RPL	Topik Skripsi
----	-----	------	----------	----	-----	---------------

1	1655201001	Ayub Christian	4	3,75	4,333	AI
2	1655201003	Junaidi	3,5	2,25	2,916	AI
3	1655201004	Andi Wibowo	3,25	2,75	4,208	RPL
4	1655201005	Ave Ibnu Yudha	3,5	3,5	4,25	AI
5	1655201006	Bayu Yuswa	4	3,75	4,666	RPL
6	1655201007	Asia Tarigan	3,5	2,75	3,583	AI
7	1655201009	Yogi Yunnizer	2,75	3,75	3,791	Jaringan
...						
150	1655201087	Zulfian	3,75	3,75	3,625	AI

Tabel 1 diatas merupakan data training yang akan digunakan dalam penelitian ini, Adapun data training tersebut terdiri dari 7 Topik Skripsi Jaringan, 21 Topik Skripsi Kecerdasan Buatan dan 22 Topik Skripsi Rekayasa Perangkat Lunak.

3.2. Data Testing

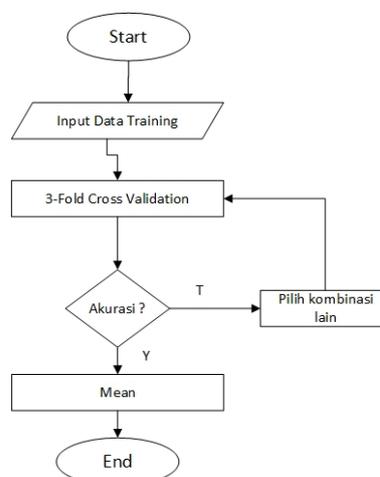
Untuk data testing, data yang digunakan data mahasiswa Angkatan 2018. Data testing yang digunakan di penelitian ini berjumlah 120 data. Adapun data testing yang digunakan dapat dilihat pada tabel dibawah ini.

Tabel 2 Data Testing

No	NIM	Nama	Jaringan	AI	RPL
1	1855201001	Dewi Puspita Sari	3,5	4	3,83
2	1855201147	Lusiana Dwi Lestari	2,75	3,5	4
3	1855201003	Aditya Wiraguna Feriza	3,75	3,75	4
4	1855201004	Aprizaldi	3,25	3	3,33
5	1855201148	Monica Sitompul	3,5	2,75	3,5
6	1855201006	Novaldy Fiqri	3,25	3	3,33
7	1855201007	Bayu Kurniawan	3,75	2,25	3,83
...					
120	1855201141	Iqbal Harifal	3,75	3,75	3,62

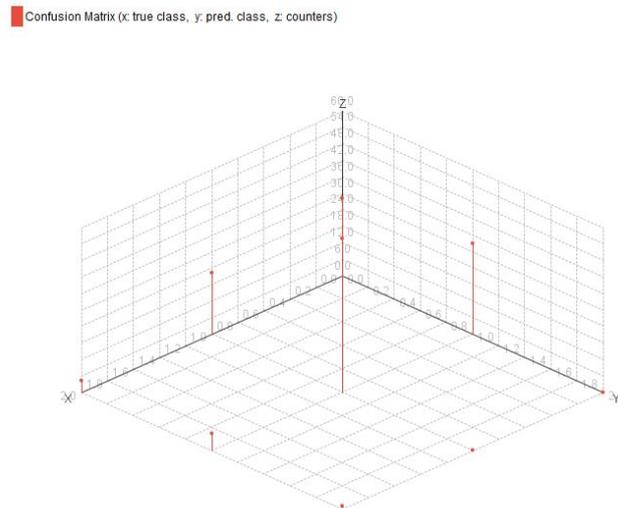
3.3. Optimasi Nilai K

Optimasi nilai k dilakukan dengan algoritma *k fold validation process*. Dengan jumlah k fold yaitu 3. Untuk algoritma k-fold validation proses dapat dilihat pada flowchart gambar 2 berikut.



Gambar 2 Flowchart Cross Validation

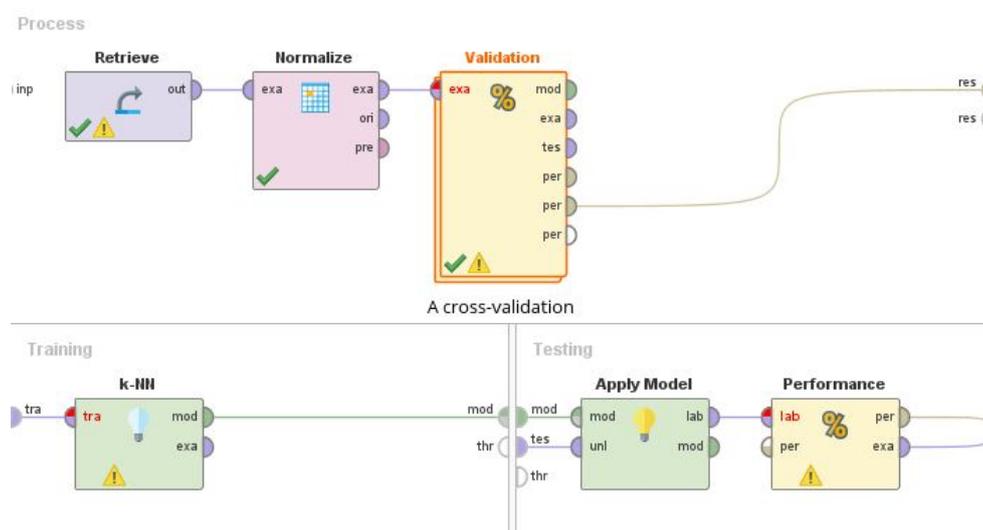
Dari hasil 2-fold cross-validation maka diperoleh nilai akurasi yaitu sebesar 56,67% dengan nilai k yaitu sebesar 5 (k=5). Nilai ini selanjutnya akan digunakan sebagai nilai k untuk untuk proses kNN. Berikut hasil dari 2-fold cross-validation (Gambar 3)



Gambar 3 Grafik Tingkat Akurasi Nilai K

3.4. Proses kNN

Proses kNN di lakukan dengan menggunakan RapidMiner. Langkah pertama melihat berapa tingkat akurasi yang dihasilkan dari 2-fold cross-validation dengan nilai k=5. Adapun tahapan cross validation dengan RapidMiner dapat dilihat pada gambar 4 berikut.



Gambar 4 Tahapan RapidMiner Melihat Tingkat Akurasi

Dari hasil 2-fold cross-validation diatas dengan nilai k=5 maka didapat tingkat akurasi yaitu sebesar 56,67% dengan Confussion Matrix seperti terlihat pada gambar 5 berikut.

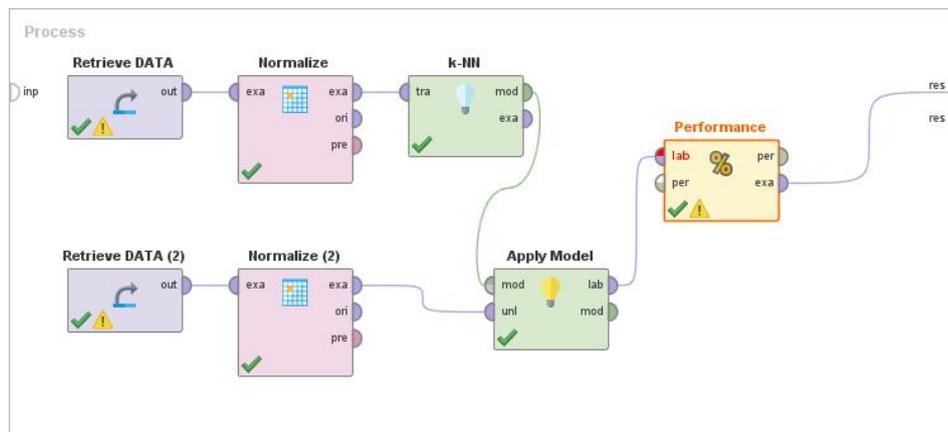
accuracy: 56.67% +/- 2.83% (micro average: 56.67%)

	true AI	true RPL	true Jaringan	class precision
pred. AI	28	22	4	51.85%
pred. RPL	33	56	6	58.95%
pred. Jaringan	0	0	1	100.00%
class recall	45.90%	71.79%	9.09%	

Gambar 5 Confussion Matrix

3.5. Prediksi

Uji prediksi dilakukan dengan menggunakan RapidMiner. Pada tahapan prediksi ini data testing akan diuji dengan data training untuk melihat hasil prediksi terhadap prestasi akademik dengan nilai $k=5$. Adapun alur diagram dari proses ini dapat dilihat pada gambar 6 berikut ini.



Gambar 6 Tahapan RapidMiner Prediksi Data

Adapun alur proses dari gambar 6 dimulai dari proses memasukkan data training dan data testing kemudian selanjutnya diikuti dengan proses normalisasi dari kedua data tersebut. Berikutnya data training akan diproses dengan algoritma kNN untuk selanjutnya kedua data ini akan di *apply* dalam sebuah model. Dari hasil *apply* ini akan dihitung nilai performance yang dihasilkan yaitu berupa nilai akurasi, tabel *confusion matrix* dan tabel prediksi untuk topik skripsi.

Adapun hasil uji prediksi dapat dilihat pada tabel 3 dibawah ini:

Tabel 3 Hasil Prediksi

No	Nama Mahasiswa	Confidence (Jaringan)	Confidence (AI)	Confidence (RPL)	Prediction (Topik Skripsi)
1	Dewi Puspita Sari	0,000	0,388	0,612	RPL
2	Lusiana Dwi Lestari	0,188	0,591	0,221	AI
3	Aditya Wiraguna Feriza	0,000	0,400	0,600	RPL
4	Aprizaldi	0,000	0,635	0,365	AI
5	Monica Sitompul	0,000	0,820	0,180	AI
6	Novaldy Fiqri	0,000	0,635	0,365	AI
7	Bayu Kurniawan	0,000	0,804	0,196	AI
8	Arief Darmawan	0,000	0,211	0,789	RPL
9	Sya'banu Ahmad	0,000	0,384	0,616	RPL
10	Farid Kurnia Akbar	0,200	0,200	0,600	RPL
...					
120	Iqbal Harifal	0,8132	0,1868	0,0000	AI

Dari tabel 3 hasil uji prediksi dapat dilihat nilai confidence dari masing masing *class* label yang terdiri atas tiga kategori *class* label yaitu jaringan, kecerdasan buatan dan rekayasa perangkat lunak. Nilai prediksi dari setiap variabel merupakan nilai *confidence* tertinggi yang diperoleh dari tertinggi dari masing masing *class* variabel.

4. Kesimpulan

Dari hasil uraian dalam pembuatan laporan ini, maka penulis dapat menyimpulkan sebagai berikut:

1. Optimasi nilai k menggunakan k -fold cross validation menghasilkan tingkat akurasi klasifikasi untuk prediksi yaitu 56,67% dengan nilai k -fold cross validation = 2 dan nilai $K=5$.
2. Dari hasil klasifikasi menggunakan Algoritma K-Nearest Neighbor hasilnya sebanyak 73 mahasiswa berminat mengambil topik skripsi rekayasa perangkat lunak (RPL), 48 mahasiswa berminat mengambil

topik skripsi kecerdasan buatan (AI) dan 0 atau tidak ada mahasiswa yang berminat mengambil topik skripsi jaringan

References

- [1] M. Rafi, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," vol. 12, no. 2, pp. 121–129, 2020.
- [2] A. Mohamed, A. Rizaner, and A. Hakan, "Using data Mining to Predict Instructor Performance," *Procedia - Procedia Comput. Sci.*, vol. 102, no. August, pp. 137–142, 2016, doi: 10.1016/j.procs.2016.09.380.
- [3] J. Elias, V. Ferreira, C. Henrique, R. M. De Miranda, and A. F. De Figueiredo, "Q uímica educación the representative elements," *Educ. Química*, vol. 26, no. 3, pp. 195–201, 2015, doi: 10.1016/j.eq.2015.05.004.
- [4] T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015, doi: 10.1016/j.patcog.2015.03.009.
- [5] W. Li, Y. Chen, and Y. Song, "Knowledge-Based Systems Boosted K-nearest neighbor classifiers based on fuzzy granules ☆," *Knowledge-Based Syst.*, no. xxxx, p. 105606, 2020, doi: 10.1016/j.knosys.2020.105606.
- [6] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001.
- [7] T. Setiyorini, R. T. Asmono, T. Informatika, and T. Informatika, "IMPLEMENTATION OF GAIN RATIO AND K-NEAREST NEIGHBOR," vol. 16, no. 1, pp. 19–24, 2020, doi: 10.33480/pilar.v16i1.813.
- [8] A. Nugroho, O. R. Riady, A. Calvin, and D. Suhartono, "Identification of Student Academic Performance using the KNN Algorithm," vol. 2, no. 3, pp. 115–122, 2020.
- [9] T. Tanner, H. Toivonen, and T. Master, "Predicting and preventing student failure – using the k -nearest neighbour method to predict student performance in an online course environment."