Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Optimasi Nilai *k* Pada Algoritma k Nearest Neighbor Untuk Prediksi Akademik Mahasiswa Yang Bekerja

Taslim[1], Yuhelmi[2], Dafwen Toresa[3]
taslim.malano@gmail.com, yuhelmi@unilak.ac.id, dafwen@unilak.ac.id
Universitas Lancang Kuning

Informasi Artikel

Diterima : 08-07-2021
Direview : 13-08-2021
Disetujui : 2909-2021

Kata Kunci

prediksi, K nearest neighbor, optimasi, k-fold cross validation

Abstrak

Sebuah lembaga pendidikan akan selalu fokus bagaimana meningkatkan kualitas akdemik dari peserta didik mereka. Penelitian ini bertujuan untuk melakukan klasifikasi dan prediksi terhadap prestasi akademik mahasiswa terutama bagi mahasiswa yang bekerja karena mereka mempunyai beban yang lebih dibanding mahasiswa yang tidak bekerja. Hasil dari prediksi ini selanjutnya dapat digunakan sebagai salah satu bahan pertimbangan bagi pihak akademik dalam mengambil kebijakan terhadap mahasiswa yang sudah bekerja. Prediksi prestasi akademik dilakukan dengan menggunakan algoritma K nearest neighbor dengan optimasi pada nilai k dengan algoritma *k-fold cross validation* dengan *5-fold cross validation*. Kelas label terdiri atas 3 kategori yaitu memuaskan, sangat memuaskan dan dengan pujian. Dari hasil penelitian didapat nilai k= 3. Uji akurasi performance menghasilkan nilai sebesar 85,71%.

Keywords

prediction, K nearest neighbor, optimization, k-fold cross validation

Abstrak

An educational institution will always focus on how to improve the academic quality of their students. This study aims to classify and predict student academic achievement, especially for students who work because they have more burdens than students who do not work. The results of this prediction can then be used as one of the considerations for academics in making policies towards students who are already working. Prediction of academic achievement is done using the K nearest neighbor algorithm with optimization on the value of k with the k-fold cross validation algorithm with 5-fold cross validation. The label class consists of 3 categories, namely satisfactory, very satisfying and with praise. From the results of the study, the value of k = 3. The performance accuracy test resulted in a value of 85.71%.

A. Pendahuluan

Fokus utama dari sebuah lembaga Pendidikan adalah bagaimana menyediakan Pendidikan yang berkualitas bagi peserta didik mereka agar dapat meningkatkan kinerja akademik peserta didik[1]. Penelitian ini bertujuan untuk melakukan klasifikasi dan prediksi terhadap prestasi akademik mahasiswa terutama bagi mahasiswa yang bekerja karena mereka mempunyai beban yang lebih dibanding mahasiswa yang tidak bekerja. Hasil dari klasifikasi dan prediksi ini selanjutnya dapat digunakan sebagai salah satu bahan pertimbangan bagi pihak akademik dalam mengambil kebijakan terhadap mahasiswa yang sudah bekerja.

Data mining adalah metode untuk menggali informasi yang sebelumnya tak terlihat dalam sebuah database yang besar. Data mining membantu dalam menganalisa pola karakter yang akan datang yang memungkinkan sebuah instansi untuk membuat keputusan yang tepat. Analisis data dimulai dari proses menganalisa, membersihkan dan memodelkan data untuk menghasilkan pengetahuan dan kesimpulan yang bermamfaat [2]

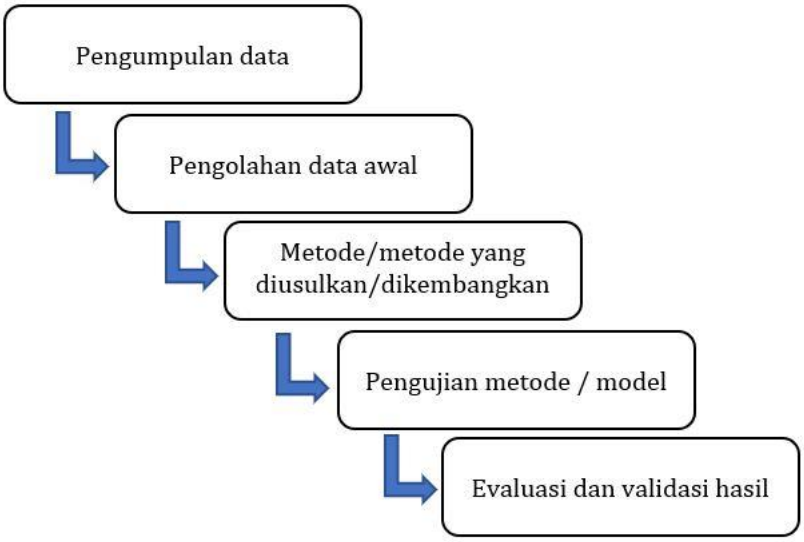
Dalam data mining salah satu metode klasifikasi yang populer adalah algoritma K-nearest neighbor. Algoritma pengelompokkan pengenalan pola yang bersifat nonparametrik, dan telah banyak digunakan pada berbagai bidang karena simple, efektif dan intuitif[3]. Algoritma ini merupakan algoritma klasifikasi yang sederhana namun efektif untuk memprediksi label kelas dari sebuah query berdasarkan informasi di sekitarnya[3]. Akan tetapi algoritma ini memiliki kelemahan pada pemilihan nilai *k.*

Pemilihan nilai k pada algoritma kNN adalah hal yang sangat penting, nilai "k", yang merupakan jumlah tetangga berdasarkan metrik jarak akan sangat mempengaruhi klasifikasi[4]. Berbagai teknik telah diusulkan untuk pemilihan nilai k seperti *cross validation* dan *heuristik*. Nilai k tidak boleh kelipatan dari jumlah kelas untuk menghindari suara denga jumlah yang sama. Disamping itu nilai k yang besar mengurangi efek noise pada klasifikasi dan membuat batas kelas menjadi kurang jelas, sedangkan nilai k yang kecil akan membuat hasil klasifikasi akan terpengruh oleh noise [5].

Salah satu algoritma yang popular untuk evaluasi kinerja set data klasifikasi adalah k-fold croos validation[6]. *K-Fold cross validation* adalah di mana kumpulan data yang diberikan dibagi menjadi sejumlah K bagian di mana setiap lipatan digunakan sebagai kumpulan pengujian di beberapa titik.

B. Metode Penelitian

Dalam penelitian ini data sumber yang digunakan adalah data yang berasal hasil mahasiswa Fakultas Ilmu Komputer yang kuliah sambal bekerja dari Angkatan 2015 sampai dengan Angkatan 2019. Data akan dibagi menjadi 2 bagian yaitu data training dan data testing. Selanjutnya dengan diakukan optimasi pada nilai k dengan menggunakan algoritma *k-fold cross validation.* Dalam penelitian ini akan dilakukan beberapa langkah-langkah atau tahapan penelitian seperti yang digambarkan pada gambar 1.



Gambar 1 . Metode Penelitian

1. Pengumpulan data

Sumber data didapat dari hasil survey terhadap mahasiswa Fakultas Ilmu Komputer Universitas Lancang Kuning yang kuliah sambal bekerja dari Angkatan 2015 sampai dengan Angkatan 2019. Data akan dibagi menjadi 2 bagian yaitu data training dan data testing. Untuk data training digunakan data mahasiswa Angkatan 2015 sampai dengan mahasiswa Angkatan 2018, sedangkan Untuk data testing adalah data mahasiswa Angkatan 2019 sebanyak 30 data. Attribut data yang dikumpulkan adalah nim, nama, IP semester 1, IP semester 2, IP semester 3, IP semester 4, IP semester 5, jumlah jam kerja perminggu dan jumlah pendapatan perbulan.

2. Pengolahan Data Awal

Data *Preprocessing* merupakan tahapan yang sangat penting untuk merancang model klasifikasi[7]. Pengolahan data awal dilakukan untuk memperoleh data yang bersih yang bebas dari noise atau oulier. Beberapa tahapan yang dilakukan yaitu :

a. Data validation
   Data validation dilakukan untuk mengidentifikasi adanya data noise atau outlier, data yang tidak lengkap dan data yang tidak konsisten.
b. Data transformation
   Dalam penelitian ini terdapat beberapa data kategorikal, selanjutnya data data ini di transformasikan kedalam bentuk data numerik.
c. Normalisasi data
   Dalam penelitian ini digunakan normalisasi min-max. Normalisasi min-max melakukan transformasi linier ke data asli[8] yang bertujuan untuk mendapatkan nilai attibut yang seimbang atau untuk menghasilkan data dalam rentang tertentu. Rentang nilai yang tidak seimbang pada setiap atribut dapat mempengaruhi kualitas hasil data mining.

3. Metode yang diusulkan
   Penelitian ini akan dilakukan menggunakan metode K-Nearest Neighbor untuk melakukan klasifikasi dan prediksi terhadap data mahasiswa Fakultas ilmu komputer Universitas Lancang Kuning yang kuliah sambil bekerja. Untuk optimasi nilai *k* dilakukan dengan *k-fold cross validation.* Cross validation silang sejauh ini merupakan salah satu metode yang paling umum digunakan untuk memperkirakan kompleksitas sebuah model[9].
4. Pengujian Model.
   Dalam pengujian model terlebih dahulu akan dilakukan optimasi terhadap nilai *k* dengan menggunakan 5-fold cross validation, selajutnya nilai tersebut akan di gunakan sebagai nilai *k* dalam proses kNN. Tahap selanjutnya dilakukan proses uji kinerja untuk melihat hasil akurasi dari klasifikasi. Tahap terkhir adalah melakukan uji predikasi terhadap data testing.
5. Evaluasi dan Validasi
   Validasi adalah tahapan yang sangat penting pemodelan, untuk melihat sejauh mana kehandalan model yang akan digunakan dalam hal pengambilan keputusan[10]. Untuk evaluasi dan validasi dari hasil klasifikasi dilihat dari hasil confusion matrik. Sedangkan untuk hasil prediksi ditampilkan dalam bentuk tabel hasil prediksi.

C. Hasil dan Pembahasan

a. Data Training

Untuk data training digunakan data mahasiswa Angkatan 2015 sampai dengan mahasiswa Angkatan 2018. Adapun data training yang digunakan dapat dilihat pada tabel 1.

Tabel 1. Data Training

No	NIM	Nama	IPS 1	IPS2	IPS3	IPS4	IPS5	Jam Kerja	Pen dap atan	IP	prestasi akademik
1	1557201049	Franko Taratilo Siahaan	2,85	2,87	2,92	2,68	2,44	2	1	2,75	memuaskan
2	1655201005	Ave Ibnu Yudha	3,60	3,86	3,75	3,63	3,50	5	4	3,67	dengan pujian
3	1655201101	Andri yuslim	2,40	2,40	1,50	2,27	2,07	5	2	2,13	memuaskan
4	1755201067	Willy prastyo	3,05	3,04	3	2,50	3,60	5	4	3,04	sangat memuaskan
5	1755201068	Taufiq Hidayat	3,55	2,88	3,14	2,86	3,15	5	3	3,12	sangat memuaskan
….	…….	…….	….	…	….	….	…..	……	….	….	…….
35	1857201080	Dira septi mulyani	3,0	2,98	3,18	3,34	3,43	1	1	3,19	sangat memuaskan

Adapun transformasi data untuk data jam kerja perminggu dan data pendapatan dapat dilihat pada tabel 2 dan tabel 3 berikut :

Tabel 2. Tranformasi data jam kerja perminggu

Jumlah jam kerja	Transformasi
1 jam s/d 10 jam perminggu	1
11 jam s/d 20 jam perminggu	2
21 jam s/d 30 jam perminggu	3
31 jam s/d 40 jam perminggu	4
lebih daro 40 jam perminggu	5

Tabel 3. Tranformasi data pendapatan

Jumlah jam kerja	Transformasi
Rp. 1.000.000 s/d Rp.1.500.000	1
Rp. 1.500.000 s/d Rp. 2.500.000	2
Rp. 2.500.000 s/d Rp.3.500.000	3
Lebih dari Rp. 3.500.000	4

b. Data Testing

Untuk data testing adalah data mahasiswa Angkatan 2019 sebanyak 30 data, adapun data testing dapat dilihat pada tabel.4 berikut :

Tabel 4. Data Testing

No	NIM	Nama	Indek Prestasi Semester 1	Indek Prestasi Semester 2	Indek Prestasi Semester 3	Jumlah Jam Kerja Perminggu	Pendapatan
1	1955201002	Sanja Saputra	3,5	3,65	3,88	1	1
2	1955201003	Dendy Ferianto	3,37	3,26	3,63	2	1
3	1955201008	Aldi Ardian Syah Daulay	2,74	3,6	3,5	5	1
4	1955201009	Mhd aswin	3,58	3,26	3,38	1	1
…	……..	……..	…..	…..	….	…..	…..
30	1957201069	Raihan Ramadhan	3,58	3,26	3,38	1	1

c. Normalisasi Data

normalisasi adalah salah satu teknik preprocessing yang digunakan untuk menangani rentang nilai atribut, agar data tersebar dengan skala yang sama[7]. Pada penelitian ini proses normalisasi dilakukan dengan metode Min Max dengan rentang nilai antara 0 dan 1. Hasil dari normalisasi terhadap data training dan data testing dapat dilihat pada tabel 5 dan tabel 6 berikut.

Tabel 5. Normalisasi Data training

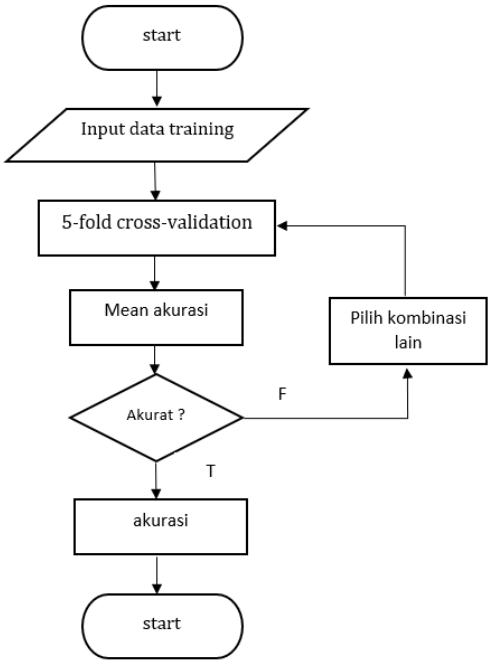
no	NIM	IPS 1	IPS 2	IPS 3	IPS 4	IPS 5	Jam Kerja Perminggu	Pendapatan	IP	prestasi akademik
1	1557201049	0.28	0.34	0.63	0.23	0.61	0.25	0	0.37	memuaskan
2	1655201005	0.75	0.99	1	0.78	0.87	1	1	0.93	dengan pujian
3	1655201101	0	0.03	0	0	0.51	1	0.33	0	memuaskan
4	1755201067	0.40	0.45	0.66	0.13	0.9	1	1	0.55	sangat memuaskan
5	1755201068	0.71	0.34	0.72	0.34	0.78	1	0.66	0.60	sangat memuaskan
….	……	…..	….	…..	…..	…..	….	….	…	……
35	1857201080	0.37	0.41	0.74	0.61	0.85	0	0	0.64	sangat memuaskan

Tabel 6. Normalisasi Data testing

No	NIM	Indek Prestasi Semester 1	Indek Prestasi Semester 2	Indek Prestasi Semester 3	Jumlah Jam Kerja Perminggu	Pendapatan
1	1955201002	0.6	0.79	0.86	0	0
2	1955201003	0.5	0.18	0.56	0.25	0
3	1955201008	0	0.71	0.41	1	0
….	…….	……	…….	……..	…….	….
30	1957201061	1	0.73	0.56	0.5	0

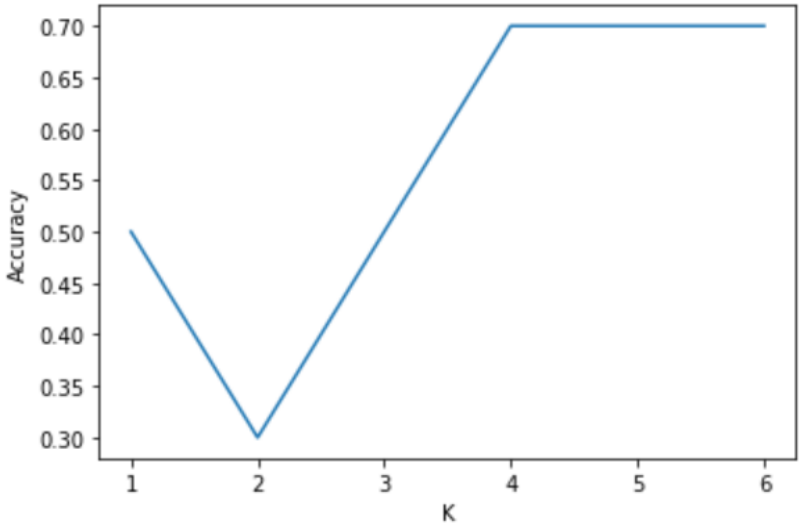
d. Optimasi nilai *k*

Optimasi nilai *k* dilakukan dengan algoritma *k fold validation proses.* Dengan jumlah k fold yaitu 5. Untuk algoritma k fold validation proses dapat dilihat pada flowchart berikut (Gambar 2) .



Gambar 2. Flowchart k-fold cross validation

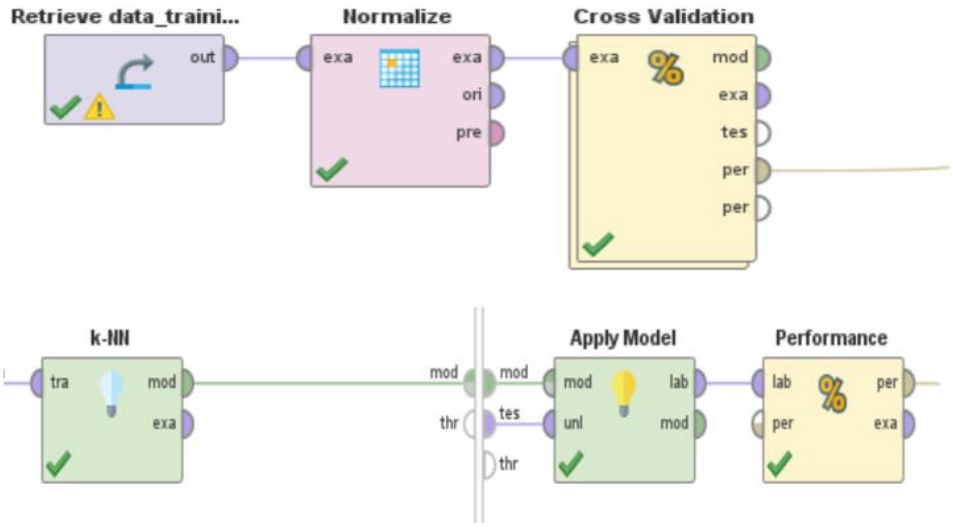
Dari hasil 5-fold cross-validation maka diperoleh nilai akurasi yaitu sebesar 0,55 dengan nilai *k* yaitu sebesar 3 (k=3). Nilai ini selanjutnya akan digunakan sebagai nilai k untuk untuk proses kNN. Berikut hasil dari 5-fold cross-validation (Gambar 3)



Gambar 3 . Grafik tingkat akurasi nilai k

e. Proses kNN

Proses kNN di lakukan dengan menggunkan RapidMiner. Langkah pertama melihat berapa tingkat akurasi yang dihasilkan dari 5-fold cross-validation dengan nilai k=3. Adapun tahapan cross validation dengan RapidMiner dapat dilihat pada gambar 4 berikut.



Gambar 4 . Tahapan Rapidminer melihat tingkat akurasi

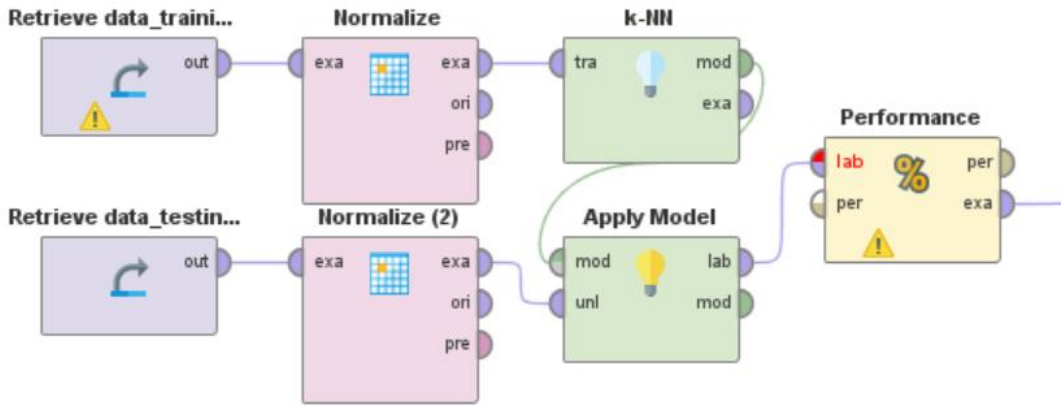
Dari hasil 5-fold cross-validation diatas dengan nilai k=3 maka didapat tingkat akurasi yaitu sebesar 85,71% dengan *Confusion Matrix* sebagai berikut (Tabel 7).

Tabel 7. Confusion Matrix

	True memuaskan	True dengan pujian	True sangat memuaskan	Class precision
Pred. memuaskan	1	0	0	100%
Pred. dengan pujian	0	10	1	90,91%
Pred. sangat memuaskan	3	1	19	82,61%
Class recall	25%	90,91%	95%	

f. 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=3. Adapun alur diagram dari proses ini dapat dilihat pada gambar 5.



Gambar 5 . Tahapan Rapidminer untuk prediksi data

Adapun hasil uji prediksi dapat dilihat pada tabel 8.

Tabel 8. Hasil uji prediksi terhadap data testing

No	nim	confidence (memuaskan)	Confidence (dengan pujian)	Confidence (sangat memuaskan)	Prediksi (prestasi akademik)
1	1955201002	0	0.602	0.397	dengan pujian
2	1955201003	0.395	0	0.604	sangat memuaskan
3	1955201008	0.607	0	0.392	memuaskan
4	1955201009	0.396	0	0.603	sangat memuaskan
.....	.....	.....	.....	.....	.....
30	1957201069	0.396	0	0.603	sangat memuaskan

D. Simpulan

Data yang digunakan pada penelitian ini adalah data mahasiswa Fakultas Ilmu Komputer Universitas Lancang Kuning yang kuliah sambil bekerja, untuk data training adalah data mahasiswa Angkatan 2015 sampai dengan Angkatan 2018 sedangkan untuk data testing adalah data mahasiswa Angkatan 2019 yang terdiri atas 3 label yaitu memuaskan, sangat memuaskan dan dengan pujian. Optimasi nilai menggunakan *5-fold cross validation* menghasilkan tingkat akurasi yaitu 0,55 dan nilai k=3. Hasil uji *performance* dengan *confusion matrik* menghasilkan nilai sebesar 85,71%.

g. Ucapan Terima Kasih

Ucapan terima kasih disampaikan kepada semua pihak yang telah mendukung terlaksananya penelitian ini, terutama kepada Fakultas Ilmu Komputer Universitas Lancang Kuning baik itu tim peneliti, pimpinan dan mahasiswa Fakultas Ilmu Komputer Universitas Lancang Kuning.

h. Referensi

[1] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2021.05.646.

[2] P. Kamath, P. Patil, S. S, Sushma, and S. S, "Crop Yield Forecasting using Data Mining," *Glob. Transitions Proc.*, pp. 0–7, 2021, doi: 10.1016/j.gltp.2021.08.008.

[3] Z. Pan, Y. Wang, and Y. Pan, "A new locally adaptive k-nearest neighbor algorithm based on discrimination class," *Knowledge-Based Syst.*, vol. 204, 2020, doi: 10.1016/j.knosys.2020.106185.

[4] A. Jhamtani, R. Mehta, and S. Singh, "Size of wallet estimation : Application of K-nearest neighbour and quantile regression," *IIMB Manag. Rev.*, pp. 0–19, 2021, doi: 10.1016/j.iimb.2021.09.001.

[5] W. Cherif, "Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis," *Procedia Comput. Sci.*, vol. 127, pp. 293–299, 2018, doi: 10.1016/j.procs.2018.01.125.

[6] 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.

[7] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst. Appl.*, vol. 106, pp. 252–262, 2018, doi: 10.1016/j.eswa.2018.04.008.

[8] A. Ali and N. Senan, "The Effect of Normalization in Violence Video Classification Performance," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 226, no. 1, 2017, doi: 10.1088/1757-899X/226/1/012082.

[9] L. Xu *et al.*, "Stochastic cross validation," *Chemom. Intell. Lab. Syst.*, vol. 175, no. October 2017, pp. 74–81, 2018, doi: 10.1016/j.chemolab.2018.02.008.

[10] S. Eker, E. Rovenskaya, S. Langan, and M. Obersteiner, "Model validation: A bibliometric analysis of the literature," *Environ. Model. Softw.*, vol. 117, no. December 2018, pp. 43–54, 2019, doi: 10.1016/j.envsoft.2019.03.009.