

PERBANDINGAN ALGORITMA MODIFIED K-NEAREST NEIGHBOR DAN NAIVE BAYES UNTUK KLASIFIKASI DATA REKAM MEDIS BERDASARKAN ICD-10 (STUDI KASUS: RSUD RAJA AHMAD TABIB KOTA TANJUNGPINANG)

Bertho Erizal¹, Martaleli Bettiza², Alena Uperiati³

Berthoerizal21@gmail.com

Program studi Teknik Informatika, Fakultas Teknik, Universitas Maritim Raja Ali Haji

Abstract

Medical record data that increases every day can experience data accumulation so that it affects the performance of medical personnel and it is difficult to know the pattern of disease tendencies suffered in community groups. Therefore, a system was created to classify medical record data based on the code ICD-10. This research aims to obtain the results of the classification of medical record data and obtain the results of a comparison between the algorithms Modified K-Nearest Neighbor and Naive Bayes based on the level of accuracy. The data used is medical record data obtained from the Raja Ahmad Tabib Hospital in Tanjungpinang City in 2019 which has variables of district, age, gender, month of patient visit and code ICD-10. The highest accuracy results from the 3 categories ICD-10 using the Modified K-Nearest Neighbor method of 70% while the Naive Bayes method of 72% obtained 64 classified categories from 89 test data and 365 datasets. So in this research it shows that the Naive Bayes method is better than the Modified K-Nearest Neighbor based on the level of accuracy.

Keywords: Modified K-Nearest Neighbor, Naive Bayes, Medical Records

I. Pendahuluan

RSUD Raja Ahmad Tabib merupakan rumah sakit kelas B non Pendidikan yang telah beroperasi sejak 29 Februari 2012 di Provinsi Kepulauan Riau, Tanjungpinang. Berdasarkan informasi yang diperoleh dari RSUD Raja Ahmad Tabib di bagian Rekam Medis (RM) diketahui bahwa terjadi penumpukan berkas rekam medis yang dapat mempengaruhi kelancaran dan kecepatan kerja tenaga kesehatan. Padahal data rekam medis dapat diolah lebih lanjut agar memudahkan tenaga kesehatan mengetahui pola kecenderungan penyakit yang diderita pada kelompok masyarakat dan memudahkan tenaga kesehatan dalam memberikan tindakan atau terapi kepada penyakit yang diderita pasien. Oleh karena itu penulis ingin melakukan klasifikasi terhadap data rekam medis berdasarkan kode penyakit internasional ICD-10 (*The International Statistical Classification of Diseases and Related Health Problems*).

Metode klasifikasi yang digunakan dalam penelitian ini adalah *Modified K-Nearest Neighbor* dan *Naive Bayes*. Algoritma *Modified K-Nearest Neighbor* merupakan peningkatan kinerja dari metode *K-Nearest Neighbor* dengan menambahkan proses validasi disetiap data latih dan pembobotan pada data uji (Parvin et al., 2008). Sedangkan algoritma *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu untuk memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya (Zuraidah & Nurfalinda, 2019). Berdasarkan permasalahan dan solusi yang ingin diterapkan, maka penelitian ini berjudul "Perbandingan Algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes* untuk Klasifikasi Data Rekam Medis Berdasarkan ICD-10".

II. Metode Penelitian

Pengumpulan data berupa data rekam medis tahun 2019 yang berisi informasi identitas dan penyakit setiap pasien sebanyak 1000 dataset yang diperoleh dari RSUD Raja Ahmad Tabib Kota Tanjungpinang. Data tersebut memiliki beberapa atribut yang digunakan dalam penelitian ini yaitu kecamatan, umur, jenis penyakit, bulan masuk dan kode *ICD-10*. Sebelum diproses menggunakan algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes*, data tersebut ditransformasikan dengan memberikan inisialisasi nilai pada setiap atribut sehingga dapat melalui proses klasifikasi.

2.2 The International Statistical Classification of Diseases and Related Health Problems (ICD-10)

The International Statistical Classification of Diseases and Related Health Problems (ICD) merupakan pengkodean atas penyakit dan tanda-tanda, gejala, temuan yang abnormal, keluhan, keadaan sosial dan eksternal yang menyebabkan cedera atau penyakit seperti yang telah diklasifikasikan oleh *World Health Organization* (WHO, 2016). *ICD-10* memuat klasifikasi diagnostik penyakit dengan standar internasional yang disusun berdasarkan sistem kategori dan dikelompokkan dalam suatu jenis penyakit. Daftar kode *ICD-10* edisi tahun 2016 berdasarkan pembagian bab setiap kode ditunjukkan pada Tabel 1.

Tabel 1. Daftar Kategori *ICD-10*

No.	Kategori <i>ICD-10</i>	Keterangan
1	A00-B99	<i>Certain infectious and parasitic diseases</i>
2	C00-D49	<i>Neoplasms</i>
3	D50-D89	<i>Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism</i>
4	E00-E89	<i>Endocrine, nutritional and metabolic diseases</i>
5	F01-F99	<i>Mental, Behavioral and Neurodevelopmental disorders</i>
6	G00-G99	<i>Diseases of the nervous system</i>
7	H00-H59	<i>Diseases of the eye and adnexa</i>
8	H60-H95	<i>Diseases of the ear and mastoid process</i>
9	I00-I99	<i>Diseases of the circulatory system</i>
10	J00-J99	<i>Diseases of the respiratory system</i>
11	K00-K95	<i>Diseases of the digestive system</i>
12	L00-L99	<i>Diseases of the skin and subcutaneous tissue</i>
13	M00-M99	<i>Diseases of the musculoskeletal system and connective tissue</i>
14	N00-N99	<i>Diseases of the genitourinary system</i>
15	O00-O9A	<i>Pregnancy, childbirth and the puerperium</i>
16	P00-P96	<i>Certain conditions originating in the perinatal period</i>
17	Q00-Q99	<i>Congenital malformations, deformations and chromosomal abnormalities</i>
18	R00-R99	<i>Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified</i>
19	S00-T88	<i>Injury, poisoning and certain other consequences of external causes</i>
20	V00-Y99	<i>External causes of morbidity</i>
21	Z00-Z99	<i>Factors influencing health status and contact with health services</i>

2.2 Normalisasi

Normalisasi dalam kegiatan data mining merupakan proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tertentu. Terdapat beberapa metode yang digunakan dalam proses normalisasi diantaranya *min-max*, *z-score*, *decimal scaling* dan *sigmoidal* (Noviandi, 2018). Adapun proses normalisasi dengan menggunakan metode *min-max* terdapat pada persamaan.

$$Newdata = \frac{(data-min) \times (newmax-newmin)}{(max-min) + newmin}$$

Dimana:

Newdata : Data hasil normalisasi
Min : Nilai minimum dari data per kolom
Max : Nilai maximum dari data per kolom
Newmin : Batas minimum yang diberikan
Newmax : Batas maximum yang diberikan

2.3 Algoritma Modified K-Nearest Neighbor

Hamid Parvin memperkenalkan algoritma pengembangan dari algoritma *K-Nearest Neighbor* yang dapat mengatasi rendahnya tingkat akurasi yang diperoleh. Algoritma *Modified K-Nearest Neighbor* (MKNN) merupakan peningkatan kinerja dari metode *K-Nearest Neighbor* (KNN) dengan menambahkan proses validasi disetiap data latih dan pembobotan pada data uji (Parvin et al., 2008). Proses validasi dalam hal ini digunakan untuk mendapatkan tetangga yang terdekat dan mencari jumlah titik yang memiliki kategori yang sama pada data latih dimana hasil dari proses validasi digunakan sebagai informasi tambahan untuk proses pembobotan. Proses perhitungan untuk klasifikasi data rekam medis menggunakan beberapa persamaan dari algoritma *Modified K-Nearest Neighbor* sebagai berikut:

1. *Euclidean distance* digunakan untuk menghitung jarak antar data sampel.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Dimana:

$d(x, y)$: Jarak *Euclidean* data x dengan data y
 x_i : Nilai fitur ke-i dari data x
 y_i : Nilai fitur ke-i dari data y

2. *Validity* digunakan sebagai informasi lebih mengenai data serta digunakan untuk variabel dalam perhitungan *weight voting*.

$$Validity(x) = \frac{1}{H} \sum_{i=1}^n S(label(x), label(N_i(x)))$$

Dimana:

H : Jumlah titik terdekat
 $label(x)$: Kelas x
 $label(N_i(x))$: Label kelas titik terdekat x

3. *Similarity* digunakan untuk menentukan kesamaan antara suatu kelas dengan kelas lainnya berdasarkan tetangga terdekat.

$$S_{(a,b)} = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

Dimana:

S : *Similarity*

a : Kategori *a* pada data latih

b : Kategori selain *a* pada data latih

4. *Weight voting* digunakan untuk memberikan nilai bobot yang dilakukan pada setiap data uji.

$$W(x) = \text{Validity}(x) \frac{1}{d + \alpha}$$

Dimana:

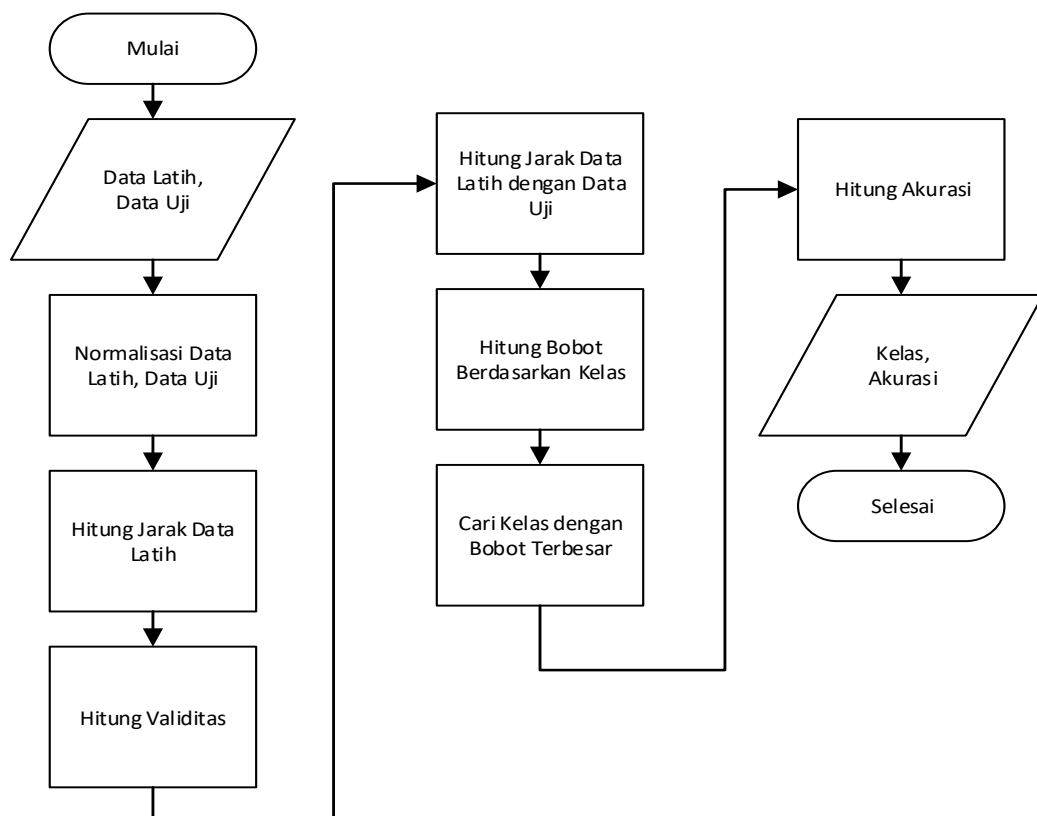
W(x) : Perhitungan *weight voting*

Validity(i) : Nilai *validity*

d : Jarak *Euclidean*

α : *Regular smoothing* akan menggunakan nilai 0.5

Adapun alur sistem untuk proses klasifikasi menggunakan metode *Modified K-Nearest Neighbor* ditunjukkan dalam bentuk *flowchart* yang ditunjukkan pada Gambar 1.



Gambar 1. Flowchart Perhitungan Modified K-Nearest Neighbor

Berdasarkan Gambar 1 langkah-langkah dalam algoritma *Modified K-Nearest Neighbor* sebagai berikut:

- 1) Menentukan Nilai *K* dan memasukkan data latih serta data uji.
- 2) Melakukan normalisasi data latih dan data uji.
- 3) Menghitung jarak *euclidean* setiap data latih dengan data latih lainnya. Hasil perhitungan diurutkan dari yang terkecil dan memilih tetangga terdekat berdasarkan nilai *K*.
- 4) Menghitung nilai *validity* (validitas) pada setiap data latih yang sebelumnya telah dilakukan *similarity* yaitu menentukan kesamaan antara suatu kategori dengan kategori lainnya berdasarkan nilai *K*.
- 5) Menghitung jarak *euclidean* antara data uji dengan data latih.
- 6) Menghitung *weight voting* untuk setiap data uji.
- 7) Menentukan kategori data uji dengan menjumlahkan *weight voting* berdasarkan kategori, kemudian mencari kategori dengan jumlah *weight voting* terbesar.
- 8) Menghitung akurasi dari setiap kesamaan data kategori lama dengan kategori baru.
- 9) Menampilkan hasil klasifikasi kategori dan akurasi.

2.4 Algoritma Naive Bayes

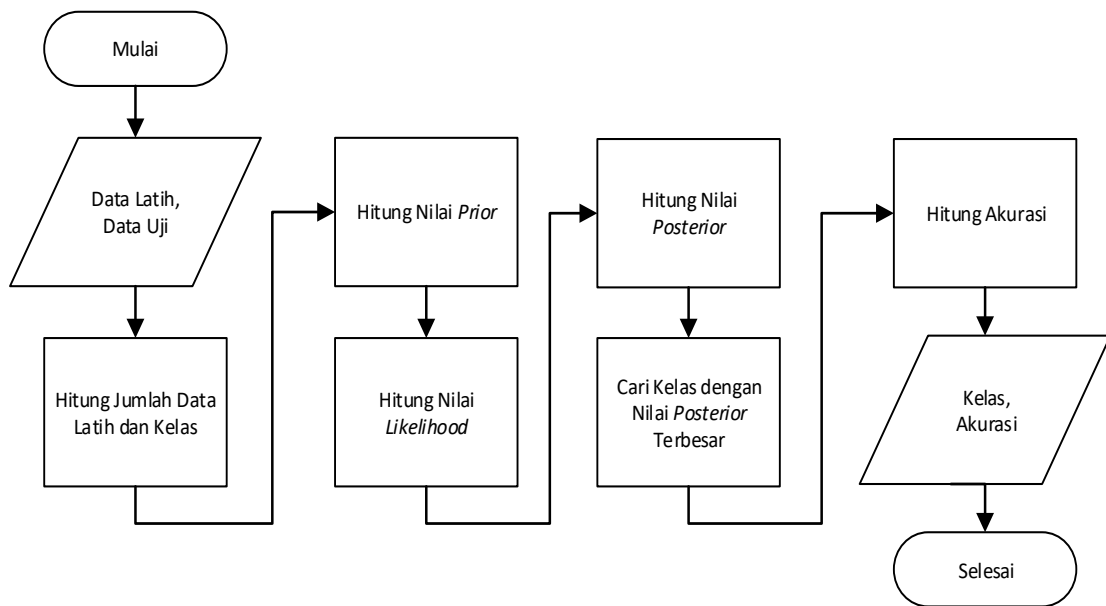
Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu untuk memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya (Zuraidah & Nurfalinda, 2019). Keuntungan dalam menggunakan metode *Naive Bayes* yaitu metode ini dapat menghasilkan akurasi yang maksimal dengan data latih yang sedikit (Devita et al., 2018). Perhitungan algoritma *Naive Bayes* menggunakan persamaan probabilitas sebagai berikut:

$$P(c|x) = \frac{P(x|c).P(C)}{P(x)}$$

Dimana:

- x* : data dengan kategori yang belum diketahui
- c* : kategori dari setiap data
- P(c|x)* : Probabilitas *c* berdasarkan kondisi *x*
- P(c)* : Probabilitas kategori
- P(x|c)* : Probabilitas *x* berdasarkan kondisi *c*
- P(x)* : Probabilitas *x*

Adapun alur sistem untuk proses klasifikasi menggunakan metode *Naive Bayes* ditunjukkan dalam bentuk *flowchart* yang ditunjukkan pada Gambar 2.



Gambar 2. Flowchart Perhitungan Naive Bayes

Berdasarkan Gambar 2 langkah-langkah dalam algoritma *Naive Bayes* sebagai berikut:

- 1) Memasukkan data latih dan data uji.
- 2) Menghitung jumlah data latih dan masing-masing kategori.
- 3) Menghitung probabilitas setiap kasus variabel (*likelihood*) dan masing-masing kategori (*prior*).
- 4) Mengalikan setiap probabilitas variabel dengan probabilitas kategorinya sehingga mendapatkan nilai *posterior*.
- 5) Mencari hasil perkalian dengan nilai (*posterior*) terbesar untuk menentukan kategori dari data uji.
- 6) Menghitung akurasi setiap kesamaan data klasifikasi.
- 7) Menampilkan hasil klasifikasi kategori dan akurasi.

2.5 Akurasi

Pengujian merupakan tahapan untuk mengetahui bagaimana kinerja algoritma yang telah dilakukan dalam proses klasifikasi. Dalam penelitian ini penulis melakukan pengujian untuk mengetahui tingkat akurasi dari algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes*. Perhitungan tingkat akurasi dari algoritma yang digunakan terdapat pada persamaan.

$$\text{Akurasi} = \frac{\text{Jumlah Kecocokan Data}}{\text{Jumlah Data}} \times 100\%$$

Dimana:

Jumlah Kecocokan Data : Jumlah kesamaan antara kategori sebelum proses klasifikasi dengan kategori setelah proses klasifikasi

Jumlah Data : Jumlah Seluruh Data Uji

III. Hasil dan Pembahasan

Penelitian ini melakukan beberapa percobaan untuk mendapatkan hasil akurasi yang bervariasi antara algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes*. Jumlah Data rekam medis yang dijadikan sebagai dataset penelitian sebanyak 1000 data yang memiliki 21 kategori kode *ICD-10*. Uji coba ini dilakukan dengan menggunakan jumlah kategori *ICD-10* yang bervariasi agar dapat melihat kinerja dari masing-masing algoritma berdasarkan banyaknya kategori yang diklasifikasikan. Hasil akurasi pada algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes* ditunjukkan pada Tabel 2.

Tabel 2. Hasil Akurasi

Percobaan	Jumlah ICD-10	Jumlah Data	Akurasi						
			Naive Bayes	Modified K-Nearest Neighbor					
				K=3	K=5	K=7	K=9	K=11	K=15
1	21	(75% : 25%)	31%	31%	24%	20%	20%	18%	20%
2	15	(75% : 25%)	32%	30%	24%	21%	21%	18%	20%
3	9	(75% : 25%)	41%	38%	34%	31%	28%	27%	26%
4	3	(75% : 25%)	72%	69%	70%	64%	60%	62%	60%

Berdasarkan Tabel 2 pengujian tingkat akurasi dilakukan dari percobaan 1 sampai dengan percobaan 4 yang memiliki dataset 75% sebagai data latih dan 25% sebagai data uji. Setiap percobaan memiliki jumlah kategori ICD-10 yang bervariasi diantaranya percobaan 1 menggunakan 21 kategori ICD-10, percobaan 2 menggunakan 15 kategori ICD-10, percobaan 3 menggunakan 9 kategori ICD-10 dan percobaan 4 menggunakan 3 kategori ICD-10. Jumlah kategori yang bervariasi tersebut ditentukan oleh penulis berdasarkan kelipatan 6 yang dimulai dari angka 3 sehingga data terlihat konsisten. Kategori ICD-10 yang diambil pada setiap percobaan adalah kategori dengan jumlah terbanyak dari seluruh dataset. Hasil klasifikasi data uji rekam medis berdasarkan ICD-10 dipercobaan ke-3 menggunakan algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes* ditunjukkan pada Tabel 4.3.

Tabel 3. Hasil Klasifikasi

No.	ICD-10	MKNN	Naive Bayes	Status MKNN	Status Naive Bayes
1.	I00-I99	I00-I99	I00-I99	Sama	Sama
2.	P00-P96	P00-P96	P00-P96	Sama	Sama
3.	I00-I99	I00-I99	I00-I99	Sama	Sama
4.	P00-P96	A00-B99	P00-P96	Tidak Sama	Sama
5.	I00-I99	A00-B99	I00-I99	Tidak Sama	Sama
6.	P00-P96	P00-P96	P00-P96	Sama	Sama
7.	P00-P96	P00-P96	P00-P96	Sama	Sama
8.	I00-I99	I00-I99	I00-I99	Sama	Sama
9.	A00-B99	A00-B99	A00-B99	Sama	Sama
10.	A00-B99	P00-P96	P00-P96	Tidak Sama	Tidak Sama
11.	A00-B99	P00-P96	P00-P96	Tidak Sama	Tidak Sama
12.	I00-I99	P00-P96	A00-B99	Tidak Sama	Tidak Sama
13.	I00-I99	I00-I99	I00-I99	Sama	Sama
14.	I00-I99	I00-I99	I00-I99	Sama	Sama
15.	I00-I99	I00-I99	I00-I99	Sama	Sama
16.	I00-I99	A00-B99	A00-B99	Tidak Sama	Tidak Sama
17.	P00-P96	P00-P96	P00-P96	Sama	Sama
18.	I00-I99	I00-I99	I00-I99	Sama	Sama
19.	I00-I99	P00-P96	A00-B99	Tidak Sama	Tidak Sama
20.	A00-B99	A00-B99	P00-P96	Sama	Tidak Sama
21.	A00-B99	P00-P96	P00-P96	Tidak Sama	Tidak Sama

22	I00-I99	A00-B99	A00-B99	Tidak Sama	Tidak Sama
23	I00-I99	I00-I99	I00-I99	Sama	Sama
24	A00-B99	I00-I99	I00-I99	Tidak Sama	Tidak Sama
25	A00-B99	A00-B99	A00-B99	Sama	Sama
...	...	...	...	...	...
89.	P00-P96	P00-P96	P00-P96	Sama	Sama

IV. Kesimpulan

Berdasarkan hasil pegujian dan analisa yang telah dilaksanakan pada penelitian ini maka dapat diambil kesimpulan bahwa tingkat akurasi pada algoritma *Modified K-Nearest Neighbor* dan *Naive Bayes* akan lebih baik jika menggunakan kategori lebih sedikit. Hasil akurasi tertinggi dari 3 kategori *ICD-10* menggunakan metode *Modified K-Nearest Neighbor* sebesar 70% sedangkan metode *Naive Bayes* sebesar 72% dari 89 data uji dan 365 dataset dipercobaan ke-3. Perbandingan nilai akurasi pada penelitian ini menunjukkan bahwa metode *Naive Bayes* lebih baik dari pada *Modified K-Nearest Neighbor*. Hal tersebut dapat terjadi karena dataset yang digunakan dalam penelitian ini masih kurang banyak untuk klasifikasi menggunakan metode *Modified K-Nearest Neighbor*.

V. Daftar Pustaka

- Afrianti, M., Bettiza, M., & Nurfalinda, 2018, Klasifikasi Kelulusan Tepat Waktu Menggunakan Metode *Naive Bayes* (Studi Kasus : Mahasiswa Informatika UMRAH), *UMRAH*, 3–2, 1–4.
- Budiyono, E. P., Nikentari, N., dan Sallu, S, 2014, Analisa Klasifikasi Kadar Karat Emas Menggunakan Metode *K-Nearest Neighbours* (KNN) Study Kasus : Toko Emas Batam City, *UMRAH*, 1–10.
- Devita, R. N., Herwanto, H. W., dan Wibawa, A. P, 2018, Perbandingan Kinerja Metode *Naive Bayes* dan *K-Nearest Neighbor* untuk Klasifikasi Artikel Berbahasa Indonesia, *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 5(4), 427.
- Fiandra, Y. A., Defit, S., dan Yuhandri, 2017, Penerapan Algoritma *C4.5* untuk Klasifikasi Data Rekam Medis berdasarkan *International Classification Diseases (ICD-10)*, 1(1), 19–25.
- Imanda, A. C., Hidayat, N., dan Furqon, M. T, 2018, Klasifikasi Kelompok Varietas Unggul Padi Menggunakan *Modified K- Nearest Neighbor*, *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(8), 2392–2399.
- Larose, D. T, 2005, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908), New Jersey: John Wiley & Sons, Inc., Hoboken.
- Lestandy, M., Syafa'ah, L., dan Faruq, A, 2020, Klasifikasi Pendonor Darah Potensial Menggunakan Pendekatan Algoritme Pembelajaran Mesin, *Jurnal Teknologi Dan Sistem Komputer*, 8(3), 217–221.
- Noviandi, 2018, *Modul Mata Kuliah Data Mining*, Fakultas Ilmu-Ilmu Kesehatan Universitas Esa Unggul.
- Parvin, H., Alizadeh, H., dan Minaei-bidgoli, B, 2008, MKNN : *Modified K-Nearest Neighbor*, *Proceedings of the World Congress on Engineering and Computer Science*, WCECS, 22–25.
- Ramaulidyah, F. N, 2020, Perbandingan Metode Klasifikasi *Naive Bayes* Dan *K-Nearest Neighbor* Pada Data Status Pembayaran Pajak Pertambahan Nilai Di Kantor Pelayanan Pajak Pratama Smarinda Ulu. *Skripsi Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas*

Mulawarman Samarinda.

- Syahputra, I. K., Bachtiar, F. A., dan Wicaksono, S. A, 2018, Implementasi Data Mining untuk Prediksi Mahasiswa Pengambil Mata Kuliah dengan Algoritme *Naive Bayes*, *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2(11), 5902–5910.
- Triyanti, E., dan Weningsih, I. R, 2018, Rekam Medis dan Informasi Kesehatan (RMIK), *E-Book Manajemen Informasi Kesehatan III*.
- WHO, 2016, *International Statistical Classification of Diseases and Related Health Problems*, 10th Revision ICD-10 : Tabular List, *World Health Organization*, 1, 332–345.
- Zuraidah, dan Nurfalinda, 2019, Diagnosis Penyakit Ikan Kakap Putih Dengan Menggunakan Metode *Naive Bayes*, *UMRAH*, 1–10.