Jurnal Kajian dan Terapan Matematika
Volume 10, Edisi 1, Bulan Maret 2024, 66 - 81
http://journal.student.uny.ac.id/ojs/index.php/jktm

PERBANDINGAN TINGKAT AKURASI METODE *GAUSSIAN NAIVE BAYES* DAN *LEARNING VECTOR QUANTIZATION* (LVQ) UNTUK KLASIFIKASI TANDA-TANDA VITAL PEROKOK

COMPARISON OF ACCURACY OF GAUSSIAN NAIVE BAYES METHOD AND LEARNING VECTOR QUANTIZATION (LVQ) FOR CLASSIFICATION OF SMOKERS' VITAL SIGNS

Ai Aas Siti Asiyah, Prodi Matematika FMIPA UNY
Sri Andayani*, Prodi Matematika FMIPA UNY
*e-mail: andayani@uny.ac.id

Abstrak
Tujuan penelitian ini untuk memperoleh gambaran hasil pemodelan klasifikasi tanda-tanda vital perokok menggunakan metode klasifikasi *Gaussian Naive Bayes* dan metode *Learning Vector Quantization* serta membandingkan kinerja kedua metode berdasarkan nilai akurasinya. Data yang digunakan pada penelitian ini adalah data tanda-tanda vital perokok yang diambil dari situs Kaggle.com dengan banyaknya data yaitu 1000. Data tersebut dilakukan tahap *preprocessing* untuk mendapatkan pemodelan terbaik kemudian data dibagi menjadi data *training* dan data *testing* dengan perbandingan 70:30, 75:25, 80:20, 85:15, 90:10, dan 95:5. Selain itu, digunakan nilai *random state* 0, 1, 30, dan 42. Pada metode *Learning Vector Quantization* digunakan nilai *learning rate* (α), pengurangan *learning rate*, dan maksimum epoh masing-masing sebesar 0,1; 0,5*α; dan 100. Setelah itu dilakukan pengklasifikasian menggunakan kedua metode kemudian dicari nilai akurasinya untuk dibandingkan. Hasil dari penelitian ini adalah pada kedua metode nilai akurasi terbaik diperoleh saat perbandingan data *training* dan data *testing* sebesar 90:10 dan nilai *random state* 42. Pada metode *Gaussian Naïve Bayes* diperoleh nilai akurasi sebesar 0,8 dan metode *Learning Vector Quantization* (LVQ) diperoleh nilai akurasi sebesar 0,809091. Hal tersebut menunjukkan bahwa kedua metode memiliki performa yang hampir sama.
Kata kunci: analisis perbandingan akurasi, tanda vital perokok, *Gaussian Naïve Bayes*, *Learning Vector Quantization*.

Abstract
*This study aimed to obtain an overview of the results of the classification modeling of smokers' vital signs using the Gaussian Naïve Bayes classification method and the Learning Vector Quantization method and compare the performance of the two methods based on their accuracy values. The data used in this study is smoker vital signs data taken from the Kaggle.com site with a lot of data, namely 1000. The data is carried out in the preprocessing stage to get the best modeling then the data is divided into training data and testing data with a ratio of 70:30, 75:25, 80:20, 85:15, 90:10, and 95:5. In addition, random state values of 0, 1, 30, and 42 are used. In the Learning Vector Quantization method, the value of learning rate (α), reduction in the learning rate, and maximum epoch of 0.1 each; 0,5*α; and 100. After that, classification is carried out using both methods and then the accuracy value is sought to be compared. The results of this study are in both methods the best accuracy value obtained when comparing training data and testing data of 90:10 and random state value of 42. In the Gaussian Naïve Bayes method, an accuracy value of 0.8 was obtained and the Learning Vector Quantization (LVQ) method obtained an accuracy value of 0.809091. This shows that both methods have almost the same performance.*
Keywords*: accuracy comparison analysis, smokers' vital signs, Gaussian Naïve Bayes, Learning Vector Quantization.*

PENDAHULUAN

Merokok merupakan salah satu kebiasaan yang banyak dilakukan oleh orang-orang di berbagai negara khususnya negara Indonesia. Berdasarkan laporan *Southeast Asia Tobacco Control Alliance* (SEATCA) pada tahun 2019 dalam *The Tobacco Control Atlas, Asean Region,* negara Indonesia merupakan negara yang memiliki jumlah perokok terbanyak di *Association of Souheast Asian Nations* (ASEAN) yakni 65,19 juta orang (Kementerian Kesehatan Republik Indonesia., 2022). Kebiasaaan merokok dapat menjadi salah satu penyebab munculnya masalah kesehatan (Mahardika et al., 2020). Terdapat lebih dari 25 jenis penyakit yang ditimbulkan di antaranya yaitu kanker mulut, esofagus, faring, laring, paru, pankreas, kandung kemih, penyakit paru obstruktif kronis, dan penyakit pembuluh darah (Nururrahmah., 2015). Penyakit-penyakit yang muncul tersebut dapat berdampak pada kematian seseorang.

Cara yang dilakukan oleh para dokter untuk mengetahui apakah seseorang merupakan perokok atau bukan perokok salah satunya yaitu dengan melalui *bio-signal* atau tanda-tanda vitalnya. Tanda-tanda vital tersebut seperti berat badan, lingkar pinggang, trigliserida, gula darah, hemoglobin, kolesterol, dan lain-lain. Zat berbahaya yang terkandung dalam rokok atau asap rokok dapat memengaruhi tanda-tanda vital tersebut. Misalnya, nikotin akan meningkatkan. level neurotransmiter yang ada di otak dan menekan nafsu makan sehingga mengurangi asupan makanan (Irianti., 2016), orang dengan kebiasaan merokok akan memiliki gigi kuning, kuku kotor, sering batuk-batuk, dan bau mulut (Dinkes Provinsi Banten., 2017).

Dibutuhkan suatu metode untuk mendiagnosis apakah seseorang merupakan perokok atau bukan. Salah satu cara yang digunakan di dunia kesehatan yaitu dengan *medical checkup*. *Medical checkup* merupakan serangkaian pemeriksaan kesehatan secara menyeluruh yang meliputi pemeriksaan laboratorium, pemeriksaan fisik, dan pemeriksaan penunjang lain yang dibutuhkan untuk mengetahui kondisi kesehatan serta mendiagnosis dan mendeteksi dini gejala penyakit yang ditemukan (Halim., 2014). Selain itu, terdapat penelitian yang dilakukan oleh Hilyah, Lestari, dan Mulqie (2020) tentang analisis kadar karbon monoksida (CO) pada perokok dan non-perokok melalui *breath test* menggunakan *smokerlyzer*. Penelitian ini dilakukan karena karbon monoksida merupakan zat kimia beracun yang terkandung dalam asap rokok dan penelitian ini memberikan hasil bahwa terdapat perbedaan kadar karbon monoksida (CO) pada tubuh perokok dan non-perokok. Kadar karbon monoksida (CO) pada tubuh perokok lebih tinggi dibandingkan non-perokok. Penelitian lain yang dilakukan oleh Nadella (2017) mengenai perbandingan pertumbuhan bakteri rongga mulut perokok dan bukan perokok di lingkungan Fakultas Kedokteran Universitas Muhammadiyah Sumatra Utara. Penelitian tersebut memberikan hasil bahwa rata-rata pertumbuhan bakteri rongga mulut perokok lebih banyak dibandingkan dengan bukan perokok. Penelitian yang telah dilakukan tersebut adalah alternatif cara yang digunakan untuk mengetahui apakah seseorang merupakan perokok atau bukan perokok.

Pada penelitian ini akan dilakukan klasifikasi apakah seseorang merupakan perokok atau bukan perokok menggunakan *machine learning*. *Machine learning* merupakan bidang studi yang berfokus pada desain dan analisis suatu algoritma yang nantinya komputer memungkinkan untuk melakukan pembelajaran, algoritma tersebut saat diberikan sejumlah data akan dapat membangun sebuah model atau aturan dari data yang diberikan (Id., 2020). Proses pembelajaran yang dilakukan yaitu suatu usaha untuk memperoleh pengetahuan melalui dua tahap antara lain *training* dan *testing* (Huang et al., 2006). Salah satu teknik dari pengaplikasian *machine learning* yaitu *supervised learning* atau pembelajaran terawasi. *Supervised learning* merupakan metode klasifikasi di mana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang belum dikenali (Roihan et al., 2020). Klasifikasi merupakan teknik yang sering digunakan oleh para peneliti untuk menemukan model klasifikasi dari sebuah data dengan tujuan agar dapat memperkirakan kelas dari suatu objek yang kelasnya belum diketahui (Tan et al., 2004). Metode-metode yang ada dalam

klasifikasi *supervised learning* di antaranya yaitu *K-nearest neighbor*, *super vector machine*, regresi logistik, *naïve bayes*, *decision tree*, *random forest* (Pamungkas et al., 2020), dan n*eural network* (Lakshmi & Sheshasaayee., 2015).

Metode-metode pengklasifikasian di atas memiliki tingkat akurasi yang berbeda-beda. Tingkat akurasi tersebut dapat menggambarkan performa sebuah model dari algoritma klasifikasi di mana performa tersebut sangat penting terhadap hasil perkiraan data yang belum diketahui kelasnya. Penelitian ini akan membandingkan dua metode klasifikasi untuk diukur tingkat akurasinya yaitu metode *Gaussian Naïve Bayes* dan metode *Learning Vector Quantization*.

Pemilihan metode *Gaussian Naïve Bayes* karena metode ini memiliki kinerja yang lebih baik daripada model lain dengan sedikit data *training*, cepat, dan efisiensi ruang sehingga cocok untuk data dalam jumlah yang banyak, dan dapat digunakan pada nilai kontinu. Pemilihan metode *Learning Vector Quantization* karena metode ini mampu melakukan peringkasan terhadap data yang berukuran besar menjadi kecil sehingga waktu pelatihan dan waktu eksekusi lebih cepat dibandingkan metode lain dalam jaringan saraf tiruan.

Adapun penelitian yang telah dilakukan sebelumnya oleh Nugroho, Saptono, dan Sulistyo (2013) mengenai perbandingan metode *naive bayesian classifier* dan *Learning Vector Quantization* (LVQ) dalam kasus klasifikasi penyakit kandungan di mana hasil menunjukkan bahwa metode *naive bayesian classifier* menghasilkan akurasi sebesar 89,6% dan metode *Learning Vector Quantization* (LVQ) sebesar 95,2%. Hal tersebut menandakan bahwa metode LVQ memiliki kinerja yang lebih baik pada penelitian ini. Namun, pada penelitian lain yaitu oleh Simatupang, Wuryandari, dan Suparti (2016) tentang klasifikasi rumah layak huni di kabupaten Brebes dengan menggunakan metode *Learning Vector Quantization* (LVQ) dan *naive bayes* menunjukkan bahwa tingkat akurasi menggunakan metode *naive bayes* lebih unggul yaitu sebesar 95,24% dibanding menggunakan metode *Learning Vector Quantization* (LVQ) yaitu sebesar 71,43%. Penelitian yang dilakukan oleh Rachmad, Oktavianto, dan Rahman (2022) mengenai perbandingan metode *K-nearest neighbor* (KNN) dan *naive bayes* untuk klasifikasi penyakit stroke memberikan tingkat akurasi yang lebih baik untuk metode *naive bayes* yaitu sebesar 74,45%, sedangkan metode KNN memiliki tingkat akurasi sebesar 68,30%.

Berdasarkan uraian di atas, maka rumusan masalah pada penelitian ini yaitu: (1) bagaimana hasil klasifikasi tanda-tanda vital perokok menggunakan metode *Gaussian Naive Bayes* yang memberikan nilai akurasi terbaik?, (2) bagaimana hasil klasifikasi tanda-tanda vital perokok menggunakan metode *Learning Vector Quantization* yang memberikan nilai akurasi terbaik?, dan (3) bagaimana perbandingan nilai akurasi metode *Gaussian Naive Bayes* dan metode *Learning Vector Quantization* (LVQ) untuk klasifikasi tanda-tanda vital perokok?

METODE

Jenis penelitian ini akan mengaplikasikan dua metode yaitu metode *Gaussian Naive Bayes* yang telah dipelajari pada mata kuliah data mining dan metode *Learning Vector Quantization* yang telah dipelajari pada mata kuliah jaringan saraf tiruan. Masing-masing metode memiliki algoritma yang berbeda kemudian akan dibandingkan metode mana yang memiliki performa paling baik untuk klasifikasi tanda-tanda vital perokok.

Berikut ini adalah alur penelitian yang dilakukan untuk membandingkan tingkat akurasi metode *GAUSSIAN NAIVE BAYES* dan metode *Learning Vector Quantization* (LVQ) untuk klasifikasi tanda-tanda vital perokok.

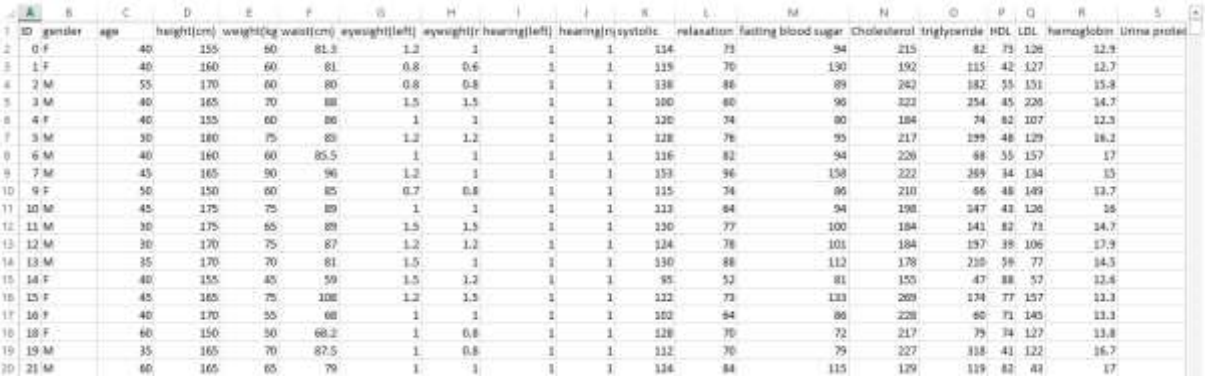
1. Pengumpulan Data
2. Pengolahan Data

   Tahap pengolahan data terdiri dari eksplorasi data dan *preprocessing* data. Tahap *preprocessing* data yang dilakukan di antaranya yaitu *cek missing value* atau data yang

kosong, mengubah kolom data atau variabel tidak numerik menjadi numerik, cek korelasi antarvariabel khususnya variabel independen dengan variabel dependen, memilih variabel independen yang memiliki nilai koefisien korelasi tinggi terhadap variabel dependen yang akan digunakan dalam proses klasifikasi, cek *outlier* data kemudian menghapusnya, cek ketidakseimbangan, normalisasi data, dan memisahkan variabel independen dan variabel dependen.

3. Klasifikasi Menggunakan Metode *Gaussian Naive Bayes* dan Metode *Learning Vector Quantization*
   a. Klasifikasi pada Metode *Gaussian Naive Bayes*
      Pembuatan model klasifikasi menggunakan metode gaussian naive bayes. Tahap-tahap yang dilakukan yaitu membagi data menjadi data training dan data testing, tahap pelatihan, tahap pengujian, kemudian melihat performa model dilihat dari tingkat akurasinya.
   b. Klasifikasi pada Metode *Learning Vector Quantization*
      Pembuatan model klasifikasi menggunakan metode learning vector quantization. Tahap-tahap yang dilakukan yaitu membagi data menjadi data training dan data testing, inisiasi parameter yang digunakan (bobot, learning rate, pengurangan learning rate, dan maksimal epoh), tahap pelatihan, tahap pengujian, kemudian melihat performa model dilihat dari tingkat akurasinya.
4. Analisis perbandingan tingkat akurasi metode *Gaussian Naive Bayes* dan metode *Learning Vector Quantization*.
   Tahap ini dilakukan proses analisis dan membandingan kinerja kedua metode dan metode mana yang memiliki tingkat akurasi terbaik. Selanjutnya, ditarik kesimpulan dari hasil penelitian.

Penelitian ini menggunakan data sinyal tubuh perokok (*body signal of smoking*) atau data tanda-tanda vital perokok. Data tersebut diambil dari platform/situs *kaggle* yaitu https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking. Data *body signal of smoking* merupakan kumpulan data sinyal biologis kesehatan dasar. Data tersebut dikumpulkan untuk mengetahui ada tidaknya kebiasaan merokok melalui *bio-signal* atau tanda-tanda vital pada tubuh. Data ini terdiri dari 55.692 data dan 27 variabel. Variabel-variabel tersebut terdiri dari variabel independen dan variabel dependen (target). Variabel bebas di antaranya ID (indeks), *gender* (jenis kelamin), *age* (umur), *height* (*cm*) (tinggi badan), *weight* (*kg*) (berat badan), *waist* (*cm*) (panjang lingkar pinggang), *eyesight* (*left*) (penglihatan kiri), *eyesight* (*right*) (penglihatan kanan), *hearing* (*left*) (pendengaran kiri), *hearing* (*right*) (pendengaran kanan), *systolic* (tekanan darah sistolik), *relaxation* (tekanan darah diastolik/relaksasi), *fasting blood sugar* (gula darah puasa), *cholesterol* (kolestrol total), *triglyceride* (trigliserida), HDL (jenis kolestrol), LDL (jenis kolestrol), hemoglobin, *urine protein* (protein urine), *serum creatinine* (kreatinin serum), AST (tipe *transaminase oksaloasetat glutamat*), ALT (tipe *transaminase oksaloasetat glutamat*), Gtp: γ-GTP, *oral* (status ujian lisan), *dental caries* (karies gigi), *tartar* (status karang gigi). Variabel target pada data ini adalah status merokok (*smoking*) yang terdiri dari dua kelas yaitu kelas perokok dan kelas bukan perokok. Adapun tampilan data dari data *body signal of smoking* ditampilkan pada Gambar 1.



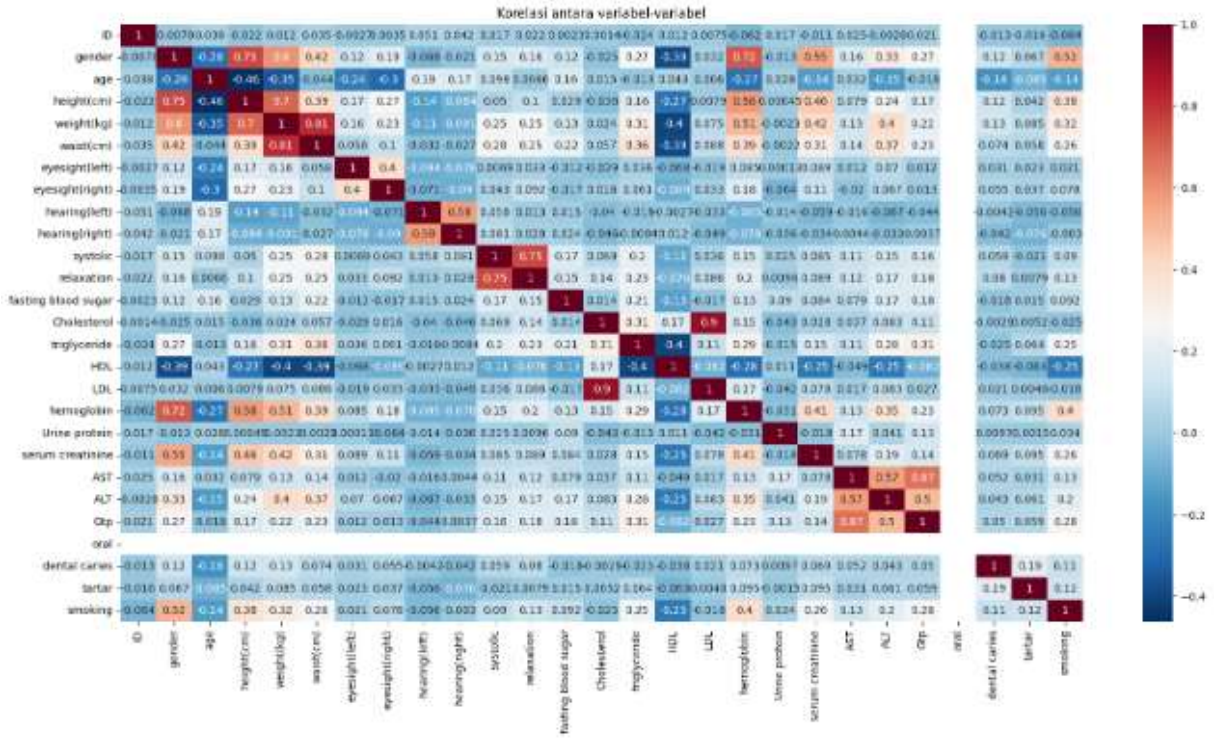
	ID	gender	age	height(cm)	weight(kg	waist(cm)	eyesight(left)	eyesight(r	hearing(left)	hearing(ri	systolic	relaxation	fasting blood sugar	Cholesterol	triglyceride	HDL	LDL	hemoglobin	Urine protei
2	0	F	40	155	60	81.3	1.2	1	1	1	114	73	94	215	82	73	126	12.9	
3	1	F	40	160	60	81	0.8	0.6	1	1	119	70	130	192	115	42	127	12.7	
4	2	M	55	170	60	80	0.8	0.8	1	1	138	86	89	242	182	55	151	15.8	
5	3	M	40	165	70	88	1.5	1.5	1	1	100	60	96	322	254	45	226	14.7	
6	4	F	40	155	60	86	1	1	1	1	120	74	80	184	74	82	107	12.3	
7	5	M	30	180	75	90	1.2	1.2	1	1	128	76	95	217	199	48	129	16.2	
8	6	M	40	160	60	85.5	1	1	1	1	116	82	94	226	68	55	157	17	
9	7	M	45	165	90	96	1.2	1	1	1	153	96	158	222	269	34	134	15	
10	9	F	50	150	60	85	0.7	0.8	1	1	115	74	86	210	66	48	149	13.7	
11	10	M	45	175	75	89	1	1	1	1	133	84	94	198	147	43	126	16	
12	11	M	30	175	65	89	1.5	1.5	1	1	130	77	100	184	141	82	73	14.7	
13	12	M	30	170	75	87	1.2	1.2	1	1	124	78	101	184	197	39	106	17.9	
14	13	M	35	170	70	81	1.5	1	1	1	130	88	112	178	210	59	77	14.5	
15	14	F	40	155	45	59	1.5	1.2	1	1	95	52	81	155	47	88	57	12.6	
16	15	F	45	165	75	108	1.2	1.5	1	1	122	73	133	269	174	77	157	13.3	
17	16	F	40	170	55	68	1	1	1	1	102	64	86	228	60	71	145	13.3	
18	18	F	60	150	50	68.2	1	0.8	1	1	128	70	72	217	79	74	127	13.8	
19	19	M	35	165	70	87.5	1	0.8	1	1	112	70	79	227	318	41	122	16.7	
20	21	M	60	165	65	79	1	1	1	1	124	84	115	129	119	82	43	17	

Gambar 1. Tampilan 20 data teratas pada data *Body Signal of Smoking*

Data *body signal of smoking* di atas berisi data klasifikasi tanda-tanda vital pada tubuh seseorang. Tanda-tanda vital tubuh seseorang tersebut diklasifikasikan pada kelas perokok dan kelas bukan perokok.

HASIL DAN PEMBAHASAN

Sebelum dilakukan tahap pembuatan model klasifikasi menggunakan metode *Gaussian Naïve Bayes* dan *Learning Vector Quantization*, dilakukan tahap *preprocessing* pada data tanda-tanda vital perokok. Tahap tersebut di antaranya yaitu dimulai dengan pengecekan nilai yang kosong dan mengubah kolom data dengan tipe data yang belum numerik menjadi kolom data dengan tipe numerik. Setelah itu dilakukan pemilihan variabel independen yang memiliki nilai koefisien korelasi tinggi dengan variabel dependen. Adapun nilai koefisien antara masing-masing variabel ditampilkan pada Gambar 2 di bawah ini.



Gambar 2. Nilai koefisien korelasi antarvariabel

Gambar 2 menunjukkan nilai koefisien korelasi antarvariabel baik variabel dependen dengan variabel dependen atau variabel dependen dengan variabel independen. Berdasarkan

Gambar 2, diperoleh nilai koefisien korelasi antara variabel independen dengan variabel dependen ditampilkan pada Tabel 1.

Tabel 1. Nilai koefisisen korelasi antara variabel independen dan variabel dependen

No	Variabel dependen	Nilai korelasi
1	ID	-0,064
2	*Gender*	0,52
3	*Age*	-0,14
4	*Height(cm)*	0,38
5	*Weight(kg)*	0,32
6	*Waist(cm)*	0,26
7	*Eyesight(left)*	0,021
8	*Eyesight(right)*	0,078
9	*Hearing(left)*	-0,056
10	*Hearing(right)*	-0,003
11	*Systolic*	0,09
12	*Relaxation*	0,13
13	*Fasting blood sugar*	0,092
14	*cholesterol*	-0,025
15	*triglyceride*	0,25
16	*HDL*	-0,25
17	*LDL*	-0,018
18	*Hemoglobin*	0,4
19	*Urine protein*	0,034
20	*Serum creatinine*	0,26
21	*AST*	0,13
22	*ALT*	0,2
23	*Gtp*	0,28
24	*Oral*	NaN
25	*Dental caries*	0,11
26	*Tartar*	0,12

Korelasi antara variabel dependen dan variabel independen menunjukkan bahwa semakin tinggi nilai korelasi maka semakin tinggi korelasinya. Tabel 1 menunjukkan bahwasannya nilai koefisien korelasi terendah yaitu -0,25 dan nilai koefisien korelasi tertinggi yaitu 0,52. Adapun Tabel 2 menampilkan batasan nilai koefisien korelasi dan banyaknya variabel.

Tabel 2. Batasan nilai koefisien korelasi dan banyaknya variabel

Nilai Koefisien Korelasi	Banyaknya Variabel	Variabel
$\geq 0{,}5$	1	*Gender*
$\geq 0{,}4$	2	*Gender, Hemoglobin*
$\geq 0{,}3$	4	*Gender, Height(cm), Weight(kg), Hemoglobin*
$\geq 0{,}2$	9	*Gender, Height(cm), Weight(kg), Waist(cm), Relaxation, Triglyceride, Hemoglobin, Serum creatinine, ALT, Gtp*
$\geq 0{,}1$	13	*Gender, Height(cm), Weight(kg), Waist(cm), Triglyceride, Hemoglobin, Serum creatinine, AST, ALT, Gtp, Dental caries, Tartar*

Berdasarkan Tabel 2 akan dipilih variabel independen yang memiliki nilai korelasi lebih besar sama dengan dua (korelasi $\geq 0{,}2$). Nilai tersebut dipilih karena banyaknya variabel tidak terlalu sedikit dan tidak terlalu banyak. Dengan demikian, diperoleh variabel independen yang memiliki nilai korelasi lebih besar sama dengan dua yaitu variabel *gender*, *height*(*cm*), *weight*(*kg*), *waist*(*cm*), *triglyceride*, *hemoglobin*, *serum creatinine*, ALT, dan Gtp. Berikut ini Tabel 3 yang merupakan tabel variabel dependen yang terpilih beserta nilai korelasinya.

Tabel 3. Variabel independen yang terpilih beserta nilai koefisien korelasinya

No	Variabel dependen	Nilai Koefisien korelasi
1	*Gender*	0,52
2	*Height(cm)*	0,38
3	*Weight(kg)*	0,32
4	*Waist(cm)*	0,26
5	*triglyceride*	0,25
6	*Hemoglobin*	0,4
7	*Serum creatinine*	0,26
8	*ALT*	0,2
9	*Gtp*	0,28

Tahap *preprocessing* selanjutnya yaitu tahap pengecekan *outlier* atau pencilan, setelah itu akan dilakukan penghapusan data yang merupakan data *outlier*. Jumlah data yang tersisa setelah dilakukan penghapusan *outlier* yaitu 819 data. Gambar 3 merupakan tampilan data tanda-tanda vital perokok terbaru setelah dilakukan penghapusan data *outlier*.

	gender	height(cm)	weight(kg)	waist(cm)	triglyceride	hemoglobin	serum creatinine	ALT	Gtp	smoking
0	0	155	60	81.3	82	12.9	0.7	19	27	0
1	0	160	60	81.0	115	12.7	0.6	19	18	0
2	1	170	60	80.0	182	15.8	1.0	16	22	1
3	1	165	70	88.0	254	14.7	1.0	26	18	0
4	0	155	60	86.0	74	12.5	0.6	14	22	0
...	...	...	...	...	...	...	...	...	...	...
995	1	165	75	89.1	114	14.5	1.0	36	31	1
996	1	175	70	83.0	35	13.2	0.8	20	21	0
997	1	175	80	95.0	89	14.9	1.0	11	55	0
998	1	165	80	88.0	81	15.6	0.8	31	61	1
999	0	155	60	90.0	162	13.5	0.8	28	56	0

819 rows × 10 columns

Gambar 3. Tampilan data tanda-tanda vital perokok terbaru dalam bentuk tabel

Tahap *preprocessing* selanjutnya yaitu pengecekan ketidakseimbangan data. Adapun Tabel 4 menampilkan informasi terkait jumlah kelas pada masing-masing kelas.

Tabel 4. Informasi jumlah kelas pada masing-masing kelas

Kelas	Jumlah kelas
0	548
1	271

Tabel 4 menunjukkan bahwasannya jumlah kelas pada kelas bukan perokok yaitu 548 data dan jumlah kelas pada kelas perokok yaitu 271 data. Masing-masing kelas memiliki jumlah data yang jauh berbeda, hal tersebut menunjukkan bahwa terdapat ketidakseimbangan pada data tanda-tanda vital perokok. Permasalahan ketidakseimbangan data dapat diatasi dengan melakukan *over-sampling*. *Over-sampling* dilakukan dengan tujuan untuk meningkatkan sampel kelas minoritas agar jumlahnya sama dengan kelas mayoritas. Kelas perokok memiliki 271 data sedangkan kelas bukan perokok memiliki 548 data. Kelas yang memiliki jumlah data paling sedikit nantinya akan memiliki jumlah data yang sama dengan kelas yang memiliki jumlah data paling banyak. Pada data tanda-tanda vital perokok, kelas yang memiliki jumlah data paling sedikit adalah kelas perokok sehingga data tersebut dipilih untuk diduplikasi. Kelas perokok akan memiliki jumlah data yang sama dengan kelas bukan perokok yaitu 548 data. Setelah dilakukan penghapusan data yang tidak seimbang, selanjutnya dicek informasi terbaru dari data tanda-tanda vital perokok yang ditampilkan pada Gambar 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1096 entries, 0 to 1095
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            1096 non-null   float64
 1   height(cm)        1096 non-null   float64
 2   weight(kg)        1096 non-null   float64
 3   waist(cm)         1096 non-null   float64
 4   triglyceride      1096 non-null   float64
 5   hemoglobin        1096 non-null   float64
 6   serum creatinine  1096 non-null   float64
 7   ALT               1096 non-null   float64
 8   Gtp               1096 non-null   float64
 9   smoking           1096 non-null   float64
dtypes: float64(10)
memory usage: 85.8 KB
```

Gambar 4. Informasi terbaru dari data tanda-tanda vital perokok

Tahap *preprocessing* selanjutnya yaitu mengubah data ke rentang nilai antara 0 sampai dengan 1 atau disebut dengan normalisasi data. Proses normalisasi yang digunakan yaitu dengan metode normalisasi min-max. Gambar 5 menampilkan nilai hasil normalisasi keseluruhan data dalam bentuk tabel.

	gender	height(cm)	weight(kg)	waist(cm)	triglyceride	hemoglobin	serum creatinine	ALT	Gtp	smoking
0	0.000000	0.300000	0.357143	0.493363	0.235521	0.337349	0.300000	0.375000	0.289855	0.0
1	0.000000	0.400000	0.357143	0.486726	0.362934	0.313253	0.200000	0.375000	0.159420	0.0
2	1.000000	0.600000	0.357143	0.464602	0.621622	0.686747	0.600000	0.312500	0.217391	1.0
3	1.000000	0.500000	0.500000	0.641593	0.899614	0.554217	0.600000	0.520833	0.159420	0.0
4	0.000000	0.300000	0.357143	0.597345	0.204633	0.289157	0.200000	0.270833	0.217391	0.0
...	...	...	...	...	...	...	...	...	...	...
1091	1.000000	0.603882	0.428571	0.507132	0.567118	0.486605	0.400000	0.275686	0.197835	1.0
1092	1.000000	0.530900	0.593500	0.672082	0.365039	0.548735	0.600000	0.729167	0.401565	1.0
1093	0.779904	0.477990	0.285714	0.193822	0.174207	0.485271	0.489952	0.241228	0.182304	1.0
1094	1.000000	0.500000	0.398951	0.440257	0.390904	0.482812	0.341468	0.536582	0.220928	1.0
1095	1.000000	0.579416	0.757351	0.854370	0.627811	0.720236	0.600000	0.717822	0.533460	1.0

1096 rows × 10 columns

Gambar 5. Menampilkan nilai hasil normalisasi keseluruhan data dalam bentuk tabel

Tahap selanjutnya yaitu memisahkan variabel dependen dan variabel independen. Variabel independen terdiri dari 9 variabel yaitu variabel *gender*, *height*(cm), *weight*(kg), *waist*(cm), *triglyceride*, *hemoglobin*, *serum creatinine*, ALT, dan Gtp, sedangkan variabel *smoking* merupakan variabel dependen. Variabel independen disimpan sebagai X dan variabel dependen disimpan sebagai y.

Setelah dilakukan tahap *preprocessing*, diperoleh bahwa banyaknya data setelah dilakukan tahap *preprocessing* yaitu 1096 data. Data tersebut kemudian akan dilakukan pemodelan klasifikasi menggunakan metode *Gaussian naïve bayes* dan *learning vector* quantization.

Data tanda-tanda vital perokok akan dibagi menjadi data *training* dan data *testing* dengan perbandingan yang akan menghasilkan nilai akurasi terbaik. Nilai akurasi terbaik akan didapatkan setelah dilakukan percobaan dengan nilai perbandingan antara jumlah data *training* dan data *testing* yang berbeda-beda. Perbandingan antara jumlah data *training* dan data *testing* yang digunakan berdasarkan persentase keseluruhan data yaitu 70:30, 75:25, 80:20, 85:15, 90:10, dan 95:5. Adapun banyaknya data *training* dan data *testing* berdasarkan persentase perbandingan yang telah ditentukan ditampilkan pada Tabel 5 di bawah ini.

Tabel 5. Hasil perhitungan banyak data *training* dan data *testing*

Persentase Data		**Banyak Data**		**Total Data**
Data *training*	**Data *testing***	**Data *training***	**Data *testing***	
70	30	767	329	1096
75	25	822	274	1096
80	20	877	219	1096
85	15	932	164	1096
90	10	986	110	1096
95	5	1041	55	1096

Pada tahap klasifikasi menggunakan program *python* akan digunakan nilai *random state* agar mendapatkan hasil yang konsisten saat melatih model. Nilai *random state* yang akan dipilih yaitu 0, 1, 30, atau 42. Nilai tersebut merupakan nilai yang umumnya digunakan untuk melatih model yang dibuat (Purwono et al, 2021; Fatoni et al, 2020; Amin et al, 2021).

Setelah dilakukan tahap pembagian data *training* dan data *testing*, dilakukan tahap pelatihan data *training* dan tahap pengujian data *testing*. Adapun nilai variabel target dari data *testing* berdasarkan prediksi yang ditampilkan pada Gambar 6.

```
[1. 0. 0. 0. 1. 0. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 0. 0. 0.
 0. 1. 1. 1. 0. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 1. 0. 1. 1. 0. 0. 0. 1. 1. 1. 0. 1. 0. 0. 1. 0. 0. 1. 0.
 1. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 0. 1.]
```

Gambar 6. Tampilan nilai variabel target dari data *testing* berdasarkan prediksi

Selanjutnya nilai variabel target asli dari data *testing* tersebut ditampilkan pada Gambar 7.

```
[1. 0. 1. 0. 0. 1. 1. 0. 0. 0. 1. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 1. 0. 0.
 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0. 0. 1. 1. 1. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0. 0. 0. 0. 1. 0. 1. 1. 0. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 1. 0. 0. 1. 0. 1. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1.
 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 0. 0. 0. 1.]
```

Gambar 7. Tampilan nilai variabel target asli dari data *testing*

Setelah dilakukan tahap pengujian, selanjutnya akan mencari nilai akurasi yang dihasilkan pada pemodelan menggunakan metode *Gaussian naïve bayes*. Perhitungan nilai akurasi akan dievaluasi menggunakan nilai *confussion matrix*. Nilai *confussion matrix* bergantung pada perbandingan data *training* dan data *testing* serta nilai *random state*. Adapun nilai akurasi yang diperoleh berdasarkan perbandingan data *training* dan data *testing* serta nilai *random state* ditampilkan pada Tabel 6.

Tabel 6. Nilai Akurasi Berdasarkan Perbandingan Data *Training* dan Data *Testing* Serta Nilai *Random State*

Nilai Perbandingan		**Nilai *Random_State***	**Nilai Akurasi**
Data *training*	**Data *testing***		
70	30	0	0,76596
		1	0,74164
		30	0,73860
		42	0,75379
75	25	0	0,76642
		1	0,74452
		30	0,72628
		42	0,74817
80	20	0	0,76818
		1	0,75909
		30	0,73181

		42	0,74090
85	15	0	0,74545
		1	0,73939
		30	0,72121
		42	0,75757
90	10	0	0,73636
		1	0,72727
		30	0,7
		42	0,8
95	5	0	0,69090
		1	0,72727
		30	0,74545
		42	0,74545

Berdasarkan Tabel 6 nilai akurasi terbaik didapatkan ketika nilai perbandingan antara data *training* dan data *testing* yaitu 90:10 dan nilai *random state* yaitu 42. Nilai akurasi terbaik pada metode *gaussian naive bayes* untuk klasifikasi tanda-tanda vital perokok yaitu sebesar 0,8.

Pada klasifikasi menggunakan metode *learning vector quantization*, tahapan yang akan dilakukan yaitu pembagian data *training* dan data *testing*, penetapan bobot (w), maksimum epoh ($MaxEpoh$), *learning rate* (α), dan pengurangan nilai *learning rate*, melakukan tahap pelatihan dengan menentukan jarak minimum dan memperbarui bobot pada epoh pertama, pada epoh berikutnya dilakukan pengurangan nilai *learning rate* terlebih dahulu untuk selanjutnya dilakukan perhitungan jarak minimum dan memperbarui bobot, hal tersebut dilakukan sampai mencapai epoh maksimum. Kemudian didapatkan nilai bobot akhir. Tahap selanjutnya yaitu melakukan tahap pengujian, tahap pengujian dilakukan dengan mencari jarak pada masing-masing bobot, kemudian menentukan jarak terpendek dan nomor dengan jarak terpendek akan menjadi kelasnya.

Pembagian dataset pada metode *learning vector quantization* sama dengan pembagian dataset pada metode *gaussian naive bayes*. Data Sinyal Tubuh Perokok akan dibagi menjadi data *training* dan data *testing* dengan perbandingan yang akan menghasilkan nilai akurasi terbaik. Perbandingan antara jumlah data *training* dan data *testing* yang digunakan yaitu 70;30, 75:25, 80:20, 85:15, 90:10, dan 95:5. Adapun banyaknya data setelah dilakukan tahap *preprocessing* yaitu 1096 data. Perhitungan manual dan perhitungan menggunakan kode program *python* untuk mencari nilai perbandingan data *training* dan data *testing* sama dengan perhitungan pada *gaussian naive bayes*. Selain itu, nilai *random state* yang digunakan juga sama. Adapun banyaknya data *training* dan data *testing* berdasarkan persentase perbandingan yang telah ditentukan ditampilkan pada Tabel 7 di bawah ini.

Tabel 1. Hasil perhitungan banyak data *training* dan data *testing*

Persentase Data		**Banyak Data**		**Total Data**
Data *training*	**Data *testing***	**Data *training***	**Data *testing***	
70	30	767	329	1096

75	25	822	274	1096
80	20	877	219	1096
85	15	932	164	1096
90	10	986	110	1096

Setelah dilakukan tahap pembagian data *training* dan data *testing*, dilakukan tahap pelatihan data *training* dan tahap pengujian data *testing*. Adapun nilai variabel target dari data *testing* berdasarkan prediksi yang ditampilkan pada Gambar 8.

```
[1. 0. 0. 0. 1. 0. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 0. 0. 0.
 0. 1. 1. 1. 0. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 1. 1. 0. 1. 1. 0. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 0. 1.]
```

Gambar 8. Tampilan nilai variabel target dari data *testing* berdasarkan prediksi

Selanjutnya nilai variabel target asli dari data *testing* tersebut ditampilkan pada Gambar 9.

```
[1. 0. 1. 0. 0. 1. 1. 0. 0. 0. 1. 0. 0. 1. 1. 1. 1. 1. 0. 0. 1. 1. 0. 0.
 0. 0. 1. 0. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0. 0. 1. 1. 1. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0. 0. 0. 0. 1. 0. 1. 1. 0. 1. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 1. 0. 0. 1. 0. 1. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1.
 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 0. 0. 0. 1.]
```

Gambar 9. Tampilan nilai variabel target asli dari data *testing*

Sama halnya dengan metode *Gaussian Naïve Bayes*, setelah dilakukan tahap pengujian, selanjutnya akan mencari nilai akurasi yang dihasilkan pada pemodelan menggunakan metode *Learning Vector Quantization*. Perhitungan nilai akurasi akan dievaluasi menggunakan nilai *confussion matrix.* Nilai *confussion matrix* bergantung pada perbandingan data *training* dan data *testing* serta nilai *random state*. Adapun nilai akurasi yang diperoleh berdasarkan perbandingan data *training* dan data *testing* serta nilai *random state* ditampilkan pada Tabel 53.

Tabel 8. Nilai akurasi berdasarkan perbandingan data *training* dan data *testing* serta nilai *random state*

Nilai Perbandingan		**Nilai *Random_State***	**Nilai Akurasi**
Data *training*	**Data *testing***		
70	30	0	0,765957
		1	0,74468
		30	0,7386
		42	0,75076
75	25	0	0,770073
		1	0,748175
		30	0,718978
		42	0,7445
80	20	0	0,77727

		1	0,75
		30	0,73636
		42	0,740901
85	15	0	0,757575
		1	0,739394
		30	0,727273
		42	0,75151
90	10	0	0,73636
		1	0,73636
		30	0,7
		42	0,809091
95	5	0	0,709091
		1	0,72727
		30	0,709091
		42	0,727273

Berdasarkan Tabel 8 nilai akurasi terbaik didapatkan ketika nilai perbandingan antara data *training* dan data *testing* yaitu 90:10 dan nilai *random state* yaitu 42. Nilai akurasi terbaik pada metode *Learning Vector Quantization* untuk klasifikasi tanda-tanda vital perokok yaitu sebesar 0,809091.

Nilai akurasi yang telah didapatkan pada metode *Gaussian Naïve Bayes* dan metode *Learning Vector Quantization* akan dibandingkan satu sama lain. Adapun nilai akurasi yang diperoleh dari kedua metode tersebut ditampilkan pada Tabel 68 di bawah ini.

Tabel 9. Nilai akurasi metode *Gaussian Naïve Bayes* dan metode *Learning Vector Quantization*

No	Metode	Nilai akurasi
1	*Gaussian naive bayes*	0,8
2	*Learning vector quantizatin*	0,809091

Berdasarkan nilai akurasi yang telah diperoleh tersebut, metode *Gaussian Naïve Bayes* memiliki tingkat akurasi sebesar 0,8 dan metode *Learning Vector Quantization* memiliki tingkat akurasi sebesar 0,809091. Metode *Learning Vector Quantization* memiliki tingkat akurasi yang lebih besar dibandingkan tingkat akurasi pada metode *Gaussian Naïve Bayes*. Namun, karena nilai akurasi yang diperoleh hanya memiliki selisih 0,09091 sehingga dapat dikatakan nilai akurasi yang diperoleh tidak berbeda jauh. Berdasarkan hal tersebut, baik metode *Gaussian Naïve Bayes* maupun metode *Learning Vector Quantization* (LVQ) memiliki performa atau kinerja yang hampir sama untuk klasifikasi tanda-tanda vital perokok dilihat dari tingkat akurasinya.

Pada pemodelan klasifikasi *Gaussian Naïve Bayes* tingkat akurasi terbaik diperoleh saat perbandingan data *training* dan data *testing* 90:10 dan nilai *random state* 42. Banyaknya data *testing* atau data uji yaitu 110 data. Hasil prediksi data *testing* yang sama dengan kelas pada

data aktual menghasilkan 56 data masuk kelas perokok dan 32 data masuk kelas bukan perokok. Dari data tersebut dilakukan perhitungan nilai rata-rata setiap variabel dari masing-masing kelasnya. Perhitungan tersebut dilakukan untuk melihat kriteria setiap variabel dari masing-masing kelas. Adapun hasil perhitungan nilai rata-rata tersebut ditampilkan pada Tabel 10 sebagai berikut.

Tabel 10. Nilai rata-rata setiap variabel dari masing-masing kelas pada hasil pemodelan *Gaussian Naïve Bayes*

Gender	Height (cm)	Weight (kg)	Waist (cm)	Triglyceride	Hemoglobin	Serum creatinine	ALT	Gtp	Kelas
0	157,0313	55,3125	74,45625	79,4375	13,25625	0,715625	16	17,09375	Bukan Perokok
1	169,3708	70,9309	85,10047	144,6132	15,41653	0,927793	24,3	34,03418	Perokok

Pada pemodelan klasifikasi *Learning Vector Quantization* (LVQ) tingkat akurasi terbaik yang diperoleh saat perbandingan data *training* dan data *testing* 90:10 dan nilai *random state* 42. Banyaknya data *testing* atau data uji yaitu 110 data. Hasil prediksi data *testing* yang sama dengan kelas pada data aktual menghasilkan 58 data masuk kelas perokok dan 31 data masuk kelas bukan perokok. Sama halnya dengan metode *Gaussian Naïve Bayes*, dari data tersebut dilakukan perhitungan nilai rata-rata setiap variabel untuk melihat kriteria setiap variabel dari masing-masing kelasnya. Hasil perhitungan nilai rata-rata tersebut ditampilkan pada Tabel 11 sebagai berikut.

Tabel 11. Nilai rata-rata setiap variabel dari masing-masing kelas pada hasil pemodelan *Learning Vector Quantization* (LVQ)

Gender	Height (cm)	Weight (kg)	Waist (cm)	Triglyceride	Hemoglobin	Serum creatinine	ALT	Gtp	Kelas
0	156,613	55,32258	74,63226	79,25806	13,31613	0,706452	16,13	17,13	Bukan Perokok
1	169,183	70,33196	84,57468	142,272	15,40931	0,928807	23,94	33,528	Perokok

SIMPULAN

Berdasarkan hasil pembahasan dan analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut.

1. Hasil klasifikasi menggunakan metode *Gaussian Naïve Bayes* yang memberikan nilai akurasi terbaik yaitu sebesar 0,8 diperoleh saat perbandingan data *training* dan data *testing* 90:10 dan nilai *random state* 42. Dari 10% data uji yaitu 110 data dan berdasarkan prediksi benar atau prediksi yang sesuai dengan data aktual, sebanyak 56 data masuk kelas perokok dan sebanyak 32 data masuk kelas bukan perokok. Data yang masuk kelas bukan perokok didominasi oleh perempuan dengan tanda-tanda vital yang dimilikinya yaitu rata-rata tinggi badan 157,0313 cm; berat badan 55,3125 kg; lingkar pinggang 74,45625 cm; trigliserida 79,4375 mg; hemoglobin 13,25625 g; kreatin serum 0,715625 mg; ALT 16 unit/liter; dan Gtp 17,09375. Data yang masuk kelas perokok didominasi oleh laki-laki dengan tanda-tanda vital yang dimilikinya yaitu rata-rata tinggi badan 169,3708 cm; berat badan 70,9309 kg; lingkar pinggang 85,10047 cm; trigliserida 144,6132 mg; hemoglobin 15,41653 g; kreatin serum 0,927793 mg; ALT 24,3 unit/liter; dan Gtp 34,03418.

2. Hasil klasifikasi menggunakan metode *Learning Vector Quantization* (LVQ) yang memberikan nilai akurasi terbaik yaitu sebesar 0,809091 diperoleh saat perbandingan data *training* dan data *testing* 90:10 dan nilai *random state* 42. Dari 10% data uji yaitu 110 data dan berdasarkan prediksi benar atau prediksi yang sesuai dengan data aktual, sebanyak 58 data masuk kelas perokok dan sebanyak 31 data masuk kelas bukan perokok. Data yang masuk kelas bukan perokok didominasi oleh perempuan dengan tanda-tanda vital yang dimilikinya yaitu rata-rata tinggi badan 156,613 cm; berat badan 55,32258 kg; lingkar pinggang 74,63226 cm; trigliserida 79,25806 mg; hemoglobin 13,31613 g; kreatin serum 0,706452 mg; ALT 16,13 unit/liter; dan Gtp 17,13. Data yang masuk kelas perokok didominasi oleh laki-laki dengan tanda-tanda vital yang dimilikinya yaitu rata-rata tinggi badan 169,183 cm; berat badan 70,33196 kg; lingkar pinggang 84,57468 cm; trigliserida 142,272 mg; hemoglobin 15,40931 g; kreatin serum 0,928807 mg; ALT 23,94 unit/liter; dan Gtp 33,528.
3. Berdasarkan nilai akurasi yang telah diperoleh, metode *Gaussian Naïve Bayes* memiliki tingkat akurasi sebesar 0,8 dan metode *Learning Vector Quantization* memiliki tingkat akurasi sebesar 0,809091. Metode *Learning Vector Quantization* memiliki tingkat akurasi yang lebih besar dibandingkan tingkat akurasi pada metode *gaussian naïve bayes*. Namun, karena nilai akurasi yang diperoleh hanya memiliki selisih 0,09091 atau dapat dikatakan nilai akurasi yang diperoleh tidak berbeda jauh. Berdasarkan hal tersebut, baik metode *Gaussian Naïve Bayes* maupun metode *Learning Vector Quantization* (LVQ) memiliki performa atau kinerja yang hampir sama untuk klasifikasi tanda-tanda vital perokok dilihat dari tingkat akurasinya.

SARAN

Pada pemodelan klasifikasi menggunakan kedua metode digunakan *train/test split* (*split validation*) untuk melakukan validasi sederhana dengan membagi dataset secara acak menjadi data latih dan data uji. Namun, teknik tersebut memiliki kelemahan yaitu pengambilan *test error* atau data uji tidak bisa mendistribusikan kelas secara terstruktur. Saran untuk penelitian selanjutnya menggunakan teknik *cross-validation*. Teknik tersebut akan melakukan validasi berulang dengan membagi dataset menjadi banyak subset atau himpunan data latih dan data untuk validasi. Validasi berulang akan menghasilkan model yang lebih optimal untuk mencari performa terbaik. Teknik tersebut lebih efisien dibandingkan dengan teknik *train*/*test split* (*split validation*).

DAFTAR PUSTAKA

Amin, Z. A., Cholil, W., Herdiansyah, M. I., & Negara, E. S. (2021). Analisa Rekam Medis Elektronik untuk Menentukan Diagnosa Medis Dalam Kategori Bab ICD 10 Menggunakan Machine Learning. *POSITIF*: *Jurnal Sistem dan Teknologi Informasi*, *7(2), 127-132.*

Fatoni, A. & Hermawan, F. A. (2020). Optimasi Aplikasi Antrian Pasien Online Menggunakan Algoritma Patient Treatment Time Prediction. *Doctoral dissertation*, Universitas 17 Agustus 1945 Surabaya.

Gagan. (31 Juli 2017). Pengertian Merokok dan Akibatnya. Dinas Kesehatan Provinsi Banten, Diambil pada tanggal 19 Maret 2023, dari https://dinkes.bantenprov.go.id/read/berita/488/PENGERTIAN-MEROKOK-DAN-AKIBATNYA.html

Halim, T. A. (2014). Penerapan Medical Check Up Berkala Sebagai Upaya Pendeteksi Dini Penyakit Akibat Kerja Di Pt. Antam (Persero) Tbk. Gold Mining Business Unit Bogor,

Jawa Barat. *Laporan Tugas Akhir*, Fakultas Kedokteran Universitas Negeri Surakarta. https://digilib.uns.ac.id/dokumen/detail/37173

Hammado, N. (2015). Pengaruh rokok terhadap kesehatan dan pembentukan karakter manusia. *Prosiding*, *1*(1), 77-84.

Hilyah, R. A., Lestari, F., & Mulqie, L. (2020). Analisis Kadar Karbon Monoksida (CO) pada Perokok dan Non-Perokok melalui Breath Test Menggunakan Smokerlyzer. *Prosiding Farmasi*, *6*(2), *371-375.*

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme Learning Machine: Theory and Applications. *Neurocomputing*, *70(1-3), 489-501.*

Id, I. D. (2021). *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python* (Vol. 1). Unri Press.

Irianti, M. T. (2016). Hubungan Antara Status Merokok Terhadap Obesitas Sentral pada Orang Dewasa Sehat Di Desa Kepuharjo Kecamatan Cangkringan Yogyakarta. *Skripsi*, Universitas Sanata Dharma.

Lakshmi, J. V. N., & Sheshasaayee, A. (2015, October). Machine Learning Approaches on Map Reduce for Big Data Analytics. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (pp. 480-484). IEEE.

Mahardhika, D. W., Cindyagita, Z. I., Akbar, M. T., & Sihaloho, E. D. (2020). Pengaruh Status Merokok Terhadap Kemampuan Kognitif Seseorang: Studi Kasus Indonesian Family Life Survey (IFLS). *Jurnal Ekonomi dan Pembangunan, 28(2), 117-129.*

Nadella, R. (2018). Perbandingan Pertumbuhan Bakteri Rongga Mulut Perokok dan Bukan Perokok di Lingkungan Fakultas Kedokteran Universitas Muhammadiyah Sumatera Utara. *Skripsi*, Universitas Muhammadiyah Sumatera Utara Medan. http://repository.umsu.ac.id/handle/123456789/261

Nugroho, P. A., Saptono, R., & Sulistyo, M. E. (2016). Perbandingan Metode Probabilistik Naive Bayesian Classifier dan Jaringan Syaraf Tiruan Learning Vector Quantization dalam Kasus Klasifikasi Penyakit Kandungan. *ITSMART: Jurnal Teknologi dan Informasi, 2(2), 21-34.*

Pamungkas, F. S., Prasetya, B. D., & Kharisudin, I. (2020, March). Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python. In *PRISMA, Prosiding Seminar Nasional Matematika*, *Vol. 3, 692-697.*

Purwono, P., Wirasto, A., & Nisa, K. (2021). Comparison of Machine Learning Algorithms for Classification of Drug Groups. *SISFOTENIKA, 11(2), 196-207.*

P2PTM Kemenkes RI. (22 Maret 2022). Webinar HTTS 2022 Seri 1: Rokok dan Pandemi COVID-19. Kementerian Kesehatan Republik Indonesia, Diambil pada tanggal 17 Maret 2023, dari https://p2ptm.kemkes.go.id/video-p2ptm/webinar-htts-2022-seri-1-rokok-dan-pandemi-covid-19

Rachmad, D. U. M., Oktavianto, H., & Rahman, M. (2022). Perbandingan Metode K-Nearest Neighbor dan Gaussian Naïve Bayes untuk Klasifikasi Penyakit Stroke. *Jurnal Smart Teknologi, 3(4), 405-412.*

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang. *Jurnal Khatulistiwa Informatika, 5(1), 76-82.*

Simatupang, F. J., Wuryandari, T., & Suparti S. (2016). Klasifikasi Rumah Layak Huni di Kabupaten Brebes dengan Menggunakan Metode Learning Vector Quantization dan Naïve Bayes. *Jurnal Gaussian, 5(1), 99-111.*

Tan, Y., Shi, L., Tong, W., Hwang, G. G., & Wang, C. (2004). Multi-Class Tumor Classification by Discriminant Partial Least Squares Using Microarray Gene Expression Data and Assessment of Classification Models. *Computational Biology and Chemistry, 28(3), 235-243.*