Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik20185983
Vol. 5, No. 5, Oktober 2018, hlm. 577-584 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

PERBANDINGAN KLASIFIKASI ANTARA KNN DAN NAIVE BAYES PADA PENENTUAN STATUS GUNUNG BERAPI DENGAN K-FOLD CROSS VALIDATION

Firman Tempola[1], Miftah Muhammad[2], Amal Khairan[3]

[13]Teknik Informatika Universitas Khairun Ternate
[2]Teknik Elektro Universitas Khairun Ternate
Email: [1]firman.tempola@unkhair.ac.id, [2]miftahmuh@unkhair.ac.id, [3]ibntawakkal@gmail.com

(Naskah masuk: 3 Agustus 2018, diterima untuk diterbitkan 29 Oktober 2018)

Abstrak

Penelitian ini membandingkan dua algoritma klasifikasi yaitu K-Nearest Neighbour dan Naive Bayes Classifier pada data-data aktivitas status gunung berapi yang ada di Indonesia. Sedangkan untuk validasi data menggunakan *k-fold cross validation*. Dalam penentuan status gunung berapi, pusat vulkanologi dan mitigasi bencana geologi melakukan dengan dua hal yaitu pengamatan visual dan faktor kegempaan. Pada penelitian ini dalam melakukan klasifikasi aktivitas gunung berapi menggunakan faktor kegempaan. Ada 5 kriteria yang digunakan dalam melakukan klasifikasi yaitu empat faktor kegempaan diantaranya gempa vulkanik dangkal, gempa tektonik jauh, gempa vulkanik dalam, gempa hembusan dan ditambah satu kriteria yaitu status sebelumnya. Ada 3 status yang di yang diklasifikasi yaitu normal, waspada dan siaga. Hasil penelitian yang dibagi kedalam 3 fold disetiap metode klasifikasi diperoleh perbandingan akurasi sistem rata-rata tertinggi pada k-nn 63,68 % dengan standar deviasi 7,47 %. Sedangkan dengan menggunakan naive bayes diperoleh rata-rata akurasi sebesar 79,71 % dengan standar deviasi 3,55 %. Selain itu, penggunaan *naive bayes* jaraknya akurasi lebih dekat dibandingan dengan k-nn.

Kata kunci: *Gunung berapi, knn, naive bayes,k-fold cross validation*

COMPARISON OF CLASSIFICATION BETWEEN KNN AND NAIVE BAYES AT THE DETERMINATION OF THE VOLCANIC STATUS WITH K-FOLD CROSS VALIDATION

Abstract

This research will compare two classification algorithms that are K-Nearest Neighbors and Naive Bayes Classifier on data of volcanic status activity in Indonesia. While for data validation use k-fold cross validation. In determining the status of volcanology center volcanology and geological disaster mitigation to do with two things: visual observation and seismic factors. In this research in doing the classification of volcanic activity using earthquake factor. There are 5 criteria used in the classification of four seismic factors such as shallow volcanic earthquakes, distant tectonic earthquakes, volcanic earthquakes in the earthquake, blast and plus one criterion that is the previous status. There are 3 statuses in which are classified ie normal, alert and alert. The results of the study are divided into 3 fold in each classification method obtained comparison of the highest average system accuracy at 63.68% k-nn with a standard deviation of 7.47%. While using naive bayes obtained an average accuracy of 79.71% with a standard deviation of 3.55%. In addition, the use of naive bayes is closer to the accuracy of k-nn.

Keywords: *Gunung berapi, knn, naive bayes, k-fold cross validation*

1. PENDAHULUAN

Bencana letusan gunung Api Di Indonesia dapat dikatakan hampir setiap tahun terjadi, hal ini dikarenakan banyak terdapat gunung api aktif di Indoensia. Tak hanya itu, posisi geografis Indonesia yang terletak di lempeng Asia dan Australia juga menjadi salah satu faktor sering terjadinya bencana tektonik yang diakibatkan dari letusan gunung berapi. Menurut Kepala Pusat Vulkanologi dan Bencana Geologi, Kementerian Energi dan Sumber Daya Mineral, dari 127 gunung api aktif di Indonesia, hanya 69 yang terpantau. Dan itupun masih jauh dari keadaan ideal, baik dari segi peralatan maupun dari segi Sumber Daya Manusia (Pratomo, 2006). Dengan

begitu, resiko bencana gunung berapi ketika terjadi letusan gunung dampaknya kepada masyarakat masih sangat besar, mengingat masih banyak gunung yang belum terpantau dengan baik. Selain itu, masih banyak warga yang menetap didaerah gunung berapi aktif.

Dalam era digital saat ini, sebagian besar permasalahan yang terjadi dalam kehidupan diselesaikan dengan pemanfaatan teknologi, tak terkecuali dalam hal bencana gunung berapi. Penelitian-penelitian dalam hal penanganan bencana gunung berapi juga sangat bervariasi ada yang membuat sistem peringatan dini, ada yang sebatas pengujian algoritma. Tergantung setiap kepakaran dari peneliti. Misalnya (Reath, et al., 2016) memprediksi erupsi gunung berapi dengan penginderaan jauh untuk menguji kefektifan menggunakan *thermal infrared* (TIR) dengan data-data yang diterapkan dari (Lara-Cuve, et al. 2016) tujuannya untuk menyeleksi fitur bentuk gelombang seismik untuk deteksi kejadian periode panjang di Gunung Api Cotopaxi. (Pratomo, 2006) melakukan klasifikasi gunung api di Indonesia agar dapat dilihat karakteristik dari setiap gunung berapi tanpa penerapan teknologi.

Di Indonesia yang sering mengeluarkan rekomendasi status gunung berapi adalah Pusat Vulkanologi dan mitigasi Bencana Geologi (PVMBG). PVMBG dalam mengeluarkan rekomendasi status aktivitas gunung berapi berdasarkan dengan data-data yang terpantau dari aktivitas setiap gunung. Ada dua cara pemantauan yang dilakukan yaitu berdasarkan pengamatan visual dan faktor kegempaan.

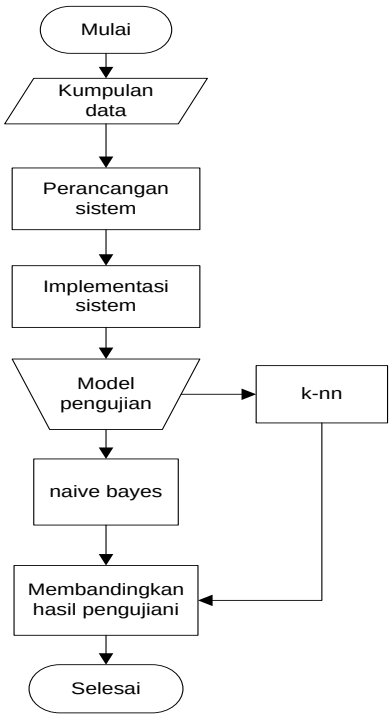
Proses penentuan keputusan status gunung berapi berdasarkan faktor kegempaan pernah dilakukan oleh (Tempola, dkk., 2017) dengan metode yang diterapkan untuk penentuan keputusan status gunung berapi dengan menggunakan metode CBR dengan hasil akurasi sistem mencapai 80, 91% ketika tanpa menggunakan validasi data, namun ketika diterapkan dengan validasi data k-fold cross validation akurasi sistem menurun, dimana yang semula akurasi sistem 80,91 % menurun menjadi 66, 64, %.

Penelitian ini membandingkan dua metode pada *machine learning* yaitu *k-nearest neighbour* (k-nn) dan *naive bayes classifier*. Kedua metode ini memiliki ciri khas masing-masing dalam proses klasifikasi ataupun prediksi. Begitupun dengan metode *machine learning* yang lain, sebagaimana yang dilakukan oleh (Sihananto dan Mahmudy, 2017) melakukan prediksi curah hujan dengan menerapkan jaringan saraf tiruan *backpropagation*.

Implementasi K-NN pernah dilakukan oleh (Kurnianingtyas. Dkk, 2017) untuk melalukan diagnosis penyakit sapi potong, akurasi sistem yang dihasilkan sebesar bahkan mencapai 100%. Berbeda dengan (Puspito, dkk. 2017) membangun sistem pendukung keputusan untuk diagnosa penyakit tanaman jeruk dengan menerapkan metode machine learning yang diterapkan yaitu *naive bayes classifier*, akurasi sistem dalam penelitian ini mencapai 90%. Untuk itu pada penelitian akan dilakukan perbandingan antara K-NN dan Naive Bayes pada datap-data aktivitas gunung berapi. Selain itu, akan dilakukan proses validasi data dengan *k-fold cross validation*. Tujuannya agar hasil prediksi atau klasifikasi tidak hanya akurasi tinggi melainkan juga valid.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah metode klasifikasi yang mana membandingkan dua metode yaitu metode K-*Nearest Neighbor* (K-NN) dan *Naive Bayes Classifier* (NBC). kedua metode ini adalah bagian dari metode supervised learning (Harrington, 2012). K-nn dikenal dengan metode yang paling sederhana sedangkan *naive bayes classifier* adalah salah satu metode yang dapat menampilkan keyakinan dengan label kelas terkait meskipun dengan data training yang sedikit (Aggarwal, 2015). Hasil dari setiap metode kemudian divalidasi dengan *k-fold cross validation*. gambaran dari alur penelitian ini seperti pada Gambar 1.



Gambar 1. Metode Penelitian

2. 1 Naive Bayes Classifier

Naive Bayesian Classifier (NBC) merupakan salah satu metode pada *probabilistic reasoning*, yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. NBC merupakan algoritma klasifikasi yang sangat efektif (mendapatkan hasil yang tepat) dan efisien (proses penalaran dilakukan

memanfaatkan *input* yang ada dengan cara yang relatif cepat). Algoritma NBC bertujuan untuk melakukan klasifikasi data pada kelas tertentu. Unjuk kerja pengklasifikasi diukur dengan nilai *predictive accuracy* (kusumadewi, 2009). Kelebihan lain dari NBC dapat menangani data baik yang bersifat diskrit maupun *continue*. Dalam proses mencari kelas terbaik ketika data berbentuk diskrit dan apabila diberikan k atribut yang saling bebas (*independence*), nilai probabilitas dapat diberikan seperti pada Persamaan 1.

$$P(x_1, , , , , x_k|C) = P(x_1|C) \times \ldots . \times P(x_k|C) \quad (1)$$

Jika atribut ke-i bersifat diskrit atau kategori, maka $P(x_i|C)$ di estimasi sebagai frekuensi relatif sampel yang memiliki nilai x_i sebagai atribut ke-i dalam kelas C. Namun, jika data yang nilai ke-i berisfat kontinu atau numerik, maka $P(x_i|C)$ dicari dengan menggunakan *densitas gauss* seperti pada Persamaan 2.

$$F(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Keterangan :

$$\sigma^2 = standar\ deviasi$$

$$\mu = mean$$

Berikut ini adalah contoh penerapan naive bayes dengan menggunakan data training berjumlah 23 data. Maka tahapan awal adalah dihitung mean dan standar deviasi setiap fitur numerik. Hasil perhitungan ditunjukkan pada Tabel 1. Selanjutnya dihitung probabilitas fitur kategori, dimana dalam kasus ini hanya terdapat satu fitur kategori yaitu status sebelumnya. Hasil perhitungan ditunjukkan pada Tabel 2. Kemudian tahapan selanjutnya adalah dihitung probabilitas setiap rekomendasi status gunung. Hasil perhitungan probabilitas dari fitru kategori dan rekomendasi status di tunjukkan pada Tabel 2.

Tabel 1. Mean dan standar deviasi setiap fitur

1.	Fitur vulkanik Dangkal		
	Status Gunung	Mean	Standar deviasi
	Normal	4	9,7125
	Waspada	20	29,7993
	Siaga	10,5	12,0208
2.	Fitur Tektonik Jauh		
	Status Gunung	Mean	Standar deviasi
	Normal	64,7143	58,8833
	Waspada	39,4286	47,0527
	Siaga	36,7778	37,7153
3.	Fitur Vulkanik Dalam		
	Status Gunung	Mean	Standar deviasi
	Normal	8,1429	8,5718
	Waspada	11,2857	13,2755
	Siaga	18,3333	18,9539
4.	Fitur Gempa Hembusan		
	Status Gunung	Mean	Standar deviasi
	Normal	5,2857	12,7111
	Waspada	10,7143	15,4450
	Siaga	102,4444	50,0053

Tabel 2. Probabilitas setiap status sebelumnya

Status sebelumnya	Jumlah kategori rekomendasi status		
	Normal	Waspada	Siaga
Normal	3	2	0
Waspada	4	3	3
Siaga	0	2	6
Jumlah	7	7	9
Status sebelumnya	Probabilitas rekomendasi status		
	Normal	Waspada	Siaga
Normal	3/7	2/7	0/9
Waspada	4/7	3/7	3/9
Siaga	0/7	2/7	6/9
Jumlah	1	1	1

Berdasarkan hasil perhitungan dari data *continue* atau data numerik maupun data diskrit. kemudian ada data inputan baru aktivitas gunung api dengan gempa vulkanik dangkal 37 kali, gempa tektonik jauh 15 kali, gempa vulkanik dalam 35 kali, gempa hembusan 45 kali dan status sebelumnya waspada. Maka langkah awal untuk mengklasifikasi gunung tersebut adalah menghitung *densitas gauss* masing-masing fitur. Fitur vulkanik dangkal = 37 maka berdasarkan Persamaan (2).

$$f(vulkanik\ dangkal = 37\ |rekomendasi\ status = normal)$$

$$\frac{1}{\sqrt{2\pi}9{,}7125} e^{\frac{-(37-4)^2}{2\,(9{,}7125)^2}} = 1{,}28 \times 10^{-4}$$

$$f(vulkanik\ dangkal = 37\ |rekomendasi\ status = waspada)$$

$$\frac{1}{\sqrt{2\pi}29{,}7993} e^{\frac{-(37-20)^2}{2\,(29{,}7993)^2}} = 0{,}0114$$

$$f(vulkanik\ dangkal = 37\ |rekomendasi\ status = siaga)$$

$$\frac{1}{\sqrt{2\pi}12{,}0208} e^{\frac{-(37-10{,}5)^2}{2\,(12{,}0208)^2}} = 2{,}921 \times 10^{-3}$$

Untuk tektonik jauh = 15 berdasarkan Persamaan (2).

$$f(tektonik\ jauh = 15\ |\ rekomendasi\ status = normal)$$

$$\frac{1}{\sqrt{2\pi}58{,}8833} e^{\frac{-(15-64{,}7143)^2}{2\,(58{,}8833)^2}} = 4{,}743 \times 10^{-3}$$

$$f(tektonik\ jauh = 15\ |\ rekomendasi\ status = waspada)$$

$$\frac{1}{\sqrt{2\pi}47{,}0527} e^{\frac{-(15-39{,}4286)^2}{2\,(47{,}0527)^2}} = 7{,}41 \times 10^{-3}$$

$$f(tektonik\ jauh = 15\ |rekomendasi\ status = siaga)$$

$$\frac{1}{\sqrt{2\pi}37{,}7153} e^{\frac{-(15-36{,}7778)^2}{2\,(37{,}7153)^2}} = 8{,}952 \times 10^{-3}$$

Untuk vulkanik dalam = 35 berdasarkan Persamaan (2).

$$f(vulkanik\ dalam = 35\ |\ rekomendasi\ status = normal)$$

$$\frac{1}{\sqrt{2\pi}8{,}5718} e^{\frac{-(35-8{,}1429)^2}{2\,(8{,}5718)^2}} = 3{,}44 \times 10^{-4}$$

$$f(vulkanik\ dalam = 35\ |\ rekomendasi\ status = waspada)$$

$$\frac{1}{\sqrt{2\pi}13{,}2755}\, e^{\frac{-(35-11{,}2857)^2}{2\,(13{,}2755)^2}} = 6{,}1 \times 10^{-3}$$

$$f(vulkanik\ dalam = 35 \mid rekomendasi\ status = siaga)$$

$$\frac{1}{\sqrt{2\pi}18{,}9539}\, e^{\frac{-(35-18{,}3333)^2}{2\,(18{,}9539)^2}} = 1{,}43 \times 10^{-2}$$

Untuk gempa hembusan = 45 berdasarkan Persamaan (2).

$$f(gempa\ hembusan = 45 \mid rekomendasi\ status = normal)$$

$$\frac{1}{\sqrt{2\pi}12{,}7111}\, e^{\frac{-(45-5{,}2857)^2}{2\,(12{,}7111)^2}} = 2{,}38 \times 10^{-4}$$

$$f(gempa\ hembusan = 45 \mid rekomendasi\ status = waspada)$$

$$\frac{1}{\sqrt{2\pi}15{,}745}\, e^{\frac{-(45-10{,}7143)^2}{2\,(15{,}745)^2}} = 2{,}37 \times 10^{-3}$$

$$f(gempa\ hembusan = 45 \mid rekomendasi\ status = siaga)$$

$$\frac{1}{\sqrt{2\pi}50{,}0053}\, e^{\frac{-(45-102{,}4444)^2}{2\,(50{,}0053)^2}} = 4{,}12 \times 10^{-3}$$

Langkah kedua hitung likelihood dari setiap kategori status gunung yang direkomendasi.

Likelihood normal = (1,28 x 10^{-4}) x (4,743 x 10^{-3}) x (3,44 x 10^{-4}) x (2,38 x 10^{-4}) x (4/7) x (7/23) = $8{,}6 \times 10^{-15}$

Likelihood waspada = (0,0114) x (7,41 x 10^{-3}) x (6,1 x 10^{-3}) x (2,37 x 10^{-3}) x (3/7) x (7/23) = $1{,}6 \times 10^{-10}$

Likelihood siaga = (2,921 x 10^{-3}) x (8,952 x 10^{-3}) x (1,43 x 10^{-2}) x (4,12 x 10^{-3}) x (3/9) x (9/23) = $2{,}01 \times 10^{-10}$

Langkah ketiga menghitung nilai probabilitas dengan cara menormalisasi likelihood tersebut sehingga jumlah nilai yang diperoleh=1.

Probabilitas normal = $\frac{8{,}6 \times 10^{-15}}{8{,}6 \times 10^{-15} + 1{,}6 \times 10^{-10} + 2{,}01 \times 10^{-10}}$ = 2, 39956 x 10^{-5}

Probabilitas waspada = $\frac{1{,}6 \times 10^{-10}}{8{,}6 \times 10^{-15} + 1{,}6 \times 10^{-10} + 2{,}01 \times 10^{-10}}$ = 0,4222

Probabilitas siaga = $\frac{2{,}01 \times 10^{-10}}{8{,}6 \times 10^{-15} + 1{,}6 \times 10^{-10} + 2{,}01 \times 10^{-10}}$ = 0,5778

Langkah terakhir adalah memilih nilai probabilitas tertinggi dari hasil perhitungan probabilitas setiap status gunung yang direkomendasi, dihasilkan nilai probabilitas tertinggi ada pada status siaga sehingga gunung lawu dapat direkomendasikan sebagai gunung dengan status siaga.

2. 2 Algoritma K-NN

Algoritma K-nearest neighbor (k-nn) merupakan salah satu algoritma paling populer dalam machine learning hal ini karena prosesnya mudah dan sederhana (Harrington, 2012). Selain itu k-nn juga salah satu dari algoritma supervised learning dengan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variabel prediktor. Dalam algoritma k-nn semua data yang dimiliki harus memiliki label, sehingga ketika ada data baru yang diberikan kemudian dibandingkan dengan data yang telah ada dan diambil data yang paling mirip dan melihat label dari data tersebut. Adapun langkah-langkah dari algoritma K-NN adalah:

1. Tentukan parameter K
2. Hitung jarak antara data uji dengan data latih. Jika data berbentuk numerik maka menggunakan *euclidean distance* seperti pada Persamaan 3.

$$D\,(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

   Keterangan :
   X_i = data training
   Y_i = data testing
   D (x_i, y_i) = jarak
   i = variabel data
   n = Dimensi data
3. Jarak tersebut kemudian diurutkan secara descending
4. Memilih jarak terdekat sampai pada parameter k
5. Memilih jumlah kelas terbanyak lalu diklasifikasikan

Berikut ini adalah contoh penerapan K-NN pada klasifikasi status gunung berapi. Misalkan data training ada 5 data sebagaimana ditunjukkan pada Tabel 3.

Tabel 3. Data training simulasi K-NN

No	VD	TJ	VA	GH	SS	RS
1	2	68	1	41	waspada	Waspada
2	0	35	10	0	Waspada	Normal
3	19	54	62	114	Waspada	Siaga
4	2	10	7	13	Siaga	Waspada
5	2	15	1	63	Waspada	Waspada

Diberikan data baru aktivitas gunung api dengan gempa vulkanik dangkal 37 kali, gempa tektonik jauh 15 kali, gempa vulkanik dalam 35 kali, gempa hembusan 45 kali dan status sebelumnya waspada. Maka langkah awal untuk mengklasifikasi gunung tersebut adalah menghitung *densitas gauss* masing-masing fitur. Fitur vulkanik dangkal = 37.

Tahapan awal untuk mengklasifikasi dengan k-nn adalah menentukan nilai K. Untuk simulasi ini ditentukan K=3. Selanjutnya dihitung jarak setiap

data dengan data testing, untuk data numerik maka menggunakan Persamaan 3

$$D\ (x_1, y_1) = \sqrt{(-35)^2 + (53)^2 + (-34)^2 + (4)^2 + (0)^2}$$

$$= \sqrt{1225 + 2809 + 1156 + 16 + 0}$$

$$= \sqrt{5206} = 72{,}15$$

$D\ (x_2, y_1) = 68{,}48$

$D\ (x_3, y_1) = 87{,}76$

$D\ (x_4, y_1) = 55{,}31$

$D\ (x_5, y_1) = 52{,}01$

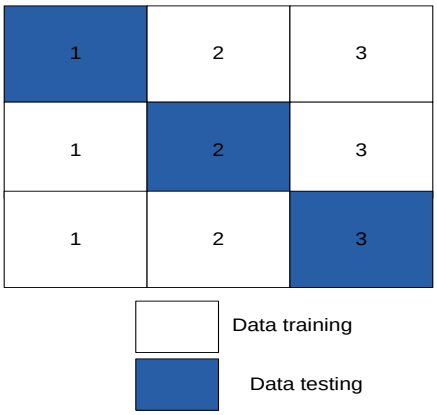
Setelah diperoleh jarak setiap data training dengan data testing selanjutnya diurutkan secara *desecending*, kemudian dipilih 3 data teratas similaritasnya, hal ini karena ditentukan nilai k= 3. Nilai similaritas 3 teratas yaitu ada pada data ke-5 (52,01). Data ke-4 (55,31) dan data ke-2 (68,48). Langkah terakhir adalah memilih kelas terbanyak, berdasarkan 3 data teratas similaritas kelas terbanyak yaitu ada pada status waspada, sehingga data testing di klasifikasikan sebagai waspada.

2. 3 K-fold cross validation

Cross-validasi atau dapat disebut estimasi rotasi adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independen. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Salah satu teknik dari validasi silang adalah *k-fold cross validation*, yang mana memecah data menjadi k bagian set data dengan ukuran yang sama. Penggunaan *k-fold cross validation* untuk menghilangkan bias pada data. Pelatihan dan pengujian dilakukan sebanyak k kali. Pada percobaan pertama, subset S1 diperlakukan sebagai data pengujian dan subset lainnya diperlakukan sebagai data pelatihan, pada percobaan kedua subset S1, S3,…Sk menjadi data pelatihan dan S2 menjadi data pengujian, dan setererusnya (Bramer, 2007).

Pada Gambar 2 merupakan penggunaan 3-fold *cross validation*. Dimana setiap data akan di eksekusi sebanyak 3 kali dan setiap subset data akan mempunyai kesempatan sebagai data testing atau data training. model pengujian seperti berikut dengan diasumsikan nama setiap pembagian data yaitu D1, D2, dan D3:

1. Percobaan pertama data D1 sebagai data testing sedangkan D2 dan D3 sebagai data training
2. Percobaan kedua data D2 sebagai data testing sedangkan data D1 dan D3 sebagai data training.
3. Pada percobaan terakhir atau percobaan ketiga data D3 sebagai data testing sedangkan D1 dan D2 sebagai data training.



Gambar 2. Model 3-fold cross validation

Untuk pengukuran performance klasifikasi yaitu dengan cara membandingkan seluruh data uji yang diklasifikasi benar dengan banyaknya data uji. Persamaan 4 adalah model yang digunakan untuk mengukur kinerja klasifikasi.

$$akurasi = \frac{\sum klasifikasi\ benar}{\sum data\ uji}\ x\ 100\% \qquad (4)$$

Selain itu, simpangan baku (*standar deviation*) juga akan dihitung, simpangan baku adalah ukuran penyebaran data yang menunjukkan jarak rata-rata dari nilai tengah ke suatu titik nilai. Semakin besar simpangan baku yang dihasilkan, maka penyebaran dari nilai tengahnya juga besar, begitu pula sebaliknya. Tujuan dihitung simpangan baku dalam penelitian ini yaitu untuk melihat jarak antara rata-rata akurasi dengan akurasi setiap percobaan. Untuk menghitung simpangan baku menggunakan Persamaan 5 (Brown, 1982).

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \qquad (5)$$

Keterangan :
N= banyaknya percobaan
μ = mean
X = percobaan ke-i
i = indeks setiap percobaan

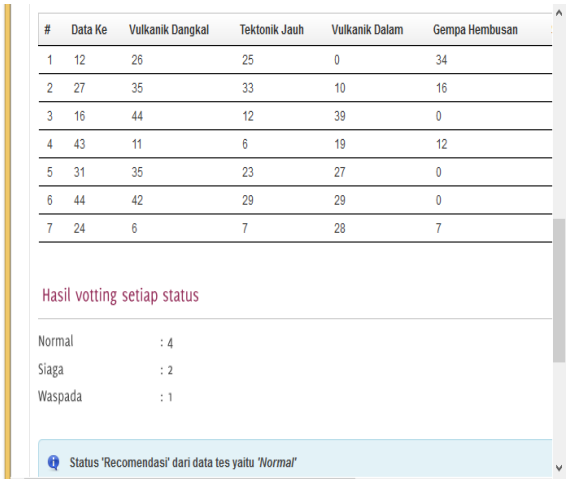
2. 4 Dataset

Dataset yang digunakan pada penelitian ini adalah dataset bersifat publik yang tersedia secara online di website pusat vulkanologi dan mitigasi bencana geologi (PVBMG). Data-data tersebut kemudian di uji pada metode klasifikasi yang telah diterapkan pada sistem. Dan dilanjutkan dengan validasi data.

3. HASIL DAN PEMBAHASAN

Pada penelitian ini dibagi menjadi dua model pengujian yaitu pengujian dengan metode k-*nearest neighbour* (k-nn) dan metode *naive bayes*. Kemudian setelah hasil klasifikasi data dari masing-masing

metode dilakukan proses validasi data. Dan dilanjutkan dengan menghitung standar deviasi dari setiap metode. Sistem ini berbasis web dengan bahasa pemograman yang digunakan adalah bahasa PHP. Pada Gambar 3 ada hasil implementasi sistem dengan menggunakan bahasa pemograman.

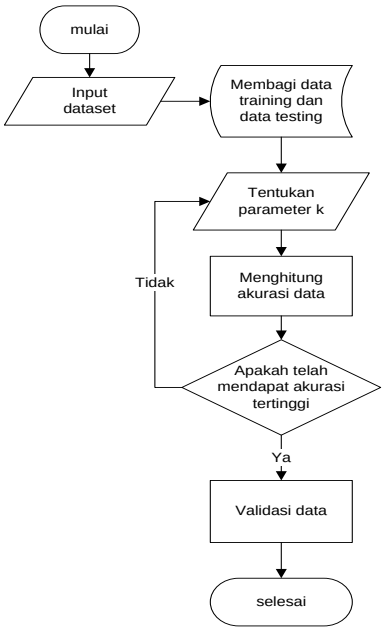


#	Data Ke	Vulkanik Dangkal	Tektonik Jauh	Vulkanik Dalam	Gempa Hembusan
1	12	26	25	0	34
2	27	35	33	10	16
3	16	44	12	39	0
4	43	11	6	19	12
5	31	35	23	27	0
6	44	42	29	29	0
7	24	6	7	28	7

Gambar 3. Hasil Implementasi sistem

3. 1 Pengujian dengan K-NN

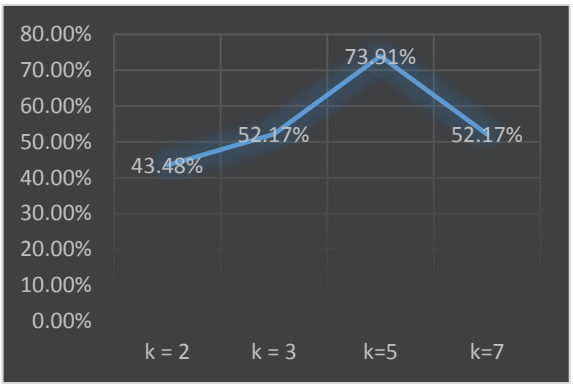
Pengujian dengan KNN dilakukan dengan menginisialisasi nilai k pada beberapa angka hasil akurasi tertinggi kemudian dipilih untuk dilakukan proses validasi data dengan *k-fold cross validation*. Flow chart dengan K-NN seperti terlihat pada Gambar 4.



Gambar 4. Flowchart uji k-nn

Hasil pengujian sistem dengan menggunakan metode klasifikasi KNN dimana terdapat 5 kriteria yaitu gempa vulkanik dangkal, gempa tektonik jauh, gempa vulkanik dalam, gempa hembusan dan status sebelumnya. Sedangkan status yang diklasifikasi ada 3 data yaitu status normal, waspada, dan siaga.

Dalam penelitian ini dataset yang dikumpulkan berjumlah 69 dataset. Sebelum diuji dengan validasi data, pertama diuji dengan menerapkan parameter k yang bervariasi, akurasi parameter k tertinggi kemudian dipilih untuk diuji dengan menggunakan validasi data. Dalam pengujian dengan variasi k, data dibagi kedalam 46 data latih dan 23 data uji. Hasil pengujian diperoleh akurasi sistem dari masing-masing nilai k yaitu k = 2 dengan akurasi 43,48 %, k = 3 dengan akurasi 52, 17 %, k = 5 dengan akurasi 73, 91 % dan k = 7 dihasilkan akurasi sebesar 52,17 %. Grafik dari masing-masing akurasi sistem setiap parameter k seperti pada Gambar 5. Berdasarkan pengujian dengan variasi nilai k akurasi tertinggi yaitu ketika nilai k = 7 sehingga ketika di uji dengan validasi data parameter di inisialisasi sama dengan 7.



Gambar 5. Perbandingan hasil akurasi k-nn

Berdasarkan hasil penentuan parameter K diperoleh akurasi tertinggi ketika nilai k di inisialisasi sama dengan 5 yaitu sebesar 73, 91%. Sehingga parameter k=5 yang dilanjutkan pada tahapan validasi data. Dalam penelitian ini dibagi kedalam 3-fold sesuai dengan karaketeristik dari *k-fold cross validasi* yaitu membagi data sama banyak. Pada pembagian data setiap data memiliki masing-masing 23 data. Model validasi telah diilustrasikan seperti pada Gambar 2. Pada pengujian pertama yaitu D1 sebagai data testing maka data trainingnya D2 dan D3 diperoleh akurasi sistem sebesar 60,87 %, kemudian pengujian kedua dengan data testing D2 dan data training D1 dan D3 dihasilkan akurasi sistem 73,91%. Selanjutnya pada pengujian ketiga dengan data testing D3 dan training D1 dan D2 diperoleh akurasi sistem sebesar 56,27 %. Dari ketiga percobaan pada metode KNN kemudian dihitung rata-rata dari akurasi sistem, maka diperoleh 63,68 %. Selanjutnya dihitung standar deviasi menggunakan Persamaan 5, gunanya untuk melihat jarak akurasi setiap percobaan dengan rata-rata akurasi.

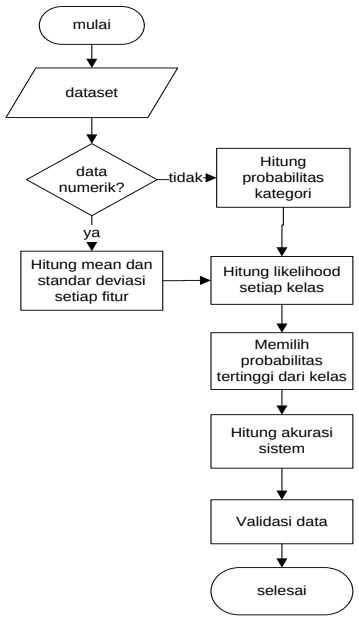
$$\sigma = \sqrt{\frac{(60{,}87 - 63{,}68)^2 + (73{,}91 - 63{,}68)^2 + (56{,}27 - 63{,}68)^2}{3}}$$

$$= 7{,}47$$

3. 2 Pengujian dengan *naive bayes classifier*

Dua metode klasifikasi memiliki ciri khas masing-masing sehingga berbeda didalam langkah-langkah untuk proses klasifikasi data. Pada Gambar 6 merupakan langkah-langkah atau algoritma yang diterapkan dalam melakukan klasifikasi data status

aktivitas gunung berapi dengan menggunakan metode *Naive Bayes classifier*.



Gambar 6. Flowchart uji *naive bayes classifier*

Sesuai dengan langkah-langkah dari klasifikasi data dengan *naive bayes*, maka tahapan pertama yang dilakukan dalam proses *training data* adalah menghitung *mean* dan *standar deviasi* bagi kriteria atau atribut yang datanya berbentuk numerik dan menghitung probabilitas setiap kategori bagi fitur yang data berbentuk kategori. Simulasi perhitungan telah ditunjukkan pada bagian 2.1

Hasil pengujian dengan naive bayes classifier dan pengukuran kinerja klasifikasi berdasarkan akurasi sistemnya dengan menggunakan Persamaan 4, dan dilanjutkan validasi data dengan *3-fold cross validation*, yang mana jumlah data sama banyak yaitu 23 data.

Ilustrasi validasi data seperti pada Gambar 2. Pada percobaan pertama dengan D1 sebagai data testing maka D2 dan D3 sebagai data training diperoleh akurasi sistem sebesar 78,26 %. Untuk percobaan kedua dengan D2 sebagai data testing maka D1 dan D3 sebagai data training dihasilkan akurasi sistem sebesar 82,61 %. Sedangkan pada percobaan ketiga dengan D3 sebagai data testing maka D1 dan D2 sebagai data training dihasilkan akurasi sistem sebesar 78,26 %.

Dari masing-masing akurasi sistem kemudian dihitung rata-rata akurasi sistem dengan cara jumlah seluruh data kemudian dibagi dengan banyaknya pembagian data dihasilkan rata-rata akurasi 79,71 %. Selanjutnya dihitung standar deviasi, tujuannya untuk melihat jarak akurasi setiap eksperimen dengan rata-rata akurasi menggunakan Persamaan 5 dihasilkan sebagai berikut.

$$\sigma = \sqrt{\frac{(78{,}26-79{,}71)^2+(78{,}26-79{,}71)^2+(82{,}61-79{,}71)^2}{3}}$$
$$= 3{,}55$$

4. KESIMPULAN

Berdasarkan pengujian dari dua metode machine learning yang telah diterapkan pada sistem tersebut, diperoleh rata-rata akurasi sistem ketika menggunakan k-nn sebesar 63,68 % dan standar deviasi 7,47. Sedangkan ketika diterapkan naive bayes classifier dihasilkan rata-rata akurasi sistem sebesar 79,71 % dan standar deviasi 3,55%. Dengan demikian ketika diterapkan dengan naive bayes classifier akurasi sistem dalam melakukan klasifikasi lebih baik dibandingkan dengan k-nn. Selain itu, Jarak akurasi setiap eksperimen dengan rata-rata akurasi lebih dekat ketika menggunakan naive bayes dibandingkan dengan KNN hal ini sesuai dengan nilai standar deviasi yang dihasilkan dari masing-masing metode.

5. DAFTAR PUSTAKA

AGGARWAL, C, C., 2015. *Data classification algorithm and application*", A Chapman & Hall/CRC book Data mining and knowledge discovery series., USA Press.

BRAMER, M., 2007. *Principles Of Data Mining*. Springer-Verlag London.

BROWN, G. W., 1982. Standard Deviation Standard Error, *Am J Dis Child, vol. 136*, 937-941.

HARRINGTON, P., 2012. *Machine Learning in Action*. USA: Manning Publication.

KURNIANINGTYAS, D., RAHARDIAN, B, A., MAHARDIKA, D, P., KARTIKA, A., dan ANGRAENI, D., 2017., Sistem Pendukung Keputusan Diagnosis Penyakit Sapi Potong Menggunakan K-Nearest Neighbour (K-NN). *Jurnal Teknologi Informasi dan Ilmu Komputer, vol.4 (2),* 122-126.

KUSUMADEWI, S., 2009. Klasifikasi Status Gizi Menggunakan *Naive Bayesian Classification*. *CommIT, vol.3 (1),* 6-11.

LARA-CUEVA, R. A., BENITEZ, D. S., CARRERA, E. V., RUIZ, M., dan ROJO-ALVAREZ, J. L., 2016. Feature selection of seismicwaveforms for long period event detection at Cotopaxi Volcano. *Journal of Volcanology and Geothermal Research*. *Vol. 316,* 34-49. Diakses dari [https://www.sciencedirect.com.ezproxy.ugm.ac.id/science/article/pii/S0377027316300099] tanggal 3 februari 2018.

PUSPITO, M, A., HIDAYAT, N., dan SUPRAPTO., 2018. Sistem Pendukung Keputusan Diagnosis Penyakit Tanaman Jeruk Menggunakan Metode Naive Bayes Classifier. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *vol. 2,(7),* 2578-2583.

PRATOMO, I., 2006. Klasifikasi Gunung Api di Indonesia, Studi Kasus Dari Beberapa Letusan Gunung Api Dalam Sejarah. *Jurnal*

Geologi Indonesia, *vol. 1,(4),* 2009-227.

REATH, K. A., RAMSEY, M. S., DEHN, J., dan WEBLEY, P, W., 2016. Predicting eruptions from precursory activity using remote sensing data hybridization. *Journal of Volcanology and Geothermal Research*. *Vol. 321*, 18-30. diakses dari [https://www.sciencedirect.com.ezproxy.ugm.ac.id/science/article/pii/S0377027316300695] tanggal 3 Februari 2018

SIHANANTO, A, N., dan MAHMUDY, W. F., 2017. Rainfall forcasting using backpropagation neural network. *Journal of Information Techonology and Computer Sciene*, 2(2), 66-76.

TEMPOLA, F., ARIEF, A., dan MUHAMMAD, M., 2017, Combination Of Case-Based Reasoning And Nearest Neighbor For Recommendation Of Volcano Status. *vol. 2, November 2017*. 348-352 , November 2017 [2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)], 978-1-5386-0658-2 / 17