Indonesian Journal of Data and Science ISSN: 2715-9930
Vol 1, No 2, Juli 2020, pp. 34-38

Klasifikasi Aroma Alkohol Menggunakan Metode KNN

Fadhila Tangguh Admojo[a,1], Ahsanawati[a,2]

[a] *Teknik Informatika, STMIK Palcomtech, Palembang, Indonesia*
[b] *Universitas Gadjah Mada, Yogyakarta, Indonesia*
[1] *Fadhila.tangguh@gmail.com;* [2] *ahsanawati@mail.ugm.ac.id*

INFORMASI ARTIKEL

Diterima : 10 - 05 - 2020
Direvisi : 17 – 06 - 2020
Diterbitkan : 31 – 07 - 2020

Kata Kunci:
alkohol
metode knn
klasifikasi
crossvalidation
analisis performa

ABSTRAK

Alkohol adalah senyawa-senyawa dimana satu atau lebih atom hidrogen dalam sebuah alkana digantikan oleh sebuah gugus -OH. Alkohol memiliki ikatan yang mirip air. Alkohol terdiri dari molekul polar. Dalam senyawa alkohol, oksigen mengemban muatan negatif parsial. Alkohol telah digunakan oleh orang di seluruh dunia, dalam makanan standar, untuk higienis / alasan medis, untuk relaksan dan efek euforia, untuk tujuan rekreasi, untuk inspirasi artistik, sebagai aphrodisiacs, dan untuk alasan lain. Alkohol memiliki beberapa jenis senyawa diantaranya adalah octanol, propanol, Butanol, propanol, dan isobutanol. Oleh karena itu dibutuhkan sensor untuk mendeteksi jenis bahan kimia pada suatu cairan berdasarkan aromanya dengan menerapkan salah satu metode klasifikasi yaitu K-Nearest Neighbor (KNN). Pengujian system ini terdiri dari pengujian pengaruh nilai K dan pengaruh nilai crossvalidation. Hasil dari pengujian pengaruh nilai K menghasilkan akurasi optimum senilai 100% pada nilai K=3 dan 100% pada nilai K=4.

I. Pendahuluan

Di Indonesia, minuman beralkohol yang diimpor diawasi peredarannya oleh negara. Dalam hal ini diamanatkan kepada Direktorat Jenderal Bea dan Cukai Kementerian Keuangan Indonesia (DJBC). Dalam istilah kepabeanan dan cukai; minuman beralkohol disebut sebagai Minuman Mengandung etil alkohol (MMEA). Impor/pemasukan MMEA dari luar negeri dilakukan oleh importir khusus.

Mengingat dampak negatif yang ditimbulkan akibat dari mengonsumsi MMEA tersebut. MMEA ini juga digolongkan dalam 3 golongan, yaitu golongan A (kurang dari 5%), golongan B (5% s.d. 20%), golongan C (lebih dari 20%). Untuk mengendalikan peredaran MMEA pemerintah melalui DJBC mengenakan tarif cukai pada tiap liter MMEA (penggunaan tarif spesifik).

Dalam kimia, alkohol (atau alkanol) adalah istilah yang umum untuk senyawa organik apa pun yang memiliki gugus hidroksil (-OH) yang terikat pada atom karbon, yang ia sendiri terikat pada atom hidrogen dan/atau atom karbon lain. Walaupun selama ini alkohol banyak disarankan untuk dihindari, sebenarnya minuman ini punya sisi baik. Misalnya saja konsumsi alkohol dalam jumlah sedang terbukti bisa melindungi jantung.

Untuk mengetahui senyawa yang terkandung disetiap jenis cairan maka dalam pengujian ini akan dilakukan perhitungan dengan menggunakan K-Nearest Neighbor (KNN). KNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Dengan menggunakan metode KNN maka dapat diketahui akurasi dari setiap senyawa yang terkandung dalam setiap. Crossvalidation adalah salah satu metode yang digunakan agar dapat mensimulasikan semua data agar setiap data dapat berkesempatan menjadi data training dan data tesing. Pada penelitian ini crossvalidation dibagi menjadi kfold = 5.

II. Metode

A. Data Mining

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat.

Data Mining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi. Proses dalam tahap data mining terdiri dari tiga langkah Utama, yaitu data Preparation Pada langkah ini, data dipilih, dibersihkan, dan dilakukan preprocessed mengikuti pedoman dan knowledge dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh. Penggunaan algoritma data mining dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai. Namun semakin besar data yang diolah maka semakin besar pula waktu prosesnya[3][4].

B. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) termasuk kelompok instance-based learning. Algoritma ini juga merupakan salah satu teknik lazy learning. kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. diperlukan suatu sistem klasifikasi sebagai sebuah sistem yang mampu mencari informasi. Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Perhitungan jarak ketetanggan menggunakan algoritma euclidien seperti yang ditunjukkan pada persamaan 1.

$$\text{euc} = \sqrt{((a\_1 - b\_1)\text{^}2 + \cdots + (a\_n - b\_n)\text{^}2)}$$

Dimana a = a1,a2, …, an, dan b = b1, b2, …, bn mewakili n nilai atribut dari dua record. Untuk atribut dengan nilai kategori..

Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Termasuk dalam *supervised learning*, dimana hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN.

Algoritma ini bekerja dengan berdasarkan pada jarak terpendek dari sample uji ke sample latih untuk menentukan KNNnya. Setelah mengumpulkan KNN, kemudian diambil mayoritas dari KNN untuk dijadikan prediksi dari sample uji. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Eucledian. Langkah-langkah untuk menghitung metode K-Nearest Neighbor antara lain:

1. Menentukan parameter K
2. Menghitung jarak antara *data training* dan *data testing*

   Perhitungan jarak yang paling umum dipakai pada perhitungan pada algoritma KNN adalah menggunakan perhitungan jarak Euclidean. Rumusannya adalah sebagai berikut:

$$euc = \sqrt{(\sum_{i=1}^{n} (pi - qi)^2)}$$

   dimana :

   p_i = sample data / *data training*

   q_i = data uji / *data testing*

   i = variabel data

   n = dimensi data
3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak terdekat sampai urutan K
5. Memasangkan kelas yang bersesuaian
6. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi

Akurasi

Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual [5]. rumus akurasi dipaparkan pada persamaan 2.

Presisi

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih[6]. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. rumus presisi ditunjukkan pada persamaan 3.

Recall

Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Rumus Recall diuraikan pada persamaan 4.

F-Measure

Measure adalah harmonic mean antara nilai presisi dan recall, F-measure juga kadang disebut dengan nama F1-Score. Rumus F-Measure dijabarkan pada persamaan 5.

$$AKURASI=((TP+TN))/((TP+TN+FP+FN))$$

$$PRESISI=TP/(TP+FP)$$

$$RECALL=TP/(TP+FN)$$

$$F\text{-}Measure=2\ (Presisi\ x\ Recall)/(Presisi+Recall)$$

Keterangan Variabel:

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative.

C. K-Fold Cross Validation

K-Fold Cross Validation adalah salah satu dari jenis pengujian cross validation yang berfungsi untuk menilai kinerja proses sebuah metode algoritme dengan membagi sampel data secara acak dan mengelompokkan data tersebut sebanyak nilai K k-fold. Kemudian salah satu kelompok k-fold tersebut akan dijadikan sebagai data uji sedangkan sisa kelompok yang lain akan dijadikan sebagai data latih.

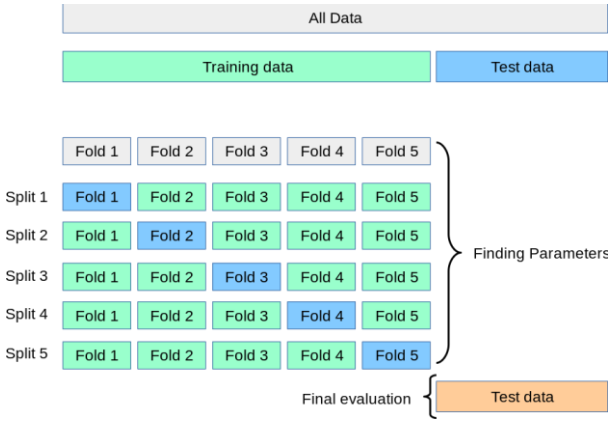
III. Hasil dan Pembahasan

Contoh dataset yang digunakan pada peneitian ini ditunjukkan pada Gambar 1.

# 0.799_0.201	# 0.799_0.201	# 0.700_0.300	# 0.700_0.300	# 0.600_0.400
0 total values	25 total values	25 total values	25 total values	25 total values
-9.4	-7.95	-21.44	-17.46	-34.39
-13.18	-12.01	-26.46	-22.75	-41.48
-18.61	-16.29	-32.84	-28.72	-49.32
-21.88	-19.81	-38.18	-33.77	-56.27
-24.84	-22.36	-42.61	-37.94	-60.47
-48.45	-56.17	-94.77	-87.56	-144.2

Gambar 1. Sampel dataset

Seperti yang telah dipaparkan sebelumnya bahwa tahapan yang dilakukan pada penelitian ini adalah dengan melakukan pembagian data training dan data testing, data yang digunakan sebanyak 500 data, dengan pembagian crossvalidation sebesar kfold=5, sejingga menjadi 80% data training dan 20% data testing disetiap tahapan. Gambar 2 mengilustrasikan crossvalidation pada penelitian ini



Gambar 2. Penerapan crossvalidation

Tahapan selanjutnya adalah menerapkan metode KNN, pemilihan nilai K pada penelitian ini yaitu nilai K=3,4 dan 5. Tabel 1. Menunjukkan hasil percobaan metode knn pada k=3 dan 4. Hasil dari penelitian ini keseluruhan data berhasil diklasifikasikan dengan baik seingga memperoleh rata-rata 95.8% untuk k=3 dan 96.4% untuk k=4. Tabel 1 menunjukkan hasil dari penelitian ini

Tabel 1. Akurasi data untuk nilai K=3 dan K=4

K-N	Akurasi	
K-3	k-fold 1	95%
	k-fold 2	98%
	k-fold 3	93%
	k-fold 4	97%
	k-fold 5	96%
K-4	k-fold 1	98%
	k-fold 2	97%
	k-fold 3	95%
	k-fold 4	95%
	k-fold 5	97%

IV. Kesimpulan

Hasil dari penelitian ini keseluruhan data berhasil diklasifikasikan dengan baik seingga memperoleh rata-rata 95.8% untuk k=3 dan 96.4% untuk k=4. Penerapan crossvalidation pada metode knn ckup baik untuk melihat perubahan nilai k pada setiap uji cobanya.

Daftar Pustaka

[1] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.

[2] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[3] A. A. Karim, H. Azis, and Y. Salim, "Kinerja Metode C4.5 dalam Penyaluran Bantuan Dana Bencana 1," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 84–87, 2018.

[4] L. Nurhayati and H. Azis, "Perancangan Sistem Pendukung Keputusan Untuk Proses Kenaikan Jabatan Struktural Pada Biro Kepegawaian," *Semin. Nas. Teknol. Inf. dan Multimed.*, pp. 6–7, 2016.

[5] H. Azis, R. D. Mallongi, D. Lantara, and Y. Salim, "Comparison of Floyd-Warshall Algorithm and Greedy Algorithm in Determining the Shortest Route," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 294–298, 2018.

[6] N. Fadhillah, Huzain Azis, and D. Lantara, "Validasi Pencarian Kata Kunci Menggunakan Algoritma Levenshtein Distance Berdasarkan Metode Approximate String Matching," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 3–7, 2018.

[7] S. Chugh, K. Arivu Selvan, and R. K. Nadesh, "Prediction of heart disease using apache spark analysing decision trees and gradient boosting algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 4, pp. 0–10, 2017.

[8] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) Untuk Mendeteksi Penyakit Jantung," *Fakt. Exacta*, vol. 7, no. September 2010, pp. 366–371, 2014.

[9] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribb. J. Sci. Technol.*, vol. 1, no. December, pp. 208–217, 2013.

[10] Rosmasari *et al.*, "Usability Study of Student Academic Portal from a User's Perspective," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 108–113, 2018.

[11] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," Int. J. Appl. Pattern Recognit., vol. 3, no. 2, p. 145, 2016, doi: 10.1504/ijapr.2016.079050.

[12] A. Tharwat, "Classification assessment methods," Appl. Comput. Informatics, 2018, doi: 10.1016/j.aci.2018.08.003.

[13] P. A. Flach and M. Kull, “Precision-Recall-Gain curves: PR analysis done right,” Adv. Neural Inf. Process. Syst., vol. 2015-Janua, pp. 838–846, 2015.

[14] J. D. Kelleher, B. Mac Namee, and A. D. Arcy, Fundamentals of Machine Learning For Predictive Data Analytics Algorithms, Worked Examples, and Case Studies. London: The MIT Press, 2015.

[15] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” Proc. - Int. Conf. Pattern Recognit., pp. 3121–3124, 2010, doi: 10.1109/ICPR.2010.764.