DATA MINING

DATA MINING
Pembahasan dan Kasus

Nisa Hanum Harani
Informatics Research Center

Kreatif Industri Nusantara

Kreatif Industri Nusantara

Penulis:
Nisa Hanum Harani

ISBN : 978-602-53897-0-2

Editor:

Penyunting:

Desain sampul dan Tata letak:

Penerbit:
Kreatif Industri Nusantara

Redaksi:
Jl. Ligar Nyawang No. 2
Bandung 40191
Tel. 022 2045-8529
Email : nisahanum@poltekpos.ac.id

Distributor:
Informatics Research Center
Jl. Sariasih No. 54
Bandung 40151
Email : irc@poltekpos.ac.id

Cetakan Pertama, 2019

CONTRIBUTORS

NISA HANUM HARANI, Informatics Research Center., Politeknik Pos Indonesia, Bandung, Indonesia

CONTENTS IN BRIEF

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

Listings

FOREWORD

Sepatah kata dari Kaprodi, Kabag Kemahasiswaan dan Mahasiswa

KATA PENGANTAR

Mempelajari Data Mining Dari Segi Permasalahan Yang Ada Di Lapangan.

N. H. HARANI

Bandung, Jawa Barat
Februari, 2019

ACKNOWLEDGMENTS

Terima kasih atas semua masukan dari para mahasiswa agar bisa membuat buku ini lebih baik dan lebih mudah dimengerti.

Terima kasih ini juga ditujukan khusus untuk team IRC yang telah fokus untuk belajar dan memahami bagaimana buku ini mendampingi proses Intership.

R. M. A.

ACRONYMS

ACGIH	American Conference of Governmental Industrial Hygienists
AEC	Atomic Energy Commission
OSHA	Occupational Health and Safety Commission
SAMA	Scientific Apparatus Makers Association

GLOSSARY

git	Merupakan manajemen sumber kode yang dibuat oleh linus torvald.
bash	Merupakan bahasa sistem operasi berbasiskan *NIX.
linux	Sistem operasi berbasis sumber kode terbuka yang dibuat oleh Linus Torvald

SYMBOLS

A Amplitude
& Propositional logic symbol
a Filter Coefficient

$\mathcal{B}$ Number of Beats

INTRODUCTION

NISA HANUM HARANI, S.KOM., M.T.

Informatics Research Center
Bandung, Jawa Barat, Indonesia

Pada era disruptif saat ini. git merupakan sebuah kebutuhan dalam sebuah organisasi pengembangan perangkat lunak. Buku ini diharapkan bisa menjadi penghantar para programmer, analis, IT Operation dan Project Manajer. Dalam melakukan implementasi git pada diri dan organisasinya..

Rumusnya cuman sebagai contoh aja biar keren[1].

$$ABC\mathcal{D}\mathcal{E}\mathcal{F}\alpha\beta\Gamma\Delta\sum_{def}^{abc} \tag{I.1}$$

BAB 1

PENDAHULUAN

1.1 Pengertian *Data Mining*

1.2 Fungsi *Data Mining*

1.3 Implementasi *Data Mining*

1.4 Rapidminer

1.5 Algoritma *Data Mining*

BAB 2

PREDIKSI

2.1 Perintah Navigasi

Perintah navigasi direktori

BAB 3

KLASIFIKASI

3.1 *Naive Bayes Classifier*

3.1.1 Pengertian *Naive Bayes Classifier*

Naive Bayes Classifier merupakan salah satu metode klasifikasi probabilistik dan statistik yang dikemukakan oleh Thomas Bayes. Algoritma *Naive Bayes* biasa digunakan untuk memprediksi peluang di masa depan dengan berdasarkan pengalaman di masa sebelumnya, biasanya disebut dengan Teorema *Bayes*. *Naive Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan

BAB 4

KLASTERING

4.1 *Clustering K-Means*

Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan *(unsupervised)*. Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data *clustering* dan *non-hierarchical* (non hirarki) data *clustering. K-Means* merupakan salah satu metode data *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok. Metode ini mempartisi data ke dalam *cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu*cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data *clustering* ini adalah untuk meminimalisasikan *objective function* yang diset dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalisasikan variasi antar *cluster*. Data *clustering* menggunakan metode *K-Means* ini secara umum dilakukan dengan algoritma dasar sebagai berikut [2]

4.1.1 *Clustering*

Clustering Analisis Pengelompokan / *Clustering* merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok lain. Potensi *Clustering* adalah dapat digunakan untuk mengetahui struktur dalam data yang dapat dipakai lebih lanjut dalam berbagai aplikasi secara luas seperti klasifikasi, pengolahan gambar, dan pengenalan pola [2].

4.1.2 *K-Means*

K-Means merupakan salah satu metode pengelompokan data non-hierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan kedalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antar kelompok.

Data clustering menggunakan metode K-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut :

1. Tentukan jumlah *Cluster*.
2. Alokasi data ke *Cluster* secara *Random*.
3. Hitung *Centroid* rata-rata dari data yang ada dari masing-masing *Cluster*.
4. alokasi masing-masing data ke *centroid*/rata-rata terdekat.
5. Kembali ke Step 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan.

Dalam tulisan ini beberapa hal terkait dengan metode *K-Means* ini berusaha untuk dijelaskan, termasuk di antaranya beberapa pengembangan yang telah dilakukan terhadap*K-Means*, beberapa permasalahan yang harus diperhitungkan dalam menggunakan metode K*K-Means* dalam pengelompokan data, ulasan mengenai keberadaan *K-Means* di antara metode pengklasifikasian dengan arahan *(supervised)* dan tanpa arahan *(unsupervised)*, ulasan singkat mengenai metode *K-Means* untuk dataset yang mempunyai bentuk khusus dan *mixture modelling*, serta algoritma dari metode-metode pengelompokan yang masih digolongkan sebagai pengembangan metode *K-Means*.

4.1.3 Pengalokasian ulang data ke dalam *Cluster*

Metode Pengalokasian Ulang Data ke Dalam Masing-Masing *Cluster* Secara mendasar, ada dua cara pengalokasian data kembali ke dalam masing-masing *cluster* pada saat proses iterasi *cluster*. Kedua cara tersebut adalah pengalokasian dengan cara tegas *(hard)*, dimana data item secara tegas dinyatakan sebagai anggota *cluster* yang satu dan tidak menjadi anggota *cluster* lainnya, dan dengan cara *cluster*, dimana masing-masing data item diberikan nilai kemungkinan untuk bisa bergabung ke setiap *cluster* yang ada. Kedua cara pengalokasian tersebut diakomodasikan pada metode *K-Means*. Perbedaan di antara kedua metode ini terletak pada asumsi yang dipakai sebagai dasar pengalokasian. *K-Means* Pengalokasian kembali data ke dalam masing-masing *cluster* dalam metode K-Means didasarkan pada perbandingan jarak antara data dengan centroid setiap *cluster* yang ada. Data dialokasikan ulang secara tegas ke*cluster* yang mempunyai *centroid* terdekat dengan data tersebut. Pengalokasian ini dapat dirumuskan sebagai berikut [3]

$$a_{ik} = \begin{cases} 1 & d = \min\{D(x_k, v_i)\} \\ 0 & lainnya \end{cases}$$

dimana:

a_{ik} : Keanggotaan data ke-k ke cluster ke-i

v_i : Nilai *centroid* cluster ke-i

Gambar 4.1 Keanggotaan data *Cluster*

4.1.4 Beberapa permasalahan yang terkait dengan *Clustering K-Means*

Beberapa Permasalahan yang Terkait Dengan *K-Means* Beberapa permasalahan yang sering muncul pada saat menggunakan metode *K-Means* untuk melakukan pengelompokan data adalah:

1. Ditemukannya beberapa model *clustering* yang berbeda
2. Pemilihan jumlah *cluster* yang paling tepat
3. Bentuk masing-masing *cluster*
4. . Masalah overlapping Keenam permasalahan ini adalah beberapa hal yang perlu diperhatikan pada saat menggunakan *K-Means* dalam mengelompokkan data. Permasalahan 1 umumnya disebabkan oleh perbedaan proses inisialisasi anggota masing-masing *cluster*. Proses initialisasi yang sering digunakan adalah proses inisialisasi secara *random*. Dalam suatu studi perbandingan, proses inisialisasi secara *random* mempunyai kecenderungan untuk memberikan hasil yang lebih baik dan *independent*, walaupun dari segi kecepatan untuk lebih lambat. Permasalahan 2 merupakan masalah laten dalam metode *K-Means*. Beberapa pendekatan telah digunakan dalam menentukan jumlah *cluster* yang paling tepat untuk suatu dataset yang dianalisa.

Satu hal yang patut diperhatikan mengenai metode-metode ini adalah pendekatan yang digunakan dalam mengembangkan metode-metode tersebut tidak sama dengan pendekatan yang digunakan oleh *K-Means* dalam mempartisi data items ke masing-masing *cluster*. Permasalahan kegagalan untuk converge, secara teori memungkinkan untuk terjadi dalam kedua metode *K-Means* yang dijelaskan di dalam tulisan ini. Kemungkinan ini akan semakin besar terjadi untuk metode *K-Means*, karena setiap data di dalam dataset dialokasikan secara tegas untuk menjadi bagian dari suatu *cluster* tertentu. Perpindahan suatu data ke suatu cluster tertentu dapat mengubah karakteristik model *clustering* yang dapat menyebabkan data yang telah dipindahkan tersebut lebih sesuai untuk berada di cluster semula sebelum data tersebut dipindahkan. Demikian juga dengan keadaan sebaliknya. Kejadian seperti ini tentu akan mengakibatkan pemodelan tidak akan berhenti dan kegagalan untuk converge akan terjadi. Untuk suatu *cluster*, walaupun ada, kemungkinan permasalahan ini untuk terjadi sangatlah kecil, karena setiap data diperlengkapi dengan membership *function K-Means* untuk menjadi anggota *cluster* yang ditemukan [4]

BAB 5

TEXT MINING

DAFTAR PUSTAKA

1. R. Awangga, "Sampeu: Servicing web map tile service over web map service to increase computation performance," in *IOP Conference Series: Earth and Environmental Science*, vol. 145, no. 1. IOP Publishing, 2018, p. 012057.
2. M. G. Sadewo, A. P. Windarto, and D. Hartama, "Penerapan datamining pada populasi daging ayam ras pedaging di indonesia berdasarkan provinsi menggunakan k-means clustering," *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, vol. 2, no. 1, pp. 60–67, 2017.
3. S. F. Pane, R. M. Awangga, and B. R. Azhari, "Qualitative evaluation of rfid implementation on warehouse management system." *Telkomnika*, vol. 16, no. 3, 2018.
4. M. Y. H. Setyawan, R. M. Awangga, and S. R. Efendi, "Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot," in *2018 International Conference on Applied Engineering (ICAE)*. IEEE, 2018, pp. 1–5.

Index