PEMBOBOTAN KORELASI PADA NAÏVE BAYES CLASSIFIER

Burhan Alfironi Muktamar[1)], Noor Akhmad Setiawan[2)], Teguh Bharata Adji[3)]

[1), 2), 3)] Teknik Elektro dan Teknologi Informasi Universitas Gadjah Mada
Jl. Grafika No.2, Kampus UGM, Yogyakarta, Daerah Istimewa Yogyakarta 55281
Email : burhanalfironimuktamar@gmail.com [1)], noorwewe@ugm.ac.id [2)], adji.tba@gmail.com [3)]

Abstrak

Naïve Bayes Classifier merupakan salah satu algoritma klasifikasi dalam data mining yang memiliki kecepatan proses yang baik dan tingkat akurasi yang cukup tinggi. Dalam proses klasifikasi, Naïve Bayes Classifier mengadopsi teorema Bayesian untuk memetakan suatu data terhadap class dengan memperhitungkan probability dari attribute data tersebut.

Sampai saat ini, algoritma Naïve Bayes Classifier hanya berdasar pada prior probability dan probability attribute. Salah satu hal yang berpotensi untuk meningkatkan akurasi dari Naïve Bayes Classifier adalah nilai korelasi attribute terhadap class. Dengan ikut memperhitungkan korelasi value attribute terhadap class, maka yang menjadi dasar ketepatan klasisifikasi bukan hanya probability melainkan juga seberapa besar hubungan (korelasi) attribute dengan class.

Dalam penelitian ini, penulis ingin meningkatkan tingkat akurasi Naïve Bayes Classifier dengan memperhitungkan nilai korelasi dari masing-masing attribute terhadap class. Dengan metode ini, Naïve Bayes Classifier menambahkan satu parameter tambahan dalam perhitungan probability untuk mencapai posterior probability yaitu korelasi value dari attribute dengan class. Hasil dari penelitian ini adalah sebuah rumusan metode baru dari algoritma Naïve Bayes Classifier yang berbasis pada probability attribute dan korelasi value attribute terhadap class yang dinamakan Correlated-Naïve Bayes Classifier.

Kata kunci: *klasifikasi, Naïve Bayes Classifier, korelasi, data mining.*

1. Pendahuluan

Data mining, sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Sehingga istilah pattern recognition sekarang jarang digunakan karena ia termasuk bagian dari data mining[1].

Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan[2].

Naïve Bayes Classifier merupakan salah satu algoritma yang digunakan dalam klasifikasi data mining. Klasifikasi atau yang disebut *supervised learning* adalah menentukan sebuah *record data* baru ke salah satu dari beberapa kategori (atau kelas) yang telah didefinisikan sebelumnya[3].

Naïve Bayes Classifier adalah algoritma yang mengadopsi teorema *Bayesian*. Bayes merupakan teknik prediksi berbasis *probabilistic* sederhana yang berdasar pada penerapan teorema bayes (atau aturan bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam *Naïve Bayes*, model yang digunakan adalah "model fitur independen". Dalam *Bayes*(terutama *Naïve Bayes*), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama[2]. Teori keputusan Bayes adalah pendekatan *statistic* yang fundamental dalam pengenalan pola (*pattern recognition*). Pendekatan ini didasarkan pada kuantifikasi trade-off antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan ongkos yang ditimbulkan dalam keputusan-keputusan tersebut. Berikut adalah persamaan-persamaan dari teorema Bayes[1]:

- Rumus kata umum dari teorema bayes

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \quad \ldots(1)$$

- Rumus teorema bayes

$$P(Y|X) = \frac{P(X|Y)\ P(Y)}{P(X)} \quad \ldots(2)$$

Kaitan antara *naïve bayes* dengan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi adalah bahwa hipotesis dalam teorema bayes merupakan label kelas yang menjadi target pemetaan dalam klasifikasi, sedangkan bukti merupakan fitur-fitur yang menjadi masukan dalam model klasifikasi. Jika X adalah *vector* masukan yang berisi fitur dan Y adalah label kelas,

Naïve Bayes dituliskan dengan P(Y|X). Notasi tersebut berarti probabilitas label kelas Y didapatkan setelah fitur-fitur X diamati. Notasi ini disebut juga probabilitas akhir (*posterior probability*) untuk Y, sedangkan P(Y) disebut probabilitas awal (*prior probability*) Y. Formulasi *Naive Bayes* untuk klasifikasi adalah[2]:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \ldots(3)$$

Naïve Bayes Classifier merupakan salah satu algoritma klasifikasi yang terkenal dan *power-full*. Pengklasifikasian data dengan *Naïve Bayes Classifier* dapat dengan mudah diinduksi dari *data set*. Namun kekuatan independen *attribute* yang kurang dan hanya menggunakan distribusi probabilitas yang mendasari algoritma tersebut dapat membuat performa klasifikasi menjadi buruk[4].

Naïve Bayes Classifier adalah algoritma klasifikasi yang efisien, mudah dipelajari dan memiliki akurasi tinggi dalam banyak *domain*. Namun, ia memiliki dua kelemahan utama: (i) akurasi klasifikasinya menurun ketika atribut tidak independen, dan (ii) tidak dapat menangani *nonparametric* atribut kontinyu[5].

Kelemahan algoritma *Naïve Bayes Classifier* lainnya adalah dalam proses klasifikasi, algoritma *Naïve Bayes Classifier* hanya berdasar pada *prior probability* dan *probability attribute*. Salah satu hal yang berpotensi untuk meningkatkan akurasi dari *Naïve Bayes Classifier* adalah nilai korelasi *attribute* terhadap *class*. Dengan memperhitungkan korelasi *value attribute* terhadap *class*, maka yang menjadi dasar ketepatan klasisifikasi bukan hanya *probability* melainkan juga seberapa besar hubungan (korelasi) *attribute* dengan *class*.

Analisa korelasi adalah alat statistik yang digunakan untuk mengetahui derajat hubungan *linear* antara variabel yang satu dengan yang lainnya. Analisa korelasi digunakan untuk mencari arah dan kuatnya hubungan antara dua variabel[6].

Koefisien korelasi adalah sebagai pengukur tinggi rendahnya derajad hubungan antara variabel-variabel yang diteliti dan koefisien korelasi merupakan pola suatu ukuran *covariability* (kovariabilitas) antara variabel x dan y tapi tidak menunjukan hubungan fungsional dari kedua variabel tersebut[7].

Tujuan analisa korelasi adalah untuk mencari hubungan variable bebas (X) dengan variable terikat (Y), dengan ketentuan data memiliki syarat-syarat tertentu. Berikut adalah persamaan untuk menentukan korelasi [8]:

$$r = \frac{n \times (\sum XY) - (\sum X) \times (\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad \ldots(4)$$

(r) memiliki ketentuan -1<= r <= 1 dan interpretasi koefisien korelasi nilai (r) dirangkum pada Tabel 1.

Tabel 1.*Tabel Koefisien Korelasi*

Interval Koefisien	Tingkat Hubungan
0 - 0.199	Sangat rendah
0.20 - 0.299	Rendah
0.4 - 0.599	Cukup
0.6 - 0.799	Kuat
0.8 - 1	Sangat kuat

Berdasarkan penjelasan sebelumnya, maka penulis merumuskan masalah sebagai berikut :

1. Bagaimana meningkatkan performa tingkat akurasi algoritma *Naïve Bayes Classifier* dengan korelasi?
2. Apakah *improve* dengan korelasi benar-benar *significant* meningkatkan akurasi algoritma *Naïve Bayes Classifier*?

Tujuan yang diinginkan dari penelitian ini adalah untuk meng-*improve* algorima *Naïve Bayes Classifier* dengan korelasi untuk mendapatkan tingkat akurasi yang lebih tinggi. *Purpose method* dengan korelasi ini kedepannya akan disebut sebagai *Correlated-Naïve Bayes Classifier*. Tujuan yang lain adalah untuk membuktikan bahwa metode baru yang ditawarkan (*Correlated-Naïve Bayes Classifier*) secara *significant* dapat meningkatkan akurasi dari metode *Naïve Bayes Classifier*. Untuk menguji hasil penelitian apakah *significant* atau tidak, maka dalam *paper* ini penulis melakukan uji hipotesis dengan uji z terhadap hasil penelitian. Formula untuk nilai Ratio Uji z (RU_Z)[9] :

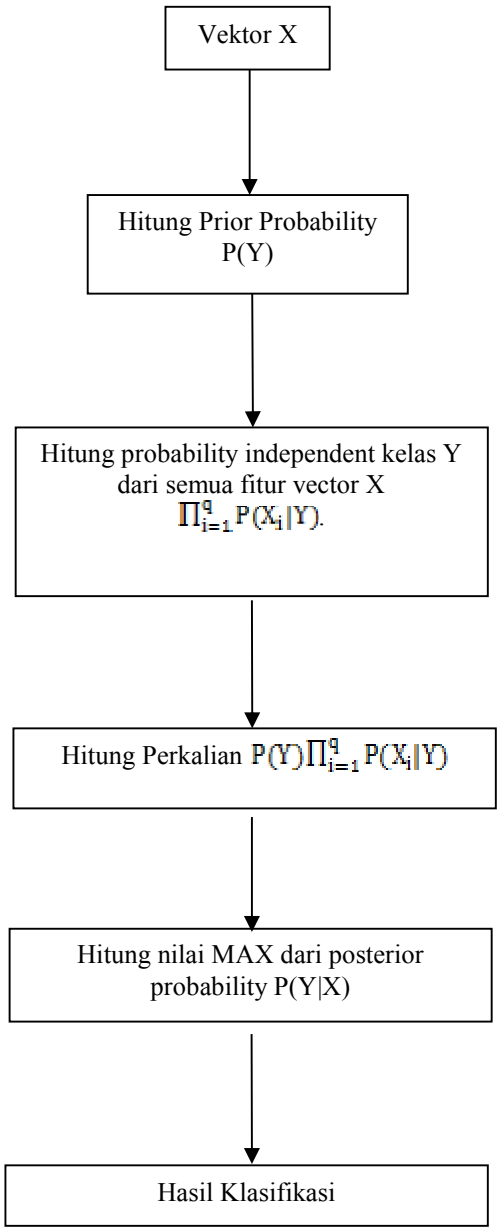
$$RU_Z = \frac{\bar{X} - \mu_{HC}}{\sigma_{\bar{X}}} \quad \ldots(5)$$

2. Pembahasan

Berdasarkan pada persamaan (1) dan persamaan (2), maka untuk mendapat *posterior probability*(*probability* akhir) teorema bayes melakukan perhitungan antara *prior probability* dengan *likelihood* atau *probability* masing-masing *attribute* terhadap *class*. Persamaan *Naïve Bayes Classifier* yang mengadopsi teorema bayes yang ditunjukkan pada persamaan (3) untuk mendapatkan *posterior probability* atau probabilitas data dengan *vector* X pada kelas Y P(Y|X) dapat diuraikan menjadi beberapa langkah sebagi berikut :

1. Menghitung *prior probability* (probability awal).
2. Menghitung probabilitas independen kelas Y dari semua fitur dalam *vector* X.
3. Karena nilai P(X) selalu tetap, maka tinggal menghitung perkalian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$.

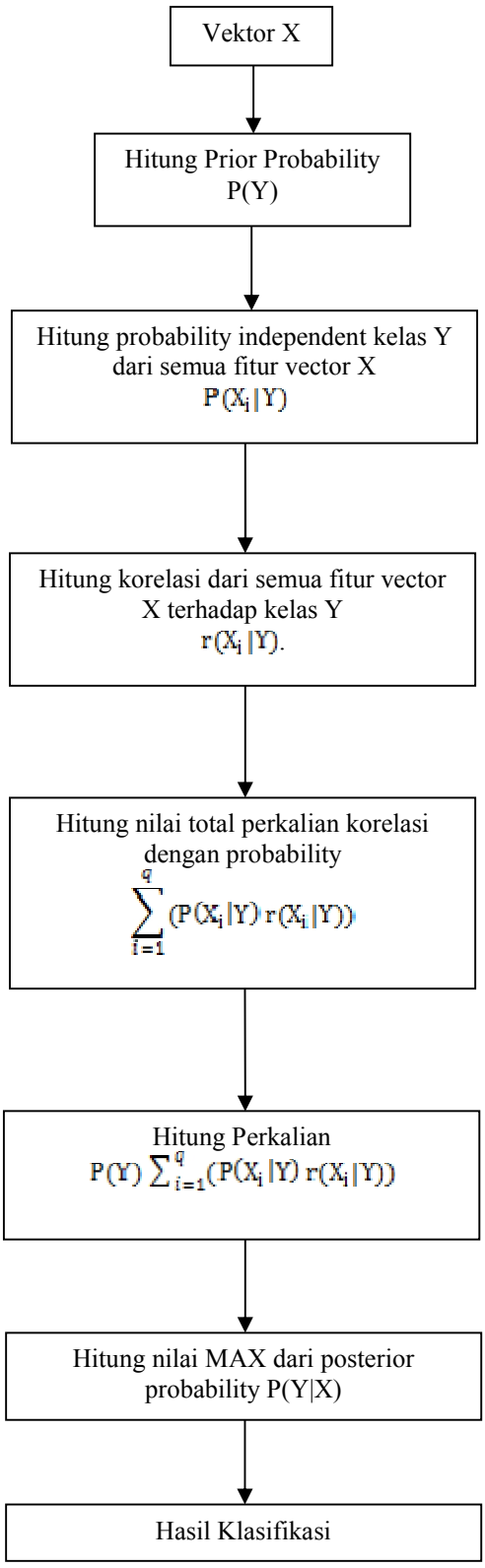
Dalam proses klasifiksi, algoritma *Naïve Bayes Classifier* dapat dijelaskan pada gambar 1.



Gambar 1.*Schematic Naïve Bayes Classifier*

Dengan melihat *schematic Naïve Bayes Classifier* pada Gambar 1 menunjukkan bahwa dalam proses klasifikasi *Naïve Bayes Classifier* berbasis pada *probability*. Dengan kata lain, untuk mendapatkan *posterior probability* P(Y|X) hanya berbasis pada *frequency* kemunculan fitur *vector* X dalam *data training*.

Untuk meningkatkan akurasi dilakukan *improve* pada saat proses perhitungan probabilitas *independent* kelas Y dari semua fitur *vector* X. *Improve* yang dilakukan adalah dengan memperhitungkan korelasi *value* dari fitur *vector* X terhadap kelas Y $r(X_i|Y)$. Korelasi dari fitur *vector* X terhadap kelas Y dapat dihitung dengan persamaan (4). Dengan memperhitungkan korelasi *value* fitur *vector* X, maka untuk mendapatkan *posterior probability* selain memperhitungkan *frequency* kemunculan fitur *vector* X dalam *data training* juga memperhitungkan seberapa besar pengaruh fitur *vector* X terhadap kelas Y. Proses klasifikasi dengan *Naïve Bayes Classifier* yang memperhitungkan korelasi *value* (*Correlated-Naïve Bayes Classifier*) dapat dilihat pada Gambar 2.



Gambar 2.*Schematic Correlated-Naïve Bayes Classifier*

Berdasarkan pada proses kerja yang ditunjukkan pada Gambar 2. Maka dapat ditentukan persamaan untuk menentukkan *posterior probability* dengan memperhitungkan korelasi *value* dari masing-masing fitur *vector* X terhadap kelas Y sebagai berikut :

$$P(Y|X) = \frac{P(Y)\sum_{i=1}^{q}(P(X_i|Y)r((X_i|Y))}{P(X)} \quad \ldots(6)$$

Dimana :

- P(Y|X) adalah probabilitas data dengan *vector* X pada kelas Y.
- P(Y) adalah probabilitas awal kelas Y.
- $\sum_{i=1}^{q}(P(X_i|Y)r((X_i|Y))$ adalah jumlah total dari perkalian antara probabilitas independent kelas Y dari fitur dalam *vector* X dengan nilai korelasi fitur dalam *vector* X terhadap kelas Y.
- P(X) adalah *evidence* atau probabilitas *vector* X.

Setelah ditemukan persamaan untuk *Correlated-Naïve Bayes Classsifier* seperti yang ditunjukkan pada persamaan (6), kemudian dilakukan uji coba untuk mengetahui tingkat akurasi algoritma *Correlated-Naïve Bayes Classifier*. Uji coba dilakukan terhadap beberapa data set dengan teknik *stage-fold* 10% dan dilakukan sebanyak 30 kali pada setiap *data set*. Teknik *stage-fold* adalah sebuah teknik pengujian yang sering digunakan. Teknik *stage-fold* 10% berarti dalam pengujian, akan diambil data secara acak sebesar 10% data dari *data set* untuk dijadikan *data testing*. Kemudian sisa dari *data set* atau data yang tidak terambil sebagai *data testing* akan digunakan sebagai *data training*. Agar dapat diketahui bahwa *improve* yang dilakukan terhadap algoritma *Naïve Bayes Classifier* benar-benar *significant* meningkatkan akurasi algoritma tersebut, maka dilakukan uji hipotesis dengan menggunakan uji z dengan *significant* level 0.01 (z = 2.325). Dimana untuk mendapatkan *z-score* atau z hitung dapat menggunakan persamaan (5). Dimana untuk hipotesis yang akan diuji adalah sebagai berikut :

H0 = akurasi *Correlated-Naïve Bayes Classifier* = akurasi *Naïve Bayes Classifier*.

H1 = akurasi *Correlated-Naïve Bayes Classifier* > akurasi *Naïve Bayes Classifier*.

Dalam pengimplementasian persamaan (6) terhadap 4 *data set* yang berbeda didapatkan hasil sebagai berikut :

A. Pengujian pertama

Hasil pengujian pertama ditunjukkan pada Tabel 2.

Tabel 2.*Tabel Hasil Pengujian Pertama*

Diskripsi	Nilai
n	30
μ akurasi NBC	91.7776667
μ akurasi C-NBC	94.22066667
σ	5.36739
z tabel α = 0.01	2.325
z hitung	2.492992326

Dari hasil yang dilihat dari Tabel 2, menunjukkan bahwa z hitung > z tabel. Hal ini berarti H0 ditolak dan H1 diterima. Sehingga pada pengujian pertama *improve* yang dilakukan dengan memperhitungkan korelasi pada *Naïve Bayes Classifier* (*Correlated-Naïve Bayes Classifier*) dapat meningkatkan akurasi sebesar 2.443 % secara significant.

B. Pengujian Kedua

Hasil pengujian kedua ditunjukkan pada Tabel 3.

Tabel 3.*Tabel Hasil Pengujian Kedua*

Diskripsi	Nilai
n	30
μ akurasi NBC	51.44433333
μ akurasi C-NBC	80.11033333
σ	4.92663
Z tabel α = 0.01	2.325
Z hitung	31.86968543

Dari hasil yang dilihat dari Tabel 3, menunjukkan bahwa z hitung > z tabel. Hal ini berarti H0 ditolak dan H1 diterima. Sehingga pada pengujian kedua *improve* yang dilakukan dengan memperhitungkan korelasi pada *Naïve Bayes Classifier* (*Correlated-Naïve Bayes Classifier*) dapat meningkatkan akurasi sebesar 28.666 % secara *significant*.

C. Pengujian Ketiga

Hasil pengujian ketiga ditunjukkan pada Tabel 4.

Tabel 4.*Tabel Hasil Pengujian Ketiga*

Diskripsi	Nilai
n	30
μ akurasi NBC	58.889
μ akurasi C-NBC	73.778
σ	10.69718
Z tabel α = 0.01	2.325
Z hitung	7.623542989

Dari hasil yang dilihat dari Tabel 4, menunjukkan bahwa z hitung > z tabel. Hal ini berarti H0 ditolak dan H1 diterima. Sehingga pada pengujian ketiga *improve* yang

dilakukan dengan memperhitungkan korelasi pada *Naïve Bayes Classifier* (*Correlated-Naïve Bayes Classifier*) dapat meningkatkan akurasi sebesar 14.889 % secara *significant*.

D. Pengujian Keempat

Hasil pengujian keempat ditunjukkan pada Tabel 5.

Tabel 5.*Tabel Hasil Pengujian Keempat*

Diskripsi	Nilai
n	30
μ akurasi NBC	77.7773333
μ akurasi C-NBC	84.889
σ	8.915
Z tabel α = 0.01	2.325
Z hitung	4.369288003

Dari hasil yang dilihat dari Tabel 5, menunjukkan bahwa z hitung > z tabel. Hal ini berarti H0 ditolak dan H1 diterima. Sehingga pada pengujian keempat *improve* yang dilakukan dengan memperhitungkan korelasi pada *Naïve Bayes Classifier* (*Correlated-Naïve Bayes Classifier*) dapat meningkatkan akurasi sebesar 7.111666667 % secara *significant*.

Dari keempat pengujian yang dilakukan, keempat pengujian tersebut memberikan hasil untuk menolak H0 dan menerima H1. Sehingga terbukti bahwa *Correlated-Naïve Bayes Classifier* atau algoritma *Naïve Bayes Classifier* yang sudah dimodifikasi dengan memperhitungkan nilai korelasi setiap fitur *vector* X secara *significant* dapat meningkatkan akurasi dalam proses klasifikasi.

3. Kesimpulan

Untuk meningkatkan akurasi pada algoritma *Naïve Bayes Classifier*, dapat dilakukan dengan memperhitungkan nilai korelasi dari masing-masing *attribute vector* X terhadap kelas Y. Sehingga yang menjadi parameter penentuan pemetaan suatu *vector* X yang belum diketahui kelasnya terhadap kelas yang sudah ditentukan menjadi dua hal, yaitu :

- *Frequency* kemunculan data dari setiap fitur *vector* X dalam *data training* (*probability*).
- Besar kecilnya pengaruh setiap fitur *vector* X terhadap kelas Y (korelasi).

Setelah melihat hasil pengujian hipotesis terhadap 4 *data set* yang berbeda, maka terbukti *improve* yang dilakukan dengan menambahkan *parameter* perhitungan korelasi pada algoritma *Naïve Bayes Classifier* dapat meningkatkan akurasi secara *significant*.

Daftar Pustaka

[1] Budi Santosa, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.

[2] Eko Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, 1st ed. CV ANDI OFFSET, 2012.

[3] F. A. Hermawati, *Data Mining*, 1st ed. Yogyakarta: CV ANDI OFFSET, 2013.

[4] A. Nurnberger, C. Borgelt, and A. Klose, "Improving naive Bayes classifiers using neuro-fuzzy learning," in *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on*, 1999, vol. 1, pp. 154–159 vol.1.

[5] M. Martinez-Arroyo and L. E. Sucar, "Learning an Optimal Naive Bayes Classifier," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 0-0 0, vol. 3, pp. 1236–1239.

[6] Mohammad Farhan Qudratullah, Sri Utami Zuliana, and Epha Diana Supandi, *STATISTIKA*, 1st ed. Yogyakarta: SUKA-Press UIN Sunan Kalijaga, 2012.

[7] Samsubar Saleh, *STATISTIK DESKRIPTIP*, 1st ed. Yogyakarta: Unit Penerbit dan Percetakan (UPP) AMP YKPN, 1998.

[8] B. D. A. Fadlisyah, *Statistika : Terapannya di Informatika*, 1st ed. Yogyakarta: Graha Ilmu, 2014.

[9] Harinaldi, *Prinsip - Prinsip Statistik untuk Teknik dan Sain*, 1st ed. Yogyakarta: Penerbit Erlangga, 2005.

Biodata Penulis

Burhan Alfironi Muktamar, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STMIK AMIKOM Yogyakarta, lulus tahun 2013. Saat ini menjadi mahasiswa pascasarjana Jurusan Teknik Elektro dan Teknologi Informasi, Falkultas Teknik di Universitas Gadjah Mada.

Noor Akhmad Setiawan, memperoleh gelar Sarjana Teknik (S.T.), Teknik Elektro Universitas Gadjah Mada Yogyakarta, lulus tahun 1998. Memperoleh gelar Magister Teknik (M.T.), Teknik Elektro Universitas Gadjah Mada Yogyakarta, lulus tahun 2003. Memperoleh gelar Doctor of Philosophy (Ph.D.), Electrical and Electronics Engineering Universiti Teknologi PETRONAS Malaysia, lulus tahun 2009. Saat ini menjadi Dosen di Jurusan Teknik Elektro dan Teknologi Informasi Universitas Gadjah Mada Yogyakarta.

Teguh Bharata Adji**,** lahir di Yogyakarta, Indonesia pada tanggal 20 September 1969. Gelar Sarjana diperoleh dari Universitas Gadjah Mada pada tahun 1995, gelar Master diperoleh dari Doshisha University, Kyoto pada tahun 2001, sedangkan gelar Doktor diperoleh dari Universiti Teknologi Petronas, Malaysia pada tahun 2010. Penelitian-penelitian dan publikasi-publikasinya banyak terdapat di bidang Pemrosesan Bahasa Alami (*Natural Language Processing*), Pengolahan Citra (*Image Processing*), Komputasi Paralel (*Parallel Computing*), Pesawat Nirawak (*Unmanned Aerial Vehicle*), Penambangan Data (*Data Mining*), dan Teknologi Animasi (Animation Technology).