ANALISIS PERBANDINGAN TINGKAT AKURASI ALGORITMA NAÏVE BAYES CLASSIFIER DENGAN CORRELATED-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

Dalam paper ini, penulis membahas tentang analisis perbandingan algoritma klasifikasi dalam data mining. Algoritma klasifikasi data mining yang akan dianalisis adalah algoritma Naïve Bayes Classifier dan Correlated-Naïve Bayes Classifier. Analisis perbandingan yang dimaksud adalah perbandingan tingkat akurasi dari kedua algoritma tersebut.

Naïve Bayes Classifier merupakan salah satu algoritma klasifikasi data mining yang berbasis pada probability value dari data set. Pola dasar Naïve Bayes Classifier berbasis pada teorema bayes. Secara garis besar cara kerja Naïve Bayes Classifier adalah merubah prior probability (probability awal) menjadi posterior probability (probability akhir) dengan melakukan perhitungan pada prior probability, likelihood (probability attribut) dan evidence.

Correlated-Naïve Bayes Classifier merupakan pengembangan dari algoritma Naïve Bayes Classifier. Naïve Bayes Classifier mengklasifikasikan class berdasarkan pada probability atau dengan kata lain bergantung pada frekuensi kemunculan data pada data set. Sedangkan Correlated-Naïve Bayes Classifier selain berdasar pada probability attribute, juga memperhitungkan nilai korelasi masing-masing attribute terhadap class. Sehingga untuk mendapat posterior probability, Correlated-Naïve Bayes Classifier memperhitungkan dua aspek yaitu tingkat kemunculan data (probability) dan tingkat hubungan attribute dengan class.

Hasil dari penelitian ini adalah sebuah pengetahuan baru tentang tingkat akurasi dari algoritma Naïve Bayes Classifier dan Correlated-Naïve Bayes Classifier. Dengan pengetahuan tersebut kita dapat membandingkan kehandalan antara kedua algoritma tersebut yang diukur berdasarkan tingkat akurasinya

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

1. Pendahuluan

Data mining merupakan proses yang mempekerjakan satu atau lebih teknik pembelajaran koputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. Definisi lainnya adalah pembelajaran berbasis induksi (*induction-based learning*) adalah proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsep-konsep yang akan dipelajari. *Knowledge Discovery in Database* (KDD) adalah penerapan metode saintifik pada data mining[1].

Data mining dibagi menjadi beberapa kelompok, salah satu pengelompokkan data mining tersebut adalah klasifikasi. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target f yang memetakan setiap set atribut (fitur) x ke satu dari sejumlah label kelas y yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori[2].

Klasifikasi mencoba untuk memetakan data ke dalam slah satu dari kelas yang telah ditetapkan atau sering disebut kategori diskrit. Klasifikasi merupakan satu set supervised method yang berarti bahwa kategori(class) sebelumnya harus didefinisikan dan dikenal oleh data yang digunakan sebagai training set[3].

Klasifikasi adalah salah satu teknik data mining yang menetapkan kasus dalam pengumpulan data untuk menargetkan kategori atau kelas. Tujuan klasifikasi adalah untuk secara akurat memprediksi kelas target untuk setiap kasus dalam data. Pertambangan

Association adalah Teknik lain data mining untuk menemukan menarik hubungan antara dua variabel[4].

Klasifikasi dibagi menjadi lima kelompok berdasarkan teori yang diadopsi atau teori yang menjadi dasar teknik klasifikasi. Lima pengelompokkan klasifikasi itu adalah *classifier Bayes's theorem*, *distance-based classifier*, *discriminant classifier*, *neural networks classifier*, dan *decision tree classifier*[5]. Pada paper ini akan focus terhadap *classifier Bayes's theorem* atau algoritma klasifikasi yang mengadopsi teorema bayes.

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[6].

Formula Teorema Bayes [6]:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad \ldots(1)$$

Naïve Bayes Classifier adalah sebuah *statistical classifier* yang sangat baik dalam data mining. Kaitan antara *naïve bayes* dengan klasifikasi, korelasi hipotesis, dan bukti 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 [2][7].
Formula *Naïve Bayes Classifier* [2]:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \ldots\ldots..(2)$$

Correlated-Naïve Bayes Classifier adalah modifikasi dan pengembangan dari algoritma *Naïve Bayes Classifier*. Metode klasifikasi dengan menggunakan algoritma *Naïve Bayes Classifier* berdasar pada nilai *probability attribute* dari suatu data yang belum diketahui *class*-nya. Sehingga teknik dasar dari klasifikasi algoritma ini berdasar pada frequensi kemunculan data pada data set. Selain berdasar pada *probability* atau tingkat kemunculan data pada data set, *Correlated-Naïve Bayes Classifier,* juga melibatkan perhitungan terhadap nilai korelasi masing-masing *attribute* terhadap *class*.

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

$$r = \frac{n\,x(\sum XY) - (\sum X)\,x\,(\sum Y)}{\sqrt{(n.\sum X^2 - (\sum X)^2)(n.\sum Y^2 - (\sum Y)^2)}} \quad \ldots(3)$$

(r) memiliki ketentuan -1<= r <= 1 dan interpretasi koefisien korelasi nilai (r) dirangkum pada Tabel 1[8].

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

Salah satu aspek yang menjadi parameter kehandalan dari suatu algoritma klasifikasi adalah tingkat akurasi. Sebuah sistem dalam melakukan klasifikasi diharapkan dapat mengklasifikasi semua set data dengan benar, tetapi tidak dipungkiri bahwa kinerja suatu sistem tidak bisa 100% akurat Berdasarkan perbedaan teknik antara algoritma *Naïve Bayes Classifier* dengan *Correlated-Naïve Bayes Classifier*, maka dalam paper ini akan dibahas mengenai perbandingan tingkat akurasi dari kedua algoritma klasifikasi tersebut[2].

- Untuk menghitung akurasi digunakan formula [2]:

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}}$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad \ldots(4)$$

- Untuk menghitung kesalahan prediksi (error) digunakan formula[2]:

$$\text{Error} = \frac{\text{Jumlah data yang diprediksi secara salah}}{\text{jumlah prediksi yang dilakukan}}$$

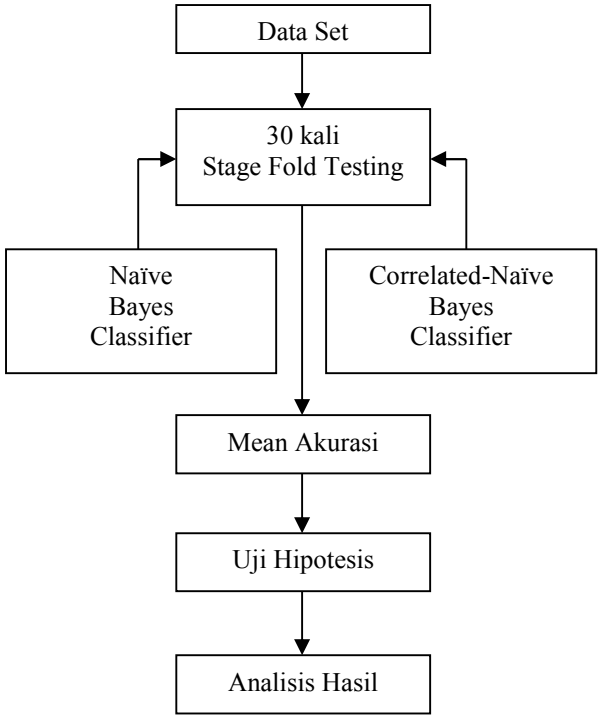
$$= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad \ldots(5)$$

Tujuan yang diinginkan dari penelitian ini adalah untuk mengetahui tingkat akurasi dari algoritma *Naïve Bayes Classifier* dan *Correlated-Naïve Bayes Classifier* serta membandingkan kehandalan antara kedua algoritma tersebut yang diukur berdasarkan tingkat akurasinya.

2. Pembahasan

Metode yang digunakan untuk membandingkan tingkat akurasi algoritma *Naïve Bayes Classifier* dengan *Correlated-Naïve Bayes Classifier* adalah dengan melakukan pengujian akurasi beberapa *data set*. Pada *paper* ini, *data set* yang digunakan diambil dari *uci repository* yaitu *: data set iris, data set balance-scale, data set haberman*, dan *data set servo*.

Pengujian algoritma *Naïve Bayes Classifier* dan *Correlated-Naïve Bayes Classifier* terhadap data set menggunakan metode *stage-fold*. Dengan metode *stage-fold* ini, *data testing* yang digunakan sebesar 10% dari *data set* yang diambil secara acak. Untuk masing-masing *data set* diuji sebanyak 30 kali kemudian diambil rata-rata akurasi untuk masing-masing algoritma. Setelah mendapatkan rata-rata akurasi dari masing-masing algorima, dilakukan uji z untuk menguji hipotesis yang ditentukan agar mendapatkan hasil yang *significant*. Untuk memperjelas alur dari pengujian, schema pengujian untuk membandingkan algoritma *Naïve Bayes Classifier* dan *Correlated-Naïve Bayes Classifier* dapat dilihat pada Gambar 1.



Gambar 1.*Schema Pengujian*

Pada pengujian pertama untuk membandingkan algoritma *Naïve Bayes Classifier* dan *Correlated-Naïve Bayes Classifier* dilakukan implementasi algoritma terhadap *data set iris*. Nilai akurasi dari hasil pengujian ditunjukkan pada Tabel 2. Dalam tabel hasil pengujian kolom Tes menunjukkan nomor pengujian, kolom NBC (%) menunjukkan akurasi klasifikasi dengan algoritma *Naïve Bayes Classfier* dalam persen, dan C-NBC(%) menunjukkan akurasi klasifikasi dengan algoritma *Correlated-Naïve Bayes Classifie*r dalam persen.

Tabel 2.*Tabel Hasil Pengujian Data Set Iris*

Tes	NBC (%)	C-NBC (%)
1	100	100
2	100	100
3	100	100
4	93.33	93.33
5	86.67	93.33
6	86.67	93.33
7	100	93.33
8	80	80
9	73.33	80
10	100	100
11	93.33	93.33
12	86.67	93.33
13	93.33	100
14	93.33	93.33
15	93.33	93.33
16	86.67	93.33
17	86.67	80
18	100	100
19	100	100
20	100	93.33
21	86.67	86.67
22	86.67	93.33
23	86.67	93.33
24	86.67	100
25	93.33	100
26	86.67	93.33
27	93.33	100
28	93.33	100
29	93.33	93.33
30	93.33	93.33

Dari data hasil pengujian data set iris pada Tabel 2, didapatkan nilai rata-rata akurasi dengan algoritma *Naïve Bayes Classifier* ($\mu 1$) adalah 91.77766667 dan nilai rata-rata akurasi dengan algoritma *Correlated-Naïve Bayes Classifier* ($\mu 2$) adalah 94.22066667. Beberapa informasi lain yang diperoleh dari data tersebut yaitu jumlah data (n) = 30, standard deviasi (s) = 5.36739.

Untuk mendapatkan hasil yang *significant*, dilakukan uji z untuk menguji hipotesis dengan *significant* level 0.01. Hipotesis yang akan diuji *significant*-nya sebagai berikut:

H0 : $\mu 2 = \mu 1$
H1 : $\mu 2 > \mu 1$

Setelah dilakukan perhitungan, didapatkan nilai z hitung = 2.492992326. Karena z hitung lebih besar dari z tabel untuk *significant* level 0.01 (z = 2.325), maka H0 ditolak dan H1 diterima. Sehingga pada pengujian *data set iris*, tingkat akurasi algoritma *Correlated-Naïve Bayes Classifier significant* lebih besar dibandingkan dengan tingkat akurasi algoritma *Naïve Bayes Classifier*.

Pada pengujian kedua, implementasi algoritma *Naïve Bayes Classifier* dilakukan pada *data set balance-scale*. Hasil pengujian *data set balance-scale* dapat dilihat pada Tabel 3.

Tabel 3.*Tabel Hasil Pengujian Data Set Balance-Scale*

Tes	NBC (%)	C-NBC (%)
1	60	70
2	70	58.33
3	55	76.67
4	63.33	68.33
5	56.67	78.33
6	56.67	68.33
7	56.67	75
8	53.33	68.33
9	55	71.67
10	53.33	70
11	55	78.33
12	50	76.67
13	55	81.67
14	48.33	78.33
15	53.33	83.33
16	45	80
17	51.67	81.67
18	43.33	83.33
19	51.67	85
20	56.67	90
21	45	86.67
22	55	90
23	43.33	83.33
24	50	88.33
25	41.67	90
26	48.33	86.67
27	43.33	93.33
28	45	90
29	36.67	83.33
30	45	88.33

Dari data hasil pengujian *data set balance scale* pada Tabel 3, didapatkan beberapa informasi yaitu : nilai rata-rata akurasi dengan algoritma *Naïve Bayes Classifier* (μ1) adalah 51.44433333, nilai rata-rata akurasi dengan algoritma *Correlated-Naïve Bayes Classifier* (μ2) adalah 80.11033333, jumlah data (n) = 30, standard deviasi (s) = 4.92663.

Langkah berikutnya adalah melakukan uji z untuk uji hipotesis dengan *significant* level 0.01. Setelah dilakukan perhitungan, didapatkan nilai z hitung = 31.86968543. Karena z hitung lebih besar dari z tabel untuk *significant* level 0.01 (z = 2.325), maka H0 ditolak dan H1 diterima. Sehingga tingkat akurasi algoritma *Correlated-Naïve Bayes Classifier significant* lebih besar dibandingkan dengan tingkat akurasi algoritma *Naïve Bayes Classifier*.

Pada pengujian ketiga, implementasi algoritma *Naïve Bayes Classifier* dan algoritma *Correlated-Naïve Bayes Classiferi* dilakukan pada *data set haberman*. Setelah dilakukan pengujian sebanyak 30 kali, hasil pengujian *data set haberman* dapat dilihat pada Tabel 4.

Tabel 4.*Tabel Hasil Pengujian Data Set Haberman*

Tes	NBC (%)	C-NBC (%)
1	33.33	73.33
2	80	86.67
3	40	73.33
4	26.67	80
5	46.67	60
6	66.67	66.67
7	66.67	73.33
8	60	66.67
9	66.67	66.67
10	80	93.33
11	60	66.67
12	46.67	60
13	60	80
14	60	80
15	73.33	73.33
16	80	93.33
17	73.33	66.67
18	73.33	80
19	60	66.67
20	60	80
21	46.67	73.33
22	46.67	66.67
23	60	80
24	60	73.33
25	13.33	60
26	53.33	66.67
27	46.67	80
28	73.33	66.67

29	80	93.33
30	93.33	93.33

Dari data hasil pengujian *data set haberman* pada Tabel 4, didapatkan beberapa informasi yaitu : nilai rata-rata akurasi dengan algoritma *Naïve Bayes Classifier* (μ1) adalah 58.889, nilai rata-rata akurasi dengan algoritma *Correlated-Naïve Bayes Classifier* (μ2) adalah 73.778, jumlah data (n) = 30, standard deviasi (s) = 10.69718.

Pada tahap uji z untuk uji hipotesis dengan *significant* level 0.01, didapatkan nilai z hitung = 7.623542989. Karena z hitung lebih besar dari z tabel untuk *significant* level 0.01 (z = 2.325), maka H0 ditolak dan H1 diterima. Sehingga tingkat akurasi algoritma *Correlated-Naïve Bayes Classifier significant* lebih besar dibandingkan dengan tingkat akurasi algoritma *Naïve Bayes Classifier*.

Pada pengujian keempat, hasil pengujian *data set servo* dapat dilihat pada Tabel 5.

Tabel 5.*Tabel Hasil Pengujian Data Set Servo*

Tes	NBC (%)	C-NBC (%)
1	46.67	80
2	73.33	86.67
3	86.67	93.33
4	80	73.33
5	73.33	100
6	73.33	80
7	86.67	80
8	73.33	100
9	80	86.67
10	80	66.67
11	73.33	66.67
12	73.33	86.67
13	80	80
14	80	93.33
15	86.67	80
16	80	86.67
17	80	93.33
18	73.33	80
19	73.33	93.33
20	93.33	86.67
21	60	73.33
22	80	100
23	66.67	80
24	93.33	80
25	66.67	93.33
26	73.33	80
27	93.33	93.33
28	86.67	86.67
29	66.67	86.67
30	100	80

Dari data hasil pengujian *data set servo* pada Tabel 5, didapatkan beberapa informasi yaitu : nilai rata-rata akurasi dengan algoritma *Naïve Bayes Classifier* (μ1) adalah 77.77733333, nilai rata-rata akurasi dengan algoritma *Correlated-Naïve Bayes Classifier* (μ2) adalah 84.889, jumlah data (n) = 30, standard deviasi (s) = 8.915.

Pada tahap uji z untuk uji hipotesis dengan *significant* level 0.01, didapatkan nilai z hitung = 4.369288003. Karena z hitung lebih besar dari z tabel untuk *significant* level 0.01 (z = 2.325), maka H0 ditolak dan H1 diterima. Sehingga tingkat akurasi algoritma Corre*lated-Naïve Bayes Classifier significant* lebih besar dibandingkan dengan tingkat akurasi algoritma *Naïve Bayes Classifier*.

Setelah dilakukan pengujian terhadap empat data set, ringkasan hasil pengujian dapat dilihat pada Tabel 6.

Tabel 6.*Tabel Hasil Pengujian Data Set Servo*

Data Set	Mean akurasi NBC (%)	Mean akurasi C-NBC (%)
Iris	91.77766667	94.22066667
Balance-Scale	51.44433333	80.11033333
Haberman	58.889	73.778
Servo	77.77733333	84.889

3. Kesimpulan

Setelah dilakukan pengujian terhadap *data set iris, data set balance-scale, data set haberman*, dan *data set servo* serta dilakukan uji z untuk mendapatkan hasil yang *significant*, didapatkan hasil bahwa tingkat akurasi algoritma *Correlated-Naïve Bayes Classifier* lebih besar dibandingkan dengan tingkat akurasi algoritma *Naïve Bayes Classifier*. Perbedaan akurasi untuk *data set iris* sebesar 2.443%, perbedaan akurasi untuk *data set balance-scale* sebesar 28,666%, perbedaan akurasi untuk *data set haberman* sebesar 14.889, dan perbedaan akurasi untuk *data set servo* sebesar 7.1116667%.

Data set yang digunakan pada pengujian pada *paper* ini memiliki jumlah attribute berkisar antara 4 sampai dengan 5 *attribute*. Untuk kedepannya diharapkan pengujian dapat dilakukan untuk *data set* yang memiliki *attribute* yang lebih banyak. Untuk menambah *significant* hasil, analisis perbandingan algoritma *Naïve Bayes Classifier* dengan *Correlated-Naïve Bayes Classifier* dapat diuji dengan *data set* yang dikelompokkan berdasarkan *attribute type*. *Attribute type* yang dimaksud meliputi *categorical, integer, real*, atau *mix* dari ketiganya.

Daftar Pustaka

[1] F. A. Hermawati, *Data Mining*, 1st ed. Yogyakarta: CV ANDI OFFSET, 2013.
[2] Eko Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, 1st ed. CV ANDI OFFSET, 2012.
[3] S. V. Stankovic, G. Rakocevic, N. Kojic, and D. Milicev, "A classification and comparison of Data Mining algorithms for Wireless Sensor Networks," in *Industrial Technology (ICIT), 2012 IEEE International Conference on*, 2012, pp. 265–270.
[4] S. D. Pandya and P. V. Virparia, "Comparing the application of classification and association rule mining techniques of data mining in an Indian university to uncover hidden patterns," in *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*, 2013, pp. 361–364.
[5] T. Justin, R. Gajšek, V. Štruc, and S. Dobrišek, "Comparison of different classification methods for emotion recognition," in *MIPRO, 2010 Proceedings of the 33rd International Convention*, 2010, pp. 700–703.
[6] Budi Santosa, *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*, 1st ed. Yogyakarta: Graha Ilmu, 2007.
[7] C. Shah and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction," in *Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference on*, 2013, pp. 1–4.
[8] B. D. A. Fadlisyah, *Statistika : Terapannya di Informatika*, 1st ed. Yogyakarta: Graha Ilmu, 2014.

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).