Identifikasi Keakuratan Data Pelanggan PT.XYZ dengan menggunakan C4.5, Naïve Bayes dan Algoritma Preprocessing

Taransa Agasya Tutupoly, Ibnu Alfarobi

Abstract *— The accuracy of customer data in a company is very important, especially in corporate loyalty, where it takes a lot of data in campaign, promotion, digital marketing, email blast, sms blast which aims to create brand awareness from the customer side and of course increase the level of sales and liveliness members in doing transactions using point, the current method used to measure the accuracy of data using Naive Bayes, C4.5 and help Algorithm Preprocessing to improve data accuracy, from the test results that the processing using the method Naive Bayes, C4.5 has an accuracy of 96.2 % with AUC 0.9627 while the test results using Naive Bayes, C4.5 and Preprocessing Algorithm obtained 99% yield, with AUC 0.9908, it can be proven that the addition of processing with Preprocessing Algorithm get more optimal results*

Intisari— Penulisan ini digunakan mencari Keakuratan data pelanggan dalam suatu perusahaan amatlah penting, khususnya pada perusahaan loyalty, dimana dibutuhkan banyak data dalam melakukan campaign, promosi, digital marketing, email blast, sms blast yang tujuannya membuat brand awareness dari sisi pelanggan dan tentunya meningkatkan tingkat penjualan dan keaktifan member dalam melakukan transaksi menggunakan point, saat ini metode yang digunakan untuk mengukur tingkat akurasi data menggunakan Naive Bayes, C4.5 dan bantuan Algoritma Preprocessing untuk meningkatkan akurasi data. Sehingga didapatkan hasil bahwa dengan proses preprocessing hasil pengukuran menjadi lebih baik.

Kata Kunci— Algoritma *Preprocessing*, Algoritma C4.5, Algoritma *Naive Bayes*, Data Pelanggan.

I. PENDAHULUAN

A. Latar Belakang Masalah

Dewasa ini keakuratan data sangatlah penting, salah satunya dalam bidang usaha yang erat dengan data pelanggan. karena itu CRM *(Customer Relationship Management)* merupakan salah satu konsep terbaik yang dapat digunakan untuk mengetahui perilaku atau pola dari pelanggan. sebelum masuk kedalam tahap tersebut identifikasi dan pemeriksaan data dari pelanggan setia sangatlah perlu dilakukan pengecekan secara detail. Karena factor *Human Error* dan Faktor Manipulasi data pada *system* pendaftaran pelanggan baru sangatlah mudah dilakukan. Penerapan *Data Mining* sangat diperlukan dalam hal ini karena dapat membantu menganalisa data yang diperoleh pada *system* sehingga dapat menggali pola – pola yang dapat dijadikan pengetahuan baru untuk proses identifikasi keakuratan data pelanggan. Teknik *Data Mining* yang digunakan untuk mencari keakuratan data adalah dengan menggunakan teknik *klasifikasi*. Dalam penelitian ini akan digunakan dua macam teknik klasifikasi C4.5 & Naïve Bayes. C4.5 sendiri memiliki tingkat keakuratan dan kecepatan yang sangat baik dibandingkan dengan algoritma klasifikasi lainnya, dari data yang beragam atau kompleks, C4.5 mampu mengubah pola yang ada dengan sangat mudah dan simple. Memangkas perhitungan – perhitungan yang tidak diperlukan karena ketika dilakukan pengujian hasilnya adalah iya atau tidak. Keuntungan menggunakan algoritma Naïve Bayes karena dapat digunakan pada berbagai macam *klasifikasi* dokumen dan dapat dengan mudah diintegrasikan dengan algoritma baru yang akan dibangun untuk melakukan *filtering*, *deteksi spam* dan *cleansing data*, dapat menangani *kuantitatif* dan *data diskrit*, hanya memerlukan sampel data yang sedikit untuk mengestimasi parameter peluang bersayarat data, cepat dan memiliki efesiensi ruang yang besar.Oleh karena itu penulis akan membandingkan hasil keakuratan data yang diolah dengan C4.5 & Naïve Bayes dengan C4.5 & Naïve Bayes dengan bantuan *algoritma preprocessing* yang akan digunakan untuk melakukan cleansing data dan filtering guna mengidentifikasikan kecurangan data registrasi pelanggan yang dilakukan karyawan PT XYZ, guna mendapatkan incentive dari perusahaan.

B. Identifikasi Permasalahan

1. Meningkatkan keakuratan data pada PT.XYZ
2. Management mengetahui secara detail perolehan anggota baru yang direcruit oleh SPG

C. Perumusan Masalah

Algoritma Naïve Bayes, C4.5 dan Algoritma Preprocessing dapat memiliki keakuratan data yang lebih tinggi dibandingkan kombinasi Naïve Bayes, C4.5 dan Algoritma Preprocessing dalam identifikasi kecurangan pegawai PT.XYZ terhadap registrasi pelanggan baru, serta algoritma Naïve Bayes, C4.5 dan Algoritma Preprocessing dapat memiliki kecepatan kinerja yang lebih tinggi dibandingkan kombinasi Naïve Bayes dan C4.5 dalam identifikasi kecurangan pegawai PT. XYZ registrasi pelanggan baru.

[1,2] Teknik Informatika*; STMIK Nusa Mandiri Jakarta; Jalan Kramat Raya No.18, RT 01/ RW 07, Kwitang, Senen, Jakarta Pusat, 10420; Telp: 021-31908575; email: taransa.tly@nusamandiri.ac.id, ibnu.iba@bsi.ac.id*

D. Maksud dan Tujuan

Untuk membuktikan apakah keakuratan data dengan metode Naïve Bayes, C4.5 dan Algoritma Preprocessing lebih baik dari pada Naïve Bayes dan C4.5 dalam identifikasi kecurangan pegawai PT. XYZ terhadap registrasi pelanggan baru serta membuktikan apakah kecepatan kinerja Naïve Bayes, C4.5 dan Algoritma Preprocessing lebih baik dari pada Naïve Bayes dan C4.5 dalam identifikasi kecurangan pegawai PT. XYZ terhadap registrasi pelanggan baru.

E. Metode Penelitian

Dalam hal ini penulis menggunakan tehnik pengumpulan data sebagai berikut:

1. Data Selection
   Memilih data yang akan digunakan dalam proses data mining. Dalam proses ini dilakukan juga pemilihan atribut-atribut yang disesuaikan dengan proses data mining.
2. Wawancara
   Melakukan wawancara langsung kepada pegawai PT. XYZ
3. Studi Pustaka
   Metode kepustakaan yaitu dengan mencari dan mempelajari buku – buku atau jurnal yang relevan guna memberi pemahaman yang lebih baik terhadap topik penulisan dan memperkaya pengetahuan tentang penelitian ilmiah.

F. Ruang Lingkup

Topik pembahasan penulisan ini hanya terbatas pada data pelanggan PT.XYZ yang dipilih secara acak selama 3 tahun terakhir.

II. BAHAN DAN METODE

A. Desain Penelitian

Terdapat empat metode penelitian yang umum digunakan yaitu *Action Reserch, Experiment, Case Study*, dan *Survey* [2]. Adapun metode penelitian yang digunakan adalah bentuk penelitian *Experiment.* Penelitian eksperimen merupakan sebuah penyelidikan hubungan kausal menggunakan tes dikendalikan oleh peneliti [2]. Dalam eksperimen biasanya terdiri dari:

1. Mendefinisikan hipotesis teoritis
2. Memilih sampel dari populasi y
3. Mengalokasikan sampel untuk kondisi
4. Mengukur sejumlah kecil variabel
5. Mengontrol semua variable

B. Pengumpulan Data

Pada tahap ini struktur basis data akan dipersiapkan sehingga mempermudah proses mining. Proses preparation ini mencakup tiga hal utama yaitu:

1. Data Selection:
   Memilih data yang akan digunakan dalam proses data mining. Dalam proses ini dilakukan juga pemilihan atribut-atribut yang disesuaikan dengan proses data mining. Dalam penelitian dilakukan seleksi data pelanggan dari tahun 2013 sampai dengan 2016 sebanyak 1000 data yang diambil secara acak.
2. Data Preprocessing
   Memastikan kualitas data yang telah dipilih pada tahap data selection, pada tahap ini masalah yang harus dihadapi adalah noisy data dan missing values. Proses pembersihan data (c*leansing*) dilakukan dengan melakukan metode-metode *query* sederhana untuk menemukan anomali-anomali data yang bisa saja masih terdapat pada sistem. Dalam Penelitian ini akan dilakukan pembersihan data dengan menggunakan PHP dan Algoritma Preprocessing.
3. Algoritma C4.5
   C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi danprediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadipohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari record pada kategori tertentu. Pohon Keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, dia sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain. Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan pada variabel tujuannya.
4. *Naive Bayes*
   Naive Bayes merupakan salah satu metode*machine learning* yang menggunakan perhitunganprobabilitas. Algoritma ini memanfaatkan metodeprobabilitas dan statistik yang dikemukakan olehilmuwan Inggris bernama Thomas Bayes, yaitumemprediksi probabilitas di masa depan berdasarkanpengalaman di masa sebelumnya. Algoritma pengklasifikasi Naive Bayes adalah pengklasifikasi yangberdasarkan probabilitas bersyarat pada teorema Bayes [1]
5. Algoritma Preprocessing
   Merupakan tahapan dalam mengolah data input sebelum memasuki tahapan utama dari metode Lantet Semantic Analysis (LSA).

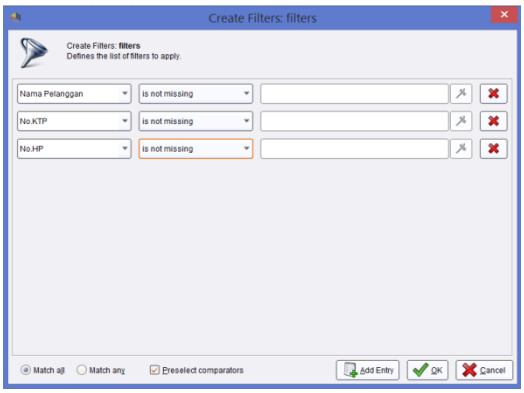
TABEL 1.
DATA PARSIAL DATABASE

Field	Keterangan
Nama Customer	Nama Customer
No. Kartu	Cek Transaksi, Bahwa Kartu valid digunakan
No. HP	SMS Blast Berhasil dilakukan
Alamat Email	Email Blast Berhasil dilakukan

Dalam proses *data preparation* dibangun suatu *data warehouse* untuk mempermudah proses *mining*. Data yang akan dilibatkan dalam penelitian ini adalah data yang berkaitan dengan transaksi penjualan.

Pada tahapan awal akan diambil 1000 sampling data excel asli pelanggan yang diambil dari PT. XYZ Data diambil secara acak selama tahun 3 tahun terakhir. Pada tahap ini dari data mentah yang sudah ada, akan dilakukan filtering dengan menggunakan tools rapid minner. Dimana kita akan mensortir seluruh data yang kosong dan tidak memiliki isi, khususnya pada bagian yang penting seperti, Nama Pelanggan, No. KTP dan No. Handphone, dimana ketiga data tersebut tidak boleh ada satupun yang kosong. Dari 1.000 data yang ada sebanyak 13 Data yang missing atau kosong.

Pada penelitian ini, perlakuan khusus yang diberikan untuk menangani *missing value* adalah dengan memberikan nilai rata-rata dari atribut. Teknik ini dapat diterapkan untuk atribut yang mempunyai nilai numerik.



Gbr 1. Filtering Data Pada Riped Minner

C. Pengujian Model

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tiga metode klasifikasi data mining. Algoritma yang akan digunakan adalah C4.5 dan Naive Bayes, serta penggabungan dari C4.5, Naive Bayes dan Algoritma Preprocessing. Setelah diolah dan menghasilkan model, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menggunakan *k-fold cross validation* dengan perbandingan antara *data testing* dan *data training* 10 : 90, 20 : 80, 30 : 70 dan mengulang pengujian tersebut beberapa kali.

Model Algoritma Klasifikasi Naïve Bayes, C4.5 dan Algoritma Preprocessing

Data hasil dari pengolahan naïve bayes dan filtering dengan php, akan diklasifikasikan dengan algoritma C4.5, berikut langkah yang dilakukan :

Menghitung data pelanggan yang VALID dan TIDAK VALID dari semua data yang ada, melakukan seluruh perhitungan entropy bedasarkan dan Tabel diatas dihitung bedasarkan sampling data yang ada.

$$\text{Entropy}(S) = \sum_{i=1}^{n} - pi * \text{Log2}\, pi \text{.....(1)}$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} \text{.....(2)}$$

D. Evaluasi

Pada tahap evaluasi, disebut tahap klasifikasi karena pada tahap ini akan ditentukan pengujian untuk akurasi. Tahap pengujiannya adalah melihat hasil akurasi pada proses klasifikasi Algoritma Naïve Bayes, C4.5 dan *Algoritma Preprocessing*.

E. Deployment

Pada tahapan *deployment*, dilakukan penerapan model algoritma klasifikasi Naïve Bayes, C4.5 dan algoritma preprocessing untuk menentukan keakuratan data pelanggan PT. XYZ yang dalam pengerjaannya dilakukan oleh SPG

III. HASIL PENELITIAN DAN PEMBAHASAN

A. Penelitian

Tujuan dari penelitian ini untuk mengetahui keakuratan data pelanggan di PT.XYZ, selain itu adalah mengembangkan metode klasifikasi yang sudah ada dengan penambahan teknik yang dilakukan, dalam penelitian ini akan membandingkan tingkat akurasi data dengan menggunakan Algoritma Naïve Bayes dan C4.5 dibandingkan dengan Naïve Bayes, C4.5 dan Algoritma Preprocessing dengan bantuan tools php dan rapid minner.

B. Pengujian dan Model Evaluasi

Algoritma Naïve Bayes dan C4.5, Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan nilai *accuracy* dan nilai AUC (*Area Under Curve*), Evaluasi model dengan *Confension Matrix*, Model *confesion matrix* akan membentuk matrix yang terdiri dari *true positif* atau tupel positif dan *true negatif* atau tupel negatif. Berikut dibawah ini merupakan hasil dari *confesion matrix* pada algoritma Naïve Bayes dan C4.5

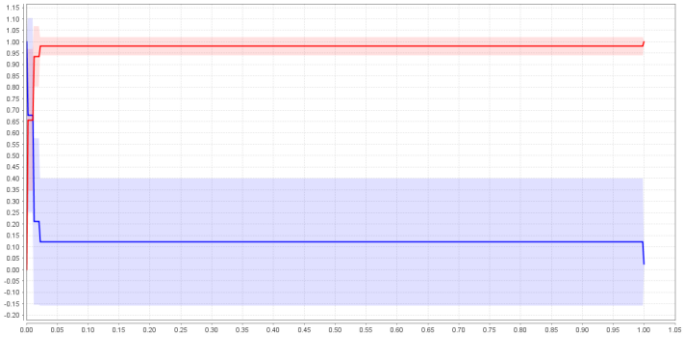
accuracy: 96.20% +/- 1.83% (mikro: 96.20%)

	true Valid	true Tidak Valid	class precision
pred. Valid	899	31	96.67%
pred. Tidak Valid	7	63	90.00%
class recall	99.23%	67.02%	

Gbr 2. Confusion Matrix Metode Algoritma Naïve Bayes & C4.5

Evaluasi dengan ROC
Pada Tabel 2 di bawah ini menunjukan grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.962. Akurasi memiliki tingkat diagnosa yaitu (Gorunescu, 2011):
Akurasi bernilai 0.90 – 1.00 = *Excellent classification*
Akurasi bernilai 0.80 – 0.90 = *Good classification*
Akurasi bernilai 0.70 – 0.80 = *Fair classification*
Akurasi bernilai 0.60 – 0.70 = *Poor classification*
Akurasi bernilai 0.50 – 0.60 = *Failure*

Sedangkan hasil yang didapat dari pengolahan ROC yang dapat dilihat pada Tabel 1 sebesar 0.962 dengan tingkat diagnosa *Excellent classification*.



Gbr 3. Nilai AUC dalam grafik Naïve Bayes dan C4.5 Algoritma *Naïve Bayes*, C4.5 dan Algoritma Preprocessing

Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan nilai *accuracy* dan nilai AUC (*Area Under Curve*) Evaluasi model dengan *Confension Matrix*

Model *confesion matrix* akan membentuk matrix yang terdiri dari *true positif* atau tupel positif dan *true negatif* atau tupel negatif. Berikut dibawah ini merupakan hasil dari *confesion matrix* pada algoritma Naïve Bayes, C4.5 dan Algoritma Preprocessing:

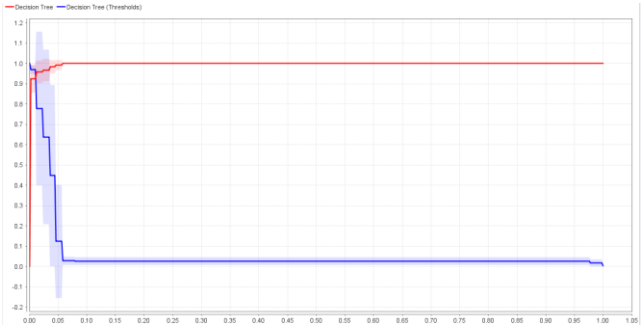
accuracy: 99.00% +/- 0.77% (mikro: 99.00%)

	true Valid	true Tidak Valid	class precision
pred. Valid	877	7	99.21%
pred. Tidak Valid	3	113	97.41%
class recall	99.66%	94.17%	

Gbr 4. *Confesion Matrix Metode Algoritma Naïve Bayes, C4.5 dan Preprocessing*

Evaluasi dengan ROC
Pada Tabel di bawah ini menunjukan grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.99 Akurasi memiliki tingkat diagnosa yaitu (Gorunescu, 2011):
Akurasi bernilai 0.90 – 1.00 = *Excellent classification*
Akurasi bernilai 0.80 – 0.90 = *Good classification*
Akurasi bernilai 0.70 – 0.80 = *Fair classification*
Akurasi bernilai 0.60 – 0.70 = *Poor classification*
Akurasi bernilai 0.50 – 0.60 = *Failure*

Sedangkan hasil yang didapat dari pengolahan ROC yang dapat dilihat pada Gambar 4.2 sebesar 0.99 dengan tingkat diagnosa Excellent *classification*.



Gbr 5. Nilai AUC dalam grafik ROC algoritma *Naïve Bayes*, C4.5 dan *Preprocessing*

Dari hasil pengujian diatas, dengan dilakukan evakuasi baik secara *confension matrik* maupun ROC *curve* ternyata terbuktu bahwa pengujian yang dilakukan dengan algoritma klasifikasi Naïve Bayes, C4.5 dan algoritma preprocessing memiliki nilai akurasi yang lebih tinggi dibanding hanya menggunakan algoritma klasifikasi Niave Bayes dan C4.5 saja. Nilai akurasi untuk model algoritma klasifikasi Naïve Bayes dan C4.5 sebesar 96,2% sedangkan nilai akurasi untuk model algoritma Naïve Bayes, C4.5 dan algoritma preprocessing berbasis PSO sebesar 99% dengan selisih akurasi sebesar 2,8%, dapat dilihat pada tabel di bawah ini:

TABEL 2.
PENGUJIAN ALGORITMA KLASIFIKASI C4.5 DAN C4.5 BERBASIS PSO

	Accuracy	AUC
Algoritma Klasifikasi Naïve Bayes + C4.5	96,2%	0,9627
Algoritma Klasifikasi Naïve Bayes + C4.5+PSO	99 %	0,9908

.

Dengan demikian algoritma Naïve Bayes, C4.5 dan bantuan pengolahan preprocessing dapat memberikan solusi untuk mengetahui penentuan keakuratan data pelanggan PT. XYZ, dan tolak ukur efisiensi dan keakuratan pekerjaan SPG.

IV. KESIMPULAN DAN SARAN

A. Kesimpulan

Dari penelitian yang telah dilakukan untuk mengetahui keakuratan Data Pelanggan PT. XYZ yang diambil dari 1000 data acak yang tersebar diseluruh indonesia dengan penggunaan metode Naïve Bayes, C4.5 dan Algoritma Preprocessing mendapatkan hasil yang optimal.

Dari hasil pengujian dengan mengukur kinerja ketiga metode tersebut menggunakan *confusion matrix* dan kurva *ROC* diketahui bahwa Naïve Bayes, C4.5 dan algoritma preprocessing menghasilkan nilai akurasi yang baik 99 % dan nilai AUC 0,9908. Sedangkan dengan perhitungan Algoritma Naïve Bayes dan C4.5 diperoleh hasil akurasi 96,2 dengan nilai AUC 0,9627 memiliki selisih 2,8 persen lebih baik hasilnya dibandingkan dengan Naïve Bayes dan C4.5. Dengan demikian metode Naïve Bayes, C4.5 dan Algoritma Preprocessing merupakan metode terbaik dalam hal keakuratan data.

B. Saran

Untuk keperluan penelitian lebih lanjut mengenai komparasi metode klasifikasi data mining dengan menggunakan data dibidang pelanggan dapat menggunakan optimasi lain seperti Genetic Algorithm, Chi Square dan sebagainya untuk mendapatkan algoritma yang paling akurat.

UCAPAN TERIMA KASIH

Sebagai penulis saya mengucapkan terima kasih kepada Tuhan Yang Maha Esa, karena atas berkatnya saya dapat menyelesaikan tulisan ini, serta keluarga yang selalu memberikan support kepada saya untuk selalu melakukan riset yang baik, serta kepada para pihak yang saya tidak dapat sebutkan satu persatu.

REFERENSI

[1] Anggarwal, Charu C. (2015). *Data Mining: The Textbook.* New York: Springer. Blaxter, L., Hughes, C., & Tight, M. (2010). *How to Research* (4th ed).

[2] Maidenhead: Open University Press. Dawson, C. W. (2009). *Projects in Computing and Information Systems a student's guide*. Harlow, UK: Addison-Wesley.

[3] Gorunescu, Florin (2011). *Data Mining: Concepts, Models, and Techniques.*Verlag Berlin Heidelberg: Springer

[4] Han, J., &Kamber.,& Pei, J. (2012). Data Mining Consepts and Techniques. San Fransisco: Morgan Kauffman.

[5] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery.Handbook.* London: Springer.

[6] Mantas, C. J., & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, *41*(10), 4625–4637. doi:10.1016/j.eswa.2014.01.017.

[7] Sammut, Claude. (2011). *Encyclopedia of Machine Learning*. Boston, MA: Springer.

[8] Setiyorini, T., Pascasarjana, P., Ilmu, M., Tinggi, S., Informatika, M., Komputer,D. a N., & Mandiri, N. (2014a). Penerapan Metode Bagging UntukMengurangi Data Noise Pada Neural Network Untuk Estimasi Kuat TekanBeton Penerapan Metode Bagging Untuk Mengurangi Data Noise Pada NeuralNetwork Untuk, 1(1), 36–41.

[9] Vercellis, C. (2009). *Business Intelligence : Data Mining and Optimization for.Decision Making.* John Wiley & Sons, Ltd.

Taransa Agasya Tutupoly Lahir di Bogor 25 Agustus 1989 , Lulusan S2 Ilmu Komputer STMIK Nusa Mandiri, saat ini aktif mengajar sebagai dosen di STMIK Nusa Mandiri Jakarta

Ibnu Alfarobi Lahir di Brebes 01 Juli 1989, Lulusan S2 Ilmu Komputer STMIK Nusa Mandiri, saat ini aktif mengajar sebagai dosen di STMIK Nusa Mandiri Jakarta