PERBANDINGAN ALGORITMA C4.5, KNN, DAN NAIVE BAYES UNTUK PENENTUAN MODEL KLASIFIKASI PENANGGUNG JAWAB BSI ENTREPRENEUR CENTER

Fuad Nurhasan[1]; Noer Hikmah[2]; Dwi Yuni Utami[3]

[1]Sistem Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
fuad.fnu@bsi.ac.id

[2]Sistem Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
noer.nhh@bsi.ac.id

[3]Sistem Informasi
Universitas Bina Sarana Informatika
www.bsi.ac.id
dwi.dyu@bsi.ac.id

Abstract - BSI Entrepreneur Center is one of the organizations engaged in entrepreneurship within the Bina Sarana Informatika University with the aim of forming students who want to become entrepreneurs. Currently BSI Entrepreneur Center has had responsibility in each campus of Bina Sarana Informatika University. But the existing human resources have not been able to fulfill the needs as the person in charge of BSI Entrepreneur Center to be placed on each campus of Bina Sarana Informatika University. Therefore, a system is needed to find appropriate human resources to be in charge of the BSI Entrepreneur Center on each campus of the Bina Sarana Informatika University. This study uses primary data as many as 300 records consisting of 12 attributes with the algorithm method C.45, KNN and Naive Bayes to classify employees according to the existing criteria. And the results of this study are suggestions from employees who are eligible to be in charge of BSI Entrepreneurs Center on each campus of the Information Technology Development University with the Naive Bayes method which has a high accuracy of 80%.

Keywords: BSI enterpreneur Center, Naive bayes method, employe

Intisari –BSI Entrepreneur Center adalah salah satu wadah yang bergerak dalam bidang kewirausahaan dilingkungan Universitas Bina Sarana Informatika dengan tujuan membentuk mahasiswa yang ingin menjadi seorang wirausaha. Saat ini BSI Entrepreneur Center telah memiliki penanggung jawab pada masing-masing kampus Universitas Bina Sarana Informatika. Tetapi karyawan yang ada saat ini belum dapat memenuhi kebutuhan sebagai penanggung jawab BSI Entrepreneur Center untuk ditempatkan pada masing-masing kampus Universitas Bina Sarana Informatika. Oleh karena itu, diperlukan suatu sistem untuk menemukan sumber daya manusiayang layak untuk menjadi penanggung jawab BSI Entrepreneur Center pada masing-masing kampus Universitas Bina Sarana Informatika. Penelitian ini menggunakan data primer sebanyak 300 record yang terdiri dari 12 atribut dengan metode algoritma C.45, KNN dan Naive Bayes untuk mengklasifikasikan karyawan yang sesuai dengan kriteria yang ada. Dan hasil dari penelitian ini adalah saran dari karyawan yang layak menjadipenanggung jawab BSI Entrepreneur Center pada masing-masing kampus Universitas Bina Sarana Informatika dengan metode Naive Bayes yang memiliki akurasi tinggi 80%.

Kata Kunci: BSI Enterpreneur Center, metode naive bayes, karyawan

PENDAHULUAN

BSI Entrepreneur Center adalah lembaga yang berdiri dibawah naungan dari Universitas Bina Sarana Informatika. BSI entrepreneur center merupakan lembaga yang dijadikan ikon oleh Universitas Bina Sarana Informatika yang bergerak dalam bidang kewirausahaan dengan tujuan untuk menciptakan bibit wirausaha muda dilingkungan Universitas Bina Sarana Informatika. Kewirausahaan merupakan salah satu jenis pelatihan yang sangat berguna bagi siswa untuk mengembangkan jiwa kewirausahaan(Jusmin, 2012) di Universitas Bina Sarana Informatika. Saat ini Universitas Bina Sarana Informatika mempunyai lebih dari 1000 karyawan. Karyawan merupakan salah satu sumber daya yang di gunakan sebagai alat penggerak dalam memajukan suatu perusahaan (Safitri, Waruwu, & Mesran, 2017). Akan tetapi dengan jumlah karyawan di Universitas Bina Sarana Informatika yang banyak tetap tidak mudah untuk menemukan karyawan yang mempunyai jiwa kewirausahaan untuk dijadikan sebagai penanggung jawab BSI Entrepreneur Center pada masing-masing kampus universitas Bina Sarana Informatika.Masalah ini terjadi dikarenakan beranekaragamnya Latar Belakang pendidikan karyawan yang ada dan minat dari karyawan sendiri yang belum mempunyai jiwa kewirausahaan. Minat dapat didefinisikan sebagai sesuatu yang membangkitkan perhatian pada suatu hal (Aprilianty, 2012). Dari permasalahan tersebut perlu diadakan klasifikasi karyawan berdasarkan kondite dan minatnya. Kondite berfungsi untuk melihat kinerja karyawan sebelumnya dan minat karyawan digunakan sebagai tolak ukur kelayakan untuk menjadi penanggungjawab BSI Entrepreneur Center. Untuk pengolahan data didapat dari data karyawan bersumber dari staf administrasi BSI Entrepreneur Center yang akan dilakukan perbandingan metode klasifikasi data mining menggunakan 3 algoritma yaitu C4.5, Naive Bayes, dan KNN (*k-Nearest Neighbor)* dengan pengujian cross validatiaon dan uji T-Test. Sehingga pola yang dihasilkan dalam penelitian ini bisa digunakan untuk mencari sumber daya manusia yang memenuhi persyaratan untuk ditempatkan menjadi penanggung jawab BSI enterpreneur center pada masing-masing kampus Universitas Bina Sarana Informatika.

BAHAN DAN METODE

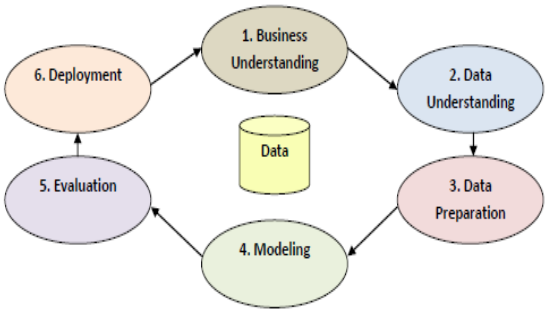
Data Mining

Pengolahan data yang dilakukan pada penelitian ini termasuk kedalam data mining. Menurut Larose dalam (Nuswantoro, 2009), data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat di lakukan, yaitu :

a. Deskripsi
   Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data.
   b. Estimasi Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori.
c. Prediksi
   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang
   d. Klasifikasi
   Dalam klasifikasi, terdapat target variabel kategori.
e. Pengklusteran
   Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (*similarity*) antara satu data dengan data yang lain. Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*).
f. Asosiasi
   Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

CRISP-DM (*Cross- Industry Standard Process for Data Mining*)

Menurut Larose dalam (Nuraeni, 2017) Data mining adalah sebuah proses, sehingga dalam melakukan prosesnya harus sesuai dengan prosedur CRISP-DM (*Cross- Industry Standard Process for Data Mining*). CRISP-DMadalah standarisasi data mining yang disusun oleh tiga pengagas *data mining market* yaitu Daimler Chrysler, SPSS, NCR(Budiman, 2012). CRISP-DM tidak menentukan standar atau karakteristik tertentu karena setiap data yang akan dianalisis akan diproses kembali pada fase-fase di dalamnya(Imtiyaz, Nasrun, & Ahmad, 2015). Berikut gambar yang menjelaskan tentang siklus hidup CRISP-DM(*Cross Industry Standar Proces for data mining*)



Sumber : (North, 2012)

Gambar 1. Konsep Model CRISP-DM

Proses *data mining* berdasarkan CRIPS-DM terdiri dari enam fase sebagai berikut (North, 2012) :

1. *Business Understanding*
   Pada tahapan pertama ini harus didefinisikan apa pengetahuan yang ingin didapatkan dalam bentuk pertanyaan-pertanyaan yang sifatnya umum, misalnya bagaimana cara meningkatkan keuntungan, bagaimana cara mengantisipasi kesalahan cacat produk, dan sebagainya.
2. *Data Understanding*
   Tahapan kedua ini bertujuan untuk mengumpulkan, mengidentifikasikan, dan memahami aset data yang kita miliki. Data tersebut juga harus dapat diverifikasi kebenaran dan realibilitasnya.
3. *Data Preparation*
   Tahapan ini meliputi banyak kegiatan, seperti membersihkan data, memformat ulang data, mengurangi jumlah data, dan sebagainya yang bertujuan untuk menyiapkan data agar konsisten sesuai format yang dibutuhkan.
4. *Modelling*
   Model adalah representasi komputasi dari hasil pengamatan yang merupakan hasil dari pencarian dan identifikasi pola-pola yang terkandung pada data.
5. *Evaluation*
   Evaluasi bertujuan untuk menentukan nilai kegunaan dari model yang telah berhasil kita buat pada langkah sebelumnya.
6. *Deployment*
   Pada tahap ini, hasil yang diperoleh dari seluruh tahapan sebelumnya digunakan secara nyata.

Dalam penelitian ini terdapat 3 algoritma metode klasifikasi data mining yaitu C4.5, Naive Bayes, KNN (*k-Nearest Neighbor).*

Algoritma C4.5

Tahapan dalam membuat pohon keputusan dengan algoritma C4.5 (Gorunescu, 2011) dalam (Wisti Dwi Septiani, 2017) yaitu:

1. Mempersiapkan data training, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon dengan menghitung nilai gain yang tertinggi dari masing-masng atribut atau berdasarkan nilai index entropy terendah. Sebelumnya dihitung terlebih dahulu nilai index entropy, dengan rumus:

$$Entropy\ (i) = -\sum_{j=1}^{m} f(i,j).log\ 2\ f[(i,j)] \ldots\ldots (1)$$

3. Hitung nilai gain dengan rumus:

$$Entropy\ split = \sum_{i=1}^{p} \binom{n1}{n}.IE(i) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

4. Ulangi langkah ke-2 hingga semua record terpartisi. Proses partisi pohon keputusan akan berhenti disaat:
   a. Semua tupel dalam record dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut dalam record yang dipartisi lagi. c. Tidak ada record di dalam cabang yang kosong.

Algoritma Naive Bayes

Naive Bayes menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary,* yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi (Hamzah, 2012).

Algoritma KNN(*k-Nearest Neighbor)*

KNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing (Wu, 2009) dalam (Leidiyana, 2013).

Untuk pengujian model penelitian ini menggunakan Cross Validation dan uji T-Test.

Cross Validation

Pengujian cross validation dilakukan untuk mengetahui kekonsistenan kinerja sistem klasifikasi dengan metode ekstraksi ciri paling invariant terhadap rotasi. Selain itu, pengujian ini juga dilakukan untuk mengetahui pengaruh variasi

data latih pada kinerja sistem klasifikasi. (Kurniawardhani, Suciati, & Arieshanti, 2014)

Uji T-Test

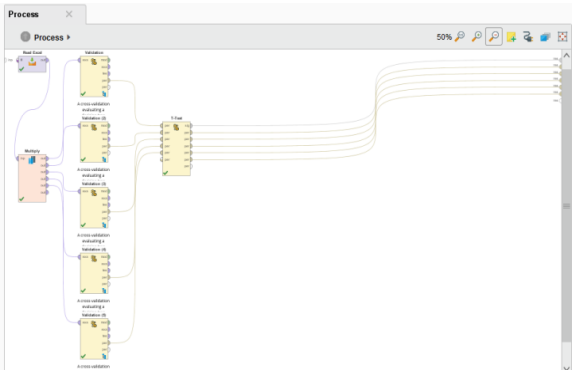
Metode T-Test adalah metode pengujian hipotesis dengan menggunakan satu individu (objek penelitian) dengan menggunakan dua perlakuan yang berbeda. Walaupun dengan menggunakan objek yang sama tetapi sampel tetap terbagi menjadi dua yaitu data dengan perlakuan pertama dan data dengan perlakuan kedua. Performance dapat diketahui dengan cara membandingkan kondisi objek penelitian pertama dan kondisi objek pada penelitian kedua (Hastuti, 2012).

Pengumpulan Data

Adapun data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh dari bagian administrasi BSI enterpreneur center. Data yang digunakan sebanyak 300 record dan terdiri dari 12 atribut. Data yang dikumpulkan adalah data yang dikumpulkan secara langsung oleh bagian administrasi BSI enterpreneur center.

HASIL DAN PEMBAHASAN

Software yang digunakan sebagai alat bantu untuk menghitung tingkat akurasi adalah Rapid Miner. Rapid Miner digunakan untuk membantu menemukan pola yang akurat (Priyana, 2015). Adapun desain menggunakan rapid miner adalah dibawah ini :



Sumber: (Nurhasan, Hikmah, & Utami, 2018)

Gambar 2. Desain Rapid Miner

Setelah dilakukan tahap desain dan pengujian maka menghasilkan performance dari masing-masing Cross Validation, yaitu:

a. Performance Vector 1 dengan algoritma C4.5

Tabel 1 .Performance Vector 1 dengan Algoritma C4.5

Accurancy : 73.33%			
	true pj	true bpj	class precision
pred.pj	18	6	75.00%
pred.bpj	2	4	66.67%
class recal	90.00%	40.00%	

Sumber: (Nurhasan, Hikmah, & Utami, 2018)

Keterangan :
Hasil performance vector dengan algoritma C4.5 menunjukkan tingkat akurasi sebesar 73,33%. Prediksi PJ kenyataan PJ = 18. Prediksi PJ kenyataan Bukan PJ = 6. Prediksi Bukan PJ kenyataan PJ = 2. Prediksi Bukan PJ kenyataan Bukan PJ = 4. Angka AUC sebesar : 0,625.

b. Performance Vector 2 dengan algoritma Naive Bayes

Tabel 2. Performance Vector 2 dengan algoritma Naive Bayes

Accurancy : 80.00%			
	true pj	tru bpj	class precison
pred.pj	17	3	85.00%
pred.bpj	3	7	70.00%
class recal	85.00%	70.00%	

Sumber: (Nurhasan, Hikmah, & Utami, 2018)

Keterangan :
Hasil performance vector dengan algoritma Naive Bayes menunjukkan tingkat akurasi sebesar 80%. Prediksi PJ kenyataan PJ = 17. Prediksi PJ kenyataan Bukan PJ = 3. Prediksi Bukan PJ kenyataan PJ = 3. Prediksi Bukan PJ kenyataan Bukan PJ = 7. Angka AUC sebesar : 0,800.

c. Performance Vector 3 dengan Algoritma KNN

Tabel 3. Performance Vector3 dengan algoritma KNN

Accurancy : 70.00%			
	true pj	tru bpj	class precision
pred.pj	16	5	76,19%
pred.bpj	4	5	55,56%
class recal	80.00%	50.00%	

Sumber: (Nurhasan, Hikmah, & Utami, 2018)

Keterangan :
Hasil performance vector dengan algoritma KNN menunjukkan tingkat akurasi sebesar 70%. Prediksi PJ kenyataan PJ = 16. Prediksi PJ kenyataan Bukan PJ = 5. Prediksi Bukan PJ

kenyataan PJ = 4. Prediksi Bukan PJ kenyataan Bukan PJ = 5. Angka AUC sebesar : 0,500.

Untuk tabel T-Test ditunjukkan seperti gambar dibawah ini :

Tabel 4. T-Test

	C4.5	NB	KNN
C4.5		0425	0,736
NB			0,281
KNN			

Sumber: (Nurhasan, Hikmah, & Utami, 2018)

KESIMPULAN

Dalam penelitian ini memperoleh metode mana yang tepat dengan melihat nilai akurasi yang lebih tinggi, dengan menggunakan metode C4.5 mempunyai nilai akurasi 73,33 % dan metode KNN mempunyai nilai akurasi 70 % dan metode naive bayes mempunyai nilai akurasi sebesar 80 %. Sehingga dari ketiga metode tersebut maka diperoleh algoritma yang paling tepat yang digunakan untuk klasifikasi menjadi penanggung jawab BSI enterpreneur center pada masing-masing kampus Universitas Bina Sarana Informatika yaitu menggunakan metode naive bayes dengan menghasilkan nilai akurasi yang paling tinggi.

REFERENSI

Aprilianty, E. (2012). Pengaruh Kepribadian Wirausaha, Pengetahuan Kewirausahaan, dan Lingkungan Terhadap Minat Berwirausaha Siswa SMK. *Pengaruh Kepribadian Wirausaha, Pengetahuan Kewirausahaan, Dan Lingkungan Terhadap Minat Berwirausaha Siswa SMK*, *2*(3), 311–324. https://doi.org/10.1007/s11365-012-0246-x

Budiman, I. (2012). *DATA CLUSTERING MENGGUNAKAN METODOLOGI CRISP-DM UNTUK PENGENALAN POLA PROPORSI PELAKSANAAN TRIDHARMA*. Universitas Diponegoro. Retrieved from http://eprints.undip.ac.id/36029/

Hamzah, A. (2012). KLASIFIKASI TEKS DENGAN NAÏVE BAYES CLASSIFIER (NBC) UNTUK PENGELOMPOKAN TEKS BERITA DAN ABSTRACT AKADEMIS. In *Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III* (pp. 269–277). https://doi.org/10.1016/j.comcom.2003.09.001

Hastuti, K. (2012). Analisis komparasi algoritma klasifikasi data mining untuk prediksi mahasiswa non aktif, *2012*(Semantik), 241–249.

Imtiyaz, M. Z., Nasrun, M., & Ahmad, U. A. (2015). ANALISIS DAN IMPLEMENTASI FRAMEWORK CRISP-DM UNTUK MENGETAHUI PERILAKU DATA TRANSAKSI PELANGGAN, *2*(1), 596–602.

Jusmin, E. (2012). Pengaruh latar belakang keluarga, kegiatan praktik di unit produksi sekolah, dan pelaksanaan pembelajaran kewirausahaan terhadap kesiapan berwirausaha siswa smk di kabupaten tanah bumbu. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *21*, 46–59.

Kurniawardhani, A., Suciati, N., & Arieshanti, I. (2014). Klafisikasi Citra Batik Menggunakan Metode Ekstraksi Ciri yang Invariant Terhadap Rotasi. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *12*(2), 48. https://doi.org/10.12962/j24068535.v12i2.a322

Leidiyana, H. (2013). Penerapan algoritma k-nearest neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*, *1*(1), 65–76.

North, M. A. (2012). *Data Mining for the Masses*. *Computer Global Text Project*. Georgia: Global Text Project.

Nuraeni, N. (2017). Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC. *Jurnal Teknik Komputer AMIK BSI*, *3*(1), 9–15.

Nurhasan, F., Hikmah, N., & Utami, D. Y. (2018). *Laporan Akhir Penelitian Mandiri*. Jakarta.

Nuswantoro, D. (2009). DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS (pp. 1–11).

Priyana, F. A. (2015). DATA MINING ASOSIASI UNTUK MENENTUKAN CROSS-SELLING PRODUK MENGGUNAKAN ALGORITMA FREQUENT PATTERN-GROWTH PADA KOPERASI KARYAWAN PT. PHAPROS SEMARANG FRISMADANI (pp. 0–1).

Safitri, K., Waruwu, F. T., & Mesran. (2017).

BERPRESTASI DENGAN MENGGUNAKAN METODE ANALYTICAL HIEARARCHY PROCESS (Studi Kasus : PT . Capella Dinamik Nusantara Takengon). *Issn 2548-8368*, *1*(1), 17–21.

Wisti Dwi Septiani. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis, *13*(1), 76–84.