INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

PERBANDINGAN ALGORITMA C 4.5 DAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PEKERJA MIGRAN INDONESIA

Saufika Sukmawati[1], Sulastri [2], Herny Februariyanti[3], Arief Jananto[4]
[1,2,3,4]Universitas Stikubank Semarang
Jl. Tri Lomba Juang No. 1 Semarang Kode Pos 50241
e-mail: hernyfeb@edu.unisbank.ac.id

ABSTRAK

Upaya pemerintah untuk pelindungan pekerja migran Indonesia dikembangkan sistem komputerisasi tenaga kerja luar negeri oleh Badan Nasional Penempatan dan Perlindungan TKI. Permasalahannya adalah adanya pekerja migran Indonesia (PMI) yang dipulangkan karena permasalahan ketenagakerjaan selama di luar negeri, sehingga dibutuhkan interpretasi pada pola data penempatan PMI yang dapat digunakan memprediksi negara tujuan. Penelitian ini membandingkan dua algoritma klasifikasi yaitu algoritma C 4.5 dan algoritma Naïve Bayes untuk mengetahui pola penempatan PMI dengan menggunakan data penempatan PMI wilayah BP3TKI Semarang. Algoritma Naïve Bayes digunakan untuk mengklasifikasikan data PMI dengan menghitung probabilitas dari data training dan data testing. Algoritma C4.5 digunakan untuk memprediksi dengan mengolah variabel usia, gender, pendidikan, staus_perkawinan, pendidikan, negara_tujuan, status_PMI, sektor_pekerjaan. *Percobaan dilakukan dengan data training 1802 dan data testing 772, menghasilkan nilai akurasi paling tinggi bagi kedua algoritma. Algoritma C 4.5 mampu memprediksi lebih baik dengan tingkat akurasi sebesar 84.84% sedangkan Algoritma Naïve Bayes menghasilkan nilai akurasi sebesar 58.29%.*

Kata kunci : *algoritma C 4.5, algoritma naïve bayes, klasifikasi, Pekerja Migran Indonesia*

ABSTRACT

The government's efforts to protect Indonesian migrant workers were developed by a computerized system for overseas workers by the National Agency for the Placement and Protection of Indonesian Migrant Workers. The problem is that there are Indonesian migrant workers (PMI) who were sent home due to labor problems while abroad, So we need an interpretation of the PMI placement data pattern that can be used as a prediction of the destination country. This study compares two classification algorithms, namely the C 4.5 algorithm and the Naïve Bayes algorithm to determine PMI placement patterns using PMI placement data in the BP3TKI Semarang area. The naïv Bayes algorithm is used to classify PMI data by calculating probabilities from training data and testing data. The C4.5 algorithm is used to predict by processing the variables of age, gender, education, marital status, education, destination_country, PMI_status, and employment_sector. The experiment was carried out with 1802 training data and 772 testing data, resulting in the highest accuracy values for both algorithms. The C 4.5 algorithm is able to predict better with an accuracy rate of 84.84% while the Naïve Bayes algorithm produces an accuracy value of 58.29%.

Keywords: *C4.5 algorithm, naïve Bayes algorithm, classification, Indonesian Migrant Workers*

1. PENDAHULUAN

Adanya layanan penempatan kerja ke luar negeri merupakan upaya pemerintah dalam mewujudkan hak masyarakat untuk mendapatkan kesempatan bekerja serta meningkatkan perekonomian negara dengan adanya transfer uang yang dilakukan pekerja migran untuk keluarganya di Indonesia.

Organisasi pemerintahan Badan Nasional Penempatan dan Perlindungan TKI (BNP2TKI) mengembangkan *e-government* dibidang penempatan dan perlindungan pekerja migran

salah satunya melalui sistem pendataan calon Pekerja Migran Indonesia (PMI) yang akan bekerja keluar negeri.

Permasalahan yang menjadi kendala adanya PMI yang dipulangkan atau mendapat permasalahan ketenagakerjaan selama diluar negeri. Sehingga dibutuhkan sebuah interpretasi pada pola data penempatan PMI yang dapat digunakan sebagai prediksi negara tujuan penempatan para calon PMI yang ingin bekerja ke luar negeri sebagai upaya pencegahan.

Proses identifikasi pola data penempatan pekerja migran dapat dilakukan dengan menerapkan data mining. (Roiger, 2017) menyatakan data mining digunakan untuk tujuan analisis berbagai jenis data dengan menggunakan tools data mining yang tersedia. Sedangkan (Kurniawan, 2018) menyebutkan bahwa konsep dari klasifikasi data dalam data mining adalah data – data yang mempunyai kemiripan struktur data akan memiliki klasifikasi yang sama kemudian dapat membentuk suatu aturan atau rule dari suatu data.

Pada penelitian ini data yang akan diteliti adalah data pekerja migran yang melakukan proses penempatan di wilayah BP3TKI Semarang dengan skema penempatan melalui badan hukum yang telah memiliki izin pemerintah dalam penyelenggaraan penempatan PMI di luar negeri. Data tersebut akan diklasifikasikan berdasarkan negara penempatan. Dari latar belakang yang telah disebutkan sebelumnya bahwa penelitian ini akan dilakukan klasifikasi dengan cara membandingkan kinerja antara Algoritma C4.5 dan Naïve Bayes. Diharapkan adanya hasil dari penelitian ini dapat digunakan sebagai pertimbangan bagi *stake holder* terkait dalam membuat kebijakan strategis perihal penempatan calon pekerja migran Indonesia.

a. Klasifikasi

Klasifikasi merupakan salah satu teknik dalam data mining yang dapat digunakan untuk melakukan pemetaan data ke dalam sebuah kelas atau kelompok data yang ditentukan sebelumnya. Klasifikasi juga merupakan metode *supervised learning*, yaitu metode untuk mengahasilkan suatu aturan dalam klasifikasi data uji ke dalam suatu kelas yang telah ditentukan membutuhkan data training yang berlabel. (Dunhan, 2003).

Algoritma untuk melakukan klasifikasi yang sering dipakai adalah *rule-based classifier, decision tree, support machine, neural-network,* serta algoritma *naïve Bayes classifier*. Algoritma klasifikasi menghasilkan pembelajaran untuk identifikasi model yang dapat memberikan relasi yang paling sesuai dan tepat diantara label kelas atau himpunan atribut dari data yang dimasukkan.

Ada 2 cara yang berbeda untuk melakukan klasifikasi teks yaitu dengan cara klasifikasi manual atau klasifikasi secara otomatis. Klasifikasi dengan cara manual, yaitu orang akan membuat catatan-catatan selanjutnya akan ditafsirkan konten teks tersebut dan selanjutnya dilakukan kategorisasi. Cara manual ini akan memberikan memberikan hasil klasifikasi teks berkualitas, akan tetapi sangat membutuhkan waktu dan biaya yang mahal. Sedangkan teknik klasifikasi secara otomatis dilakukan dengan cara penerapan pemrosesan bahasa alami, pembelajaran mesin, dan teknik lain dengan cara otomatis untuk proses klasifikasi teks yang mana teknik yang digunakan lebih cepat serta menghemat biaya. Beberapa pendekatan yang dapat dilakukan untuk teknik klasifikasi teks secara otomatis, yaitu :

1. *Machine Learning Based Systems*

Teknik klasifikasi teks dengan menggunakan *machine learning* akan dibuat klasifikasi dengan cara pengamatan yang dilakukan sebelumnya. Dengan cara memberikan label sebelumnya sebagai contoh data latih. Klasifikasi dengan *machine learning* dapat dipelajari asosiasi yang beda antara bagian teks dan output tertentu (tag) yang dipakai sebagai input tertentu (teks).

Langkah awal dalam melatih klasifikasi menggunakan pembelajaran mesin yaitu dengan ekstraksi fitur: menggunakan metode untuk merubah tiap dokumen menjadi representasi angka (numerik) menjadi vektor. Seringkali digunakan pendekatan *bag of words*, dimana digunakan vektor yang akan mewakili frekuensi suatu kata di dalam kamus kata yang telah ditentukan sebelumnya.

Setelah dilakukan proses data latih dengan menggunakan contoh data latih yang cukup, teknik *machine learning* dapat mulai dibuat suatu prediksi yang akurat. Untuk mengubah teks yang tidak dapat dilihat dengan menggunakan proses ekstraksi fitur yang sama agar dapat menjadi sekumpulan fitur dengan memasukkan ke dalam teknik klasifikasi agar didapatkan prediksi pada tag.

2. *Rule-Based Systems*

Yaitu pendekatan klasifikasi dengan basis aturan dengan diklasifikasikan teks ke dalam kelompok yang terorganisir dengan digunakan seperangkat aturan bahasa/linguistik buatan

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

tangan. Aturan-Sistem akan diinstruksi dengan aturan-aturan yang diberikan untuk menggunakan elemen teks yang sesuai secara semantik untuk dapat diidentifikasi kategori yang sesuai berdasar konten. Setiap aturan meliputi pola dan kategori yang akan diprediksi.

3. *Hybrid systems*

Teknik Sistem *hybrid* menggunakan cara digabungkannya klasifikasi dasar yang akan dilatih dengan menggunakan *rule-based system* dengan *machine learning*, untuk dapat meningkatkan hasil. Sistem hybrid dapat dengan mudah menyesuaikan dengan cara ditambahkan aturan khusus pada tag yang bentrok dan belum dimodelkan dengan benar oleh sistem klasifikasi.

b. Algoritma C4.5

Algoritma C4.5 adalah salah satu cara untuk memberikan solusi menyelesaikan masalah kasus dalam proses klasifikasi. Output dari algoritma C4.5 berupa *decision tree* sama halnya dengan algoritma klasifikasi yang lain. Suatu *decision tree* merupakan struktur yang dapat digunakan untuk melakukan pembagian kumpulan data dengan ukuran yang sangat mesar menjadi suatu himpunan *record* yang kecil dengan menggunakan aturan keputusan. Dimana setiap himpunan hasil yang mirip antara satu dengan himpunan lainnya. (Han, 1996)

Menurut (Purushottam et al., 2016) Algoritma C4.5 adalah suatu algoritma klasifikasi yang dapat digunakan untuk kontruksi pohon keputusan, yang dapat digunakan untuk membentuk *decision tree* (pengambilan keputusan). Algoritma C4.5 merupakan algoritma yang dikembangkan oleh J. Ross Quinlan, yaitu merupakan pengembangan dari *decision tree* yang sering disebut dengan ID3 (*Iterative Dichotomiser* 3), yaitu kekurangan dalam algoritma ID3 disempurnakan dengan algoritma C4.5.

Sebuah set data yang memiliki beberapa pengamatan dengan *missing value* yaitu suatu *record* dengan nilai variable yang tidak ditemukan, dengan terbatasnya jumlah pengamatan, maka nailai rata-rata dari variabel dapat menggantikan atribut dengan *missing value*. (Santoso & Sekardiana, 2019). Ada beberapa elemen dalam menyelesaian kasus dalam Algoritma C4.5 yaitu :

a. *Entropy*
b. *Gain Entropy* (S) merupakan perkiraan kebutuhan jumlah bit yang dapat melakukan ekstraksi sebuah himpunan atau kelas (positif atau negatif) dari sejumlah data acak pada ruang sampel S.

Entropy merupakan kebutuhan bit untuk menyatakan sebuah kelas. Nilai suatu entropy semakin kecil maka akan semakin besar nilai gain entropy dalam proses ekstraksi sebuah kelas.

Berikut adalah perbedaan dari algoritma C 4.5 dengan algoritma: (Elisa, 2017)

a. *Robust* (tahan) terhadap suatu data *noise*,
b. Dapat menyelesaikan variabel-variabel dengan tipe diskrit maupunpun kontinue,
c. Dapat menyelesaikan variabel-variabel dengan *missing value*,
d. Dapat meringkas/memangkas cabang dari pohon keputusan

Algoritme C4.5 memiliki masukkan berupa *training sample serta samples*. *Training samples* merupakan salah satu contoh data yang telah diuji kebenarannya untuk membangun pohon keputusan (*decision treeI*). *Samples* merupakan *field* dari sebuah data yang digunakan sebagai parameter dalam melakukan proses klasifikasi.

Tahapan dalam proses algoritma C 4.5 menurut (Lakshmi et al., 2013), adalah sebagai berikut

a. Persiapan *training data*.
b. Hitung nilai *entropy*. Nilai *Entropy* digunakan untuk mengukur ketidakpastian suatu data, merupakan perbedaan keputusan terhadap nilai suatu atribut tertentu. Nilai *entropy* semakin tinggi, maka semakin tinggi pula perbedaan keputusan (ketidakpastian). Rumus untuk menghitung nilai *Entopy* adalah :

$$Entropy = -\sum p_i \; x \; \log_2 p_i \, k_i \qquad (1)$$

S merupakan himpunan kasus, sedangkan *pi* merupakan nilai probabilitas yang dihasilkan dari *sum* dibagi total kasus.

c. Hitung nilai *gain*, merupakan tahap dalam memilih sebuah atribut yang akan digunakan dalam memilih atribut data test dari setiap simpul pada pohon keputusan. *Gain* merupakan teknik untuk mengukur pengaruh suatu atribut terhadap sebuah keputusan atau disebut ukuran efektifitas dari suatu variabel dalam proses klasifikasi. Rumus menghitung nailai *Gain* adalah sbb :

$$Gain(S, A) = Entropy(S) - \sum |S_i||S| \; xEntropy(S_i)k_i = 1 \qquad (2)$$

S adalah himpunan khasus, sedangkan A disebut sebagai atribut, |*Si*| merupakan jumlah dari kasus ke-*i*, dan |*S*| disebut sebagai jumlah kasus dalam *S*. Algoritma C 4.5 untuk menentukan variabel mana yang akan dijadikan suatu node dari pohon keputusan dalakukan dengan menghitung nilai *gain*. Jika variabel tersebut memiliki nilai *gain* tertinggi

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

maka akan dijadikan sebagai *node* di pohon keputusan.

d. Rumus menghitung nilai *split info* :

$$SplitInfo(S,A) = -\sum S_j S\,x\ \log_2 S_j Sk_j \quad (3)$$

S merupakan ruang dari suatu *sample*, sedangkan A *merupakan* atribut, dan Sj merupakan jumlah dari *sample* untuk atribut ke-j.

e. Menghitung nilai *gain ratio* dengan rumus sebagai berikut :

$$GainRatio\ (S,A) = Gain\ \ (S,A) Split(S,A) \quad (4)$$

Gain(S,A) diseut sebagai *information gain* yang ada pada atribut *(S,A)*, dimana *A* merupakan suatu atribut, dan *Split (S,A)* merupakan *split information* yang berada pada atribut *(S,A)*.
Sedangkan atribut root (akar) akan dibuat berdasarkan nilai *gain ratio* tertinggi, hingga terbentuk pohon keputusan sebagai *node* 1.

f. Selanjutnya ulang langkah ke-b sampai semua cabang akan memiliki kelas atau himpunan yang sama. Proses percabangan akan berhenti apabila:
   - semua khasus di dalam simpul *n* akan menghasilkan himpunan/kelas yang sama.
   - sudah tidak ada lagi variabel independen (berdiri sendiri) dalam khasus yang akan dipartisi.
   - tidak ada lagi khasus di cabang yang kosong.

c. Algoritma Naïve Bayes

Naïve Bayes Classifier (NBC) adalah salah satu algoritma untuk mengklasifikasikan suatu data. Klasifikasi data, yang merupakan tugas dari data mining adalah melakukan pemetaan (mengklasifikasi) data ke dalam satu atau lebih kelas yang telah didefinisikan sebelumnya(Lorena., 2016). Algoritma *Naïve Bayes Classifier* adalah salah satu algoritma dari teknik *machine learning* dengan menggunkan perhitungan probabilitas dan statistik yang ditemukan oleh Thomas Bayes, yaitu dengan melakukan prediksi probabilitas di masa depan dengan berdasar pada history sebelumnya. Rumus Naïve Bayes dalam pemrograman :

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad (5)$$

Peluang dari kejadian A sebagai B akan ditentukan peluang B pada A, peluang A, serta peluang B. Pada saat implementasi rumus akan diubah menjadi :

$$P(C_i|D) = (P(D|C_i) * P(C_i))/P(D) \quad (6)$$

Alur algoritma *Naive Bayes* dapat disajikan pada gambar 1.

Gambar 1. Alur algoritma Naïve Bayes

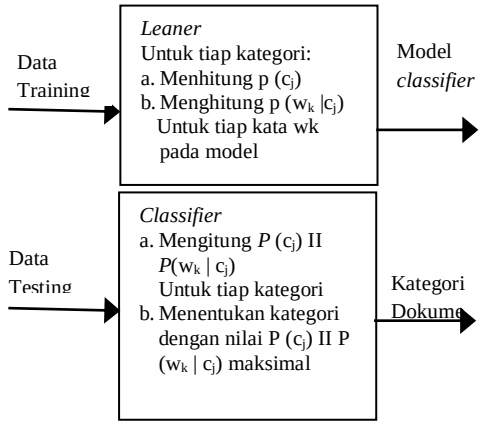
Cara kerja dari algoritma *Naïve Bayes Classifier* adalah dengan dua tahap, yaitu : (Lorena., 2016)

a. Tahap Pembelajaran (Learning)
Naïve Bayes merupakan suatu teknik yang masuk dalam *supervised learning*, maka membutuhkan pengetahuan awal untuk mengambil suatu keputusan dengan langkah sebagai berikut :
Langkah 1 : Bentuk vocabulary disetiap dokumen sebagai data training
Langkah 2 : perhitungan probabilitas disetiap kategori $P(v_j)$.
Langkah 3 : menentukan frekuensi dari setiap kata w_k disetiap kategori $P(w_k|v_j)$

b. Tahap Klasifikasi (*Classify*)
Langkahnya sebagai berikut :

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

Langkah 1 : Lakukan perhitungan $P(v_j)\Pi P(w_k \mid v_j)$ disetiap kategori.
Langkah 2 : menentukan kategori dengan nilai $P(v_j)\Pi P(w_k \mid v_j)$ secara maksimal.



Gambar 2. Proses Klasifikasi *Naïve Bayes Classifier*

2. METODOLOGI PENELITIAN

a. Metode Analisis Data

Analisa data pada penelitian ini dilakukan dengan metode *Knowledge Discovery in Database* atau KDD. Proses KDD penelitian ini dijelaskan sebagai berikut:

1. Data Selection
   Data penempatan PMI yang diperoleh memiliki atribut data yang cukup kompleks. Atribut data yang bersifat bersifat pribadi atau tidak mempengaruhi proses prediksi negara tujuan PMI tidak digunakan dalam proses data mining. Sehingga atribut data yang akan digunakan dapat dilihat pada table berikut:

Tabel 1. Penjelasan atribut dataset penempatan PMI

No	Atribut data	Keterangan
1.	NIK	Nomor Induk Kependudukan PMI
2.	Gender	Jenis kelamin PMI
3.	Pendidikan	Pendidikan terakhir PMI
4.	Status Kawin	Status perkawinan PMI
5.	Negara Tujuan	Negara penempatan PMI
6.	BLK	Tempat Balai Latihan Kerja
7.	TUK	Tempat uji kompetensi
8.	Sektor	Sektor pekerjaan PMI
9.	Tanggal PAP	Tanggal PMI melakukan pembekalan akhir sebelum berangkat ke luar negeri

2. *Data Cleaning*
   Proses *data cleaning* dilakukan dengan memeriksa data yang tidak konsisten dan data yang tidak lengkap kemudian dihilangkan. Setelah melalui proses *cleaning* diperoleh record data sebanyak 2575 dari sebelumnya 2711 data.
3. Transformasi Data
   Proses transformasi data pada data penempatan PMI pada BP3TKI Semarang dilakukan dengan mentransformasi atribut NIK dan TANGGAL PAP menjadi USIA, atribut BLK dan TUK ditransformasikan menjadi STATUS PMI yang akan dijelaskan sebagai berikut:
   a. NIK (Nomor Induk Kependudukan)
      Atribut NIK berisi nomor induk kependudukan milik PMI, sebagaimana dikutip dari website Disdukcapil Pemprov Kalimantan Barat bahwa 16 digit angka pada NIK memiliki arti sebagai berikut:

Gambar 3. Penjelasan struktur NIK

Berdasarkan penjelasan tersebut dalam atribut NIK terdapat informasi tanggal kelahiran PMI sehingga bisa dihitung usia PMI saat berangkat ke luar negeri. Atribut NIK & tanggal PAP dihilangkan kemudian digantikan atribut USIA.

   b. Atribut BLK dan TUK
      BLK atau Balai Latihan Kerja adalah tempat pelatihan untuk mendpatkan ketrampilan atau yang ingin mendalami keahlian dibidang tertentu. Sedangkan

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

TUK adalah tempat pelaksanaan asesmen/uji kompetensi oleh lembaga sertifikasi profesi. Berdasarkan surat edaran kepala BNP2TKI nomor B.205/KA/XI/2015 perihal Registrasi Calon PMI Purna Penempatan bagi PMI yang sudah pernah bekerja di negara yang sama selama minimal 2 tahun dan berada di Indonesia kurang dari 1 tahun maka tidak diwajibkan mengikuti pelatihan kerja di BLKLN dan tidak wajib mengikuti Uji Kompetensi.

Sedangkan bagi PMI sudah pernah bekerja di negara yang sama selama minimal 2 tahun dan berada di Indonesia lebih dari 1 tahun maka tidak diwajibkan mengikuti pelatihan kerja di BLKLN dan namun wajib mengikuti Uji Kompetensi.Berdasarkan hal tersebut akan ditaambah atribut baru dengan nama STATUS PMI dengan ketentuan pengisian sebagai berikut :

Tabel 2 Ketentuan Pengisian Atribut Status PMI

No.	BLK	TUK	Atribut	Ketentuan
1.	Ada	Ada	New	PMI baru yang belum pernah bekerja di negara penempatan
2.	NOBLK	NOLUK	PR1	PMI yang sudah pernah bekerja di negara penempatan minimal 2 tahun dan sebelum bekerja kembali berada di Indonesia selama kurang dari 1 tahun
3.	NOBLK	Ada	PR2	PMI yang sudah pernah bekerja di negara penempatan minimal 2 tahun dan sebelum bekerja kembali berada di Indonesia selama lebih dari 1 tahun

Hasil transformasi data menghasilkan data yang siap diolah terdiri dari 7 (tujuh) atribut sebagaimana terlihat pada tabel berikut:

Tabel 3 Atribut Dataset

No	Atribut data	Keterangan		Penjelasan
1.	Usia	*Predictor*	Usia PMI saat berangkat keluar negeri	Sesuai usia PMI
2.	Gender	*Predictor*	Jenis kelamin PMI	"P" untuk Perempuan "L" untuk Laki – Laki
3.	Pendidikan	*Predictor*	Pendidikan terakhir PMI	SD ; SMP; SMU; SMK, Diploma; Sarjana.
4.	Status Kawin	*Predictor*	Status perkawinan PMI	Kawin; Belum Kawin; Cerai.
5.	Negara Tujuan	*Class*	Negara penempatan PMI	Sesuai negara penempatan PMI
6.	Status PMI	*Predictor*	Status keberangkatan PMI	New (Baru Sektor Informal); PR1 (purna di bawah 1 tahun), PR2 (purna diatas 1 tahun)
7.	Sektor	*Predictor*	Sektor pekerjaan PMI	"Formal" untuk PMI yang bekerja di perusahaan berbadan hukum; "Informal" untuk PMI yang bekerja di perseorangan

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

Tabel 4 Sampel Hasil Transformasi Data

Usia	Gender	Status_Perkawinan	Pendidikan	Negara	Status_pmi	Sector
24	P	SMU	Kawin	Hongkong	New	Informal
25	P	SMU	Kawin	Hongkong	New	Informal
32	P	SMP	Belum Kawin	Hongkong	PR2	Informal
35	P	SMU	Cerai	Hongkong	PR2	Informal
46	P	SD	Cerai	Hongkong	PR1	Informal
34	P	SD	Cerai	Malaysia	New	Informal
42	P	SMP	Kawin	Malaysia	New	Informal
48	P	SMP	Kawin	Brunai Darussalam	New	Informal
48	P	SMP	Kawin	Brunai Darussalam	New	Informal

c. Data Mining
Pada penelitian ini tahap data mining dilakukan dengan mengimplementasikan metode klasifikasi dengan algoritma C.45 dan algortima *Naïve Bayes* menggunakan tools Rstudio yang akan menghasilkan rule klasifikasi dan nilai maximum *posteriory* hypothesis untuk memprediksi negara tujuan penempatan pekerja migran Indonesia.

d. Interpretasi dan Evaluasi
Setelah didapatkan model klasifikasi dari Algoritma C 4.5 dan *Naïve Bayes* langkah selanjutnya adalah menafsirkan model klasifikasi yang terbentuk *rule* dan melakukan evaluasi terhadap pemodelan yang ada dengan *confusion matrix*.

3. HASIL DAN PEMBAHASAN

a. Pengujian Dataset dengan Tools Rstudio

Pengujian menggunakan Rstudio dilakukan sebanyak 3 kali percobaan dengan komposisi *data training* dan *data testing* sebagaimana tabel berikut:

Tabel 5 Jumlah *data training* dan *data testing*

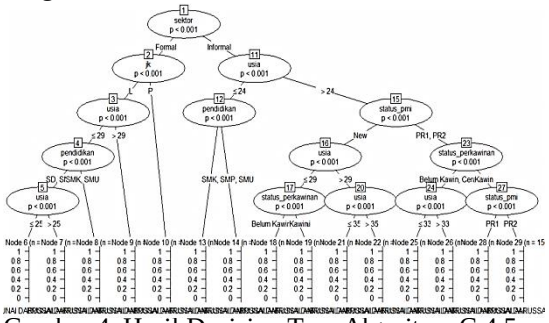
Percobaan ke	Jumlah *Data Training*	Jumlah *Data Testing*
1	1805	774
2	1934	645
3	2063	516

Pada penelitian ini *package* Rstudio yang digunakan akan dijelaskan sebagai berikut:

a. Package "Partykit" yang digunakan untuk mempresentasikan model klasifikasi dengan pohon keputusan yang terstruktur dalam hal ini Algoritma C.45.
b. Package "e1071" digunakan untuk menghitung probabilitas untuk prediksi dalam klasifikasi algoritma *Naive Bayes*.
c. Package "caret" digunakan untk merampingkan proses pembuatan model prediksi *Naïve Bayes*.
d. Package "caTools" adalah package yang digunakan sebagai pembagi data menjadi *data training*. *Data training* tersbut akan digunakan dalam pembuatan model klasifikasi, sedangkan *data testing* adalah data yang akan digunakan untuk melakukan pengujian akurasi dari masing–masing model klasifikasi.

a. Implementasi Dengan Algoritma C.45

Percobaan dilakukan menggunakan total dataset sebanyak 2574 obs data dengan 7 variabel. Jumlah data training sebanyak 1802 obs dan data testing sebanyak 772 obs dengan hasil sebagai berikut:



Gambar 4. Hasil Decision Tree Algoritma C 4.5

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

```
Model formula:
negara ~ sektor + usia + jk + pendidikan + status_perkawinan +
    status_pmi

Fitted party:
[1] root
|   [2] sektor in Formal
|   |   [3] jk in L
|   |   |   [4] usia <= 29
|   |   |   |   [5] pendidikan in SD, SMP
|   |   |   |   |   [6] usia <= 25: BRUNAI DARUSSALAM (n = 47, err = 6.4%)
|   |   |   |   |   [7] usia > 25: MALAYSIA (n = 29, err = 0.0%)
|   |   |   |   [8] pendidikan in SMK, SMU: BRUNAI DARUSSALAM (n = 80, err =
0.0%)
|   |   |   [9] usia > 29: MALAYSIA (n = 99, err = 0.0%)
|   |   [10] jk in P: MALAYSIA (n = 319, err = 0.0%)
|   [11] sektor in Informal
|   |   [12] usia <= 24
|   |   |   [13] pendidikan in SD: MALAYSIA (n = 13, err = 38.5%)
|   |   |   [14] pendidikan in SMK, SMP, SMU: HONGKONG (n = 261, err = 0.0%)
|   |   [15] usia > 24
|   |   |   [16] status_pmi in New
|   |   |   |   [17] usia <= 29
|   |   |   |   |   [18] status_perkawinan in Belum Kawin, Cerai: T A I W A N
(n = 26, err = 3.8%)
|   |   |   |   [19] status_perkawinan in Kawin: T A I W A N (n = 80, err
= 1.2%)
|   |   |   [20] usia > 29
|   |   |   |   [21] usia <= 35: SINGAPURA (n = 148, err = 2.7%)
|   |   |   |   [22] usia > 35: MALAYSIA (n = 214, err = 36.9%)
|   |   |   [23] status_pmi in PR1, PR2
|   |   |   |   [24] status_perkawinan in Belum Kawin, Cerai
|   |   |   |   |   [25] usia <= 33: HONGKONG (n = 90, err = 48.9%)
|   |   |   |   |   [26] usia > 33: T A I W A N (n = 79, err = 11.4%)
|   |   |   |   [27] status_perkawinan in Kawin
|   |   |   |   |   [28] status_pmi in PR1: SINGAPURA (n = 161, err = 15.5%)
|   |   |   |   |   [29] status_pmi in PR2: HONGKONG (n = 156, err = 53.8%)

Number of inner nodes:    14
Number of terminal nodes: 15
```

Gambar 5a. Hasil Rule Algoritma C 4.5

```
Confusion Matrix and Statistics
                    aktual
prediksi            BRUNAI DARUSSALAM HONGKONG MALAYSIA SINGAPURA T A I W A N
  BRUNAI DARUSSALAM                 3        0       21         0           0
  HONGKONG                          0      117        2         0          39
  MALAYSIA                         47        3      158         3           0
  SINGAPURA                        22       43       52       123          23
  T A I W A N                      16       14       22        15          49
Overall Statistics
               Accuracy : 0.5829
                 95% CI : (0.5472, 0.618)
    No Information Rate : 0.3303
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.4621
 Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: BRUNAI DARUSSALAM Class: HONGKONG Class: MALAYSIA
Sensitivity                         0.034091         0.6610          0.6196
Specificity                         0.969298         0.9311          0.8975
Pos Pred Value                      0.125000         0.7405          0.7488
Neg Pred Value                      0.886364         0.9023          0.8271
Prevalence                          0.113990         0.2293          0.3303
Detection Rate                      0.003886         0.1516          0.2047
Detection Prevalence                0.031088         0.2047          0.2733
```

Gambar 5b. Hasil Rule Algoritma C 4.5

Pada gambar 4 dapat dilihat bahwa root node pada model klasifikasi data penempatan PMI adalah atribut SEKTOR. SEKTOR menjadi variabel yang paling berpengaruh ketika seseorang hendak menentukan negara tujuan bekerja ke luar negeri. Sehingga dari hasil decision tree di atas didapatkan aturan atau rule yang dapat dilihat pada gambar 5a dan 5b.

Setelah didapatkan model klasifikasi langkah selanjutnya adalah menguji model klasifikasi tersebut dengan data testing yang telah ditentukan sebelumnya yaitu sebanyak 772 data dengan hasil prediksi seperti terlihat pada gambar 6 berikut:

```
> hasil_tree1= cbind(class_aktual=as.character(testing
30$negara),class_prediksi=as.character(predict_tree1))
> hasil_tree1
     class_aktual       class_prediksi
[1,] "HONGKONG"         "HONGKONG"
[2,] "HONGKONG"         "HONGKONG"
[3,] "HONGKONG"         "HONGKONG"
[4,] "HONGKONG"         "HONGKONG"
[5,] "HONGKONG"         "HONGKONG"
[6,] "HONGKONG"         "HONGKONG"
[7,] "HONGKONG"         "HONGKONG"
[8,] "HONGKONG"         "HONGKONG"
[9,] "HONGKONG"         "HONGKONG"
```

Gambar 6. Hasil Prediksi Algoritma C 45

Setelah menguji data testing selanjutnya melakukan evaluasi model klasifikasi dengan confusion matrix. Setelah dilakukan evaluasi dengan confusion matrix maka ditemukan tingkat akurasi dan tingkat kesalahan dari model klasifikasi yang telah dibuat sebelumnya, seperti terlihat pada gambar 7.

Berdasarkan hasil pengujian 1 di atas dapat dilihat bahwa dari data testing sejumlah 772 data terdapat sebanyak 77 data yang diprediksi tidak benar dan sebanyak 695 data diprediksi dengan benar.

```
Naive Bayes Classifier for Discrete Predictors
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)
A-priori probabilities:
Y
BRUNAI DARUSSALAM          HONGKONG          MALAYSIA          SINGAPURA
T A I W A N
        0.1143174         0.2286349         0.3301887         0.1831299
0.1437292
Conditional probabilities:
                    sektor
Y                       Formal  Informal
  BRUNAI DARUSSALAM 0.6019417 0.3980583
  HONGKONG          0.0000000 1.0000000
  MALAYSIA          0.7563025 0.2436975
  SINGAPURA         0.0000000 1.0000000
  T A I W A N       0.0000000 1.0000000
                    usia
Y                       [,1]     [,2]
  BRUNAI DARUSSALAM 31.93689 9.419280
  HONGKONG          26.18932 5.659699
  MALAYSIA          27.85714 9.401166
  SINGAPURA         34.59394 3.920792
  T A I W A N       32.38996 6.470450
```

Gambar 7. Hasil Accuracy dan Error Rate Percobaan 1

Pada class Brunei Darussalam dari 51 data yang diprediksi benar sebanyak 50 data, pada class Hongkong dari 211 data yang diprediksi benar sebanyak 157 data, pada class Malaysia dari 294 data yang diprediksi benar sebanyak 254 data, pada class Singapura dari 142 data yang diprediksi benar sebanyak 125 data dan pada class Taiwan dari 74 data yang diprediksi benar sebanyak 69 data. Sehingga dari percobaan pertama dapat dilihat bahwa tingkat akurasi dari model klasifikasi ini adalah 85% sedangkan error rate sebesar 15%.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 14 No. 1, Mei 2022
eISSN : 2580-3042
pISSN : 1979-0694

b. Implementasi Algoritma Naïve Bayes

Pembuatan model prediksi dengan algoritma *Naïve Bayes* menggunakan data training sebanyak 1802 dengan hasil pemodelan sebagai berikut:

```
                    jk
Y                        L         P
  BRUNAI DARUSSALAM 0.6019417 0.3980583
  HONGKONG          0.0000000 1.0000000
  MALAYSIA          0.2201681 0.7798319
  SINGAPURA         0.0000000 1.0000000
  T A I W A N       0.0000000 1.0000000
                    pendidikan
Y                         SD        SMK        SMP        SMU
  BRUNAI DARUSSALAM 0.33980583 0.15533981 0.27184466 0.23300971
  HONGKONG          0.12135922 0.06553398 0.51699029 0.29611650
  MALAYSIA          0.21680672 0.26386555 0.33781513 0.18151261
  SINGAPURA         0.37575758 0.01515152 0.49393939 0.11515152
  T A I W A N       0.27799228 0.03861004 0.52123552 0.16216216
                    status_perkawinan
Y                         Belum Kawin BELUM KAWIN       Cerai       Kawin
  BRUNAI DARUSSALAM 0.461165049 0.000000000 0.072815534 0.466019417
  HONGKONG          0.524271845 0.000000000 0.070388350 0.405339806
  MALAYSIA          0.584873950 0.001680672 0.063865546 0.349579832
  SINGAPURA         0.036363636 0.000000000 0.048484848 0.915151515
  T A I W A N       0.185328185 0.000000000 0.351351351 0.463320463
                    status_pmi
Y                          New       PR1       PR2
  BRUNAI DARUSSALAM 1.0000000 0.0000000 0.0000000
  HONGKONG          0.6456311 0.1407767 0.2135922
  MALAYSIA          1.0000000 0.0000000 0.0000000
  SINGAPURA         0.4363636 0.4303030 0.1333333
  T A I W A N       0.4054054 0.4054054 0.1891892
```

Gambar 8. Hasil Model klasifikasi dengan Algoritma Naïve Bayes

Setelah didapatkan probabiltas prior dan probabilitas bersyarat langkah selanjutnya adalah melakukan pengujian model klasifikasi tersebut. Pada pengujian ini data testing yang digunakan sebanyak 772 data dengan hasil prediksi pada gambar 8.

	class_aktual	class_prediksi
[1,]	"HONGKONG"	"HONGKONG"
[2,]	"HONGKONG"	"HONGKONG"
[3,]	"HONGKONG"	"HONGKONG"
[4,]	"HONGKONG"	"SINGAPURA"
[5,]	"HONGKONG"	"SINGAPURA"
[6,]	"HONGKONG"	"SINGAPURA"
[7,]	"HONGKONG"	"SINGAPURA"
[8,]	"HONGKONG"	"SINGAPURA"
[9,]	"HONGKONG"	"SINGAPURA"
[10,]	"HONGKONG"	"SINGAPURA"

Gambar 9. Hasil Prediksi Algoritma Naive Bayes

Pada gambar 9 terlihat bahwa class prediksi dan class aktual bisa menghasilkan nilai yang berbeda, artinya algoritma *Naïve Bayes* artinya beberapa data memang tidak diprediksi dengan tepat oleh pemodelan Naïve Bayes. Maka perlu dilakukan uji tingkat akurasi dengan menggunakan metode *confusion matrix* sehingga ditemukan *error rate* pada hasil prediksi pemodelan ini sebagaimana gambar 10 berikut:

```
Confusion Matrix and Statistics
                 aktual
prediksi          BRUNAI DARUSSALAM HONGKONG MALAYSIA SINGAPURA T A I W A N
  BRUNAI DARUSSALAM                 3        0       21         0         0
  HONGKONG                          0      117        2         0        39
  MALAYSIA                         47        3      158         3         0
  SINGAPURA                        22       43       52       123        23
  T A I W A N                      16       14       22        15        49
Overall Statistics
               Accuracy : 0.5829
                 95% CI : (0.5472, 0.618)
    No Information Rate : 0.3303
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.4621
 Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: BRUNAI DARUSSALAM Class: HONGKONG Class: MALAYSIA
Sensitivity                         0.034091         0.6610          0.6196
Specificity                         0.969298         0.9311          0.8975
Pos Pred Value                      0.125000         0.7405          0.7488
Neg Pred Value                      0.886364         0.9023          0.8271
Prevalence                          0.113990         0.2293          0.3303
Detection Rate                      0.003886         0.1516          0.2047
Detection Prevalence                0.031088         0.2047          0.2733
                     Class: SINGAPURA Class: T A I W A N
Sensitivity                    0.8723            0.44144
Specificity                    0.7781            0.89864
Pos Pred Value                 0.4677            0.42241
Neg Pred Value                 0.9646            0.90549
Prevalence                     0.1826            0.14378
Detection Rate                 0.1593            0.06347
Detection Prevalence           0.3407            0.15026
Balanced Accuracy              0.8252            0.67004
> akurasirb1 <- percent(sum(diag(tabrb1))/sum(tabrb1))
> akurasirb1
[1] "58%"
> errorrb1 <- percent(1-sum(diag(tabrb1))/sum(tabrb1))
> errorrb1
[1] "42%"
```

Gambar 10. Hasil *Confussion Matrix* Algoritma *Naïve Bayes*

Berdasarkan hasil pengujian Algoritma Naïve Bayes di atas dapat dilihat bahwa dari data testing sebanyak 772 data terdapat sebanyak 322 data yang diprediksi tidak benar dan sebanyak 450 data diprediksi dengan benar. Sehingga dari percobaan ini dapat dilihat bahwa tingkat akurasi dari model klasifikasi *Naïve Bayes* sebesar 58% sedangkan error rate sebesar 42%. Selain itu penulis juga melakukan dua kali pengujian algoritma klasifikasi sebelumnya. Pada percobaan kedua dan ketiga dapat kita lihat pada table berikut:

Tabel 6. Perbandingan Hasil Percobaan Algoritma C 4.5 dan *Naïve Bayes*

No	Kegiatan	Jumlah Data Training	Jumlah Data Testing	Tingkat Akurasi (%)		Tingkat Error (%)	
				C 4.5	NB	C 4.5	NB
1	Percobaan I	1805	772	84,84	58,29	15,16	41,71
2	Percobaan II	1934	643	84,60	57,39	15,40	42,61

3	Percobaan III	2063	515	84,47	56,70	43,30	15,53

Jika dilihat dari tabel 6 pengujian pertama dengan Algoritma C 4.5 memiliki tingkat akurasi tertinggi yaitu 84.84% dan presentasi error rate paling rendah yaitu 15.16%. Untuk percobaan dengan Algoritma *Naïve Bayes* pada percobaan pertama juga memiliki nilai akurasi tertinggi yaitu 58.29% dan error rate sebesar 41.71%.

4. KESIMPULAN

Berdasarkan hasil penelitian klasifikasi data penempatan Pekerja Migran Indonesia pada wilayah BP3TKI Semarang menggunakan Algoritma C 4.5 dan Algoritma *Naïve Bayes* dapat disimpulkan bahwa hasil pengujian dengan akurasi tertinggi dihasilkan pada percobaan pertama dengan *data training* sebanyak 1802 data dan *data testing* sebanyak 772. Agoritma C 4.5 memiliki tingkat akurasi sebesar 84.84% dan presentasi *error rate* sebesar 15.16% sedangkan Algoritma *Naïve Bayes* memiliki nilai akurasi sebesar 58.29% dan error rate sebesar 41.71%.

5. REFERENSI

Dunhan, M. (2003). Data Mining: Introductory and Advanced Topics. Prentice Hall. *Engineering*, 1–89. http://www.general.nsysu.edu.tw/gena/gena02/dm/part2.pdf

Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. *Jurnal Online Informatika*, *2*(1), 36. https://doi.org/10.15575/join.v2i1.71

Han, J. (1996). Data mining techniques. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *25*(2), 545. https://doi.org/10.1145/235968.280351

Kurniawan, Y. I. (2018). Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 455. https://doi.org/10.25126/jtiik.201854803

Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. (2013). An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data. *International Journal of Modern Education and Computer Science*, *5*(5), 18–27. https://doi.org/10.5815/ijmecs.2013.05.03

Lorena., S. (2016). Teknik Data Mining Menggunakan Metode Bayes Classifier Untuk Optimalisasi Pencarian Aplikasi Perpustakaan. *Jurnal Teknik Komputer*, *4*(2), 17–20.

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. *Procedia Computer Science*, *85*, 962–969. https://doi.org/10.1016/j.procs.2016.05.288

Roiger, R. J. (2017). Data Mining: A Tutorial-Based Primer. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

Santoso, T. B., & Sekardiana, D. (2019). Penerapan Algoritma C4.5 untuk Penentuan Kelayakan Pemberian Kredit. *Jurnal Algoritma, Logika Dan Komputasi*, *II*(1), 130–137.

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 64–71.