KAJIAN *MACHINE LEARNING* DENGAN KOMPARASI KLASIFIKASI PREDIKSI DATASET TENAGA KERJA NON-AKTIF

Neutrino Sae B. Kusrorong[1], Derwin R. Sina[2], Nelci Dessy Rumlaklak[3]
Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana
Email: kneotrino@yahoo.com[1], derwinilkom@gmail.com[2], dessyrumlaklak@gmail.com[3]

INTISARI

Studi komparasi tehadap pembelajaran mesin dilakukan dengan tujuan untuk menentukan basis metode terbaik berdasarkan kemampuan memprediksikan dengan data benar. Studi dilakukan terhadap dataset tenaga kerja bertujuan untuk mengekstrak informasi pilihan pegawai instansi untuk keluar atau tidak. Metode yang digunakan dalam studi perbandingan yaitu *K-Nearest Neighbors* (KNN) dari basis kemiripan, *Naïve Bayes* (NB) dari basis probabilitas, dan C4.5 dari basis pohon keputusan. Perancangan dan pembangunan aplikasi dilakukan dengan cara menerima input data tenaga kerja, dataset dibagi menjadi data latih dan data uji, data latih untuk pelatihan dan model sedangkan data uji digunakan saat klasifikasi oleh model. Proses klasifikasi dilakukan dengan skenario pelatihan persediaan dan validasi silang terhadap 14.999 data. Hipotesis awal C4.5 adalah metode terbaik dengan tolak ukur akurasi. Pembuktian hipotesis awal akan bernilai benar jika mayortitas akurasi terbaik dimiliki oleh metode C4.5 dengan skenario pelatihan persediaan dan validasi silang. Hasil olahan data klasifikasi menemukan akurasi C4.5 unggul dalam setiap parameter pembagian data skenario pelatihan persediaan dan parameter k-fold 3,5,7, dan 9 dari skenario validasi silang dengan demikian metode terbaik klasifikasi tenaga kerja non-aktif adalah C4.5.

Kata Kunci: Pembelajaran Mesin, Komparasi, K-Nearest Neighbors (KNN), Naïve Bayes, Pohon Keptusan C4.5, dan Tenaga Kerja Non-Aktif.

ABSTRACT

Comparative studies of machine learning are carried out with the aim of determining the best method base based on the ability to predict with true data. The study carried out on the labor dataset aims to extract information on the choice of agency employees to exit or not. The method used in the comparative study is K-Nearest Neighbors (KNN) from the basis of similarity, Naïve Bayes (NB) from the probability base, and C4.5 from the basis of the decision tree. Application design and construction is done by receiving input labor data, the dataset is divided into training data and test data, training data for training and models while the test data is used when classifying by model. The classification process is carried out using supply training scenarios and cross validation of 14,999 data. The initial hypothesis C4.5 is the best method with an accuracy measure. Proof of the initial hypothesis will be true if the best accuracy majority is owned by the C4.5 method with supply trainning scenarios and cross validation. The results of the classification data analysis found that the C4.5 accuracy was superior in each parameter of the inventory training scenario data distribution and the k-fold parameter was 3. 5. 7, and 9 of the cross validation scenario so that the best method of non-active labor classification was C4.5.

Keywords: Machine Learning, Comparison, K-Nearest Neighbors (KNN), Naïve Bayes, Decision tree C4.5, Relevancy, and Non-Active Labor Force.

I. PENDAHULUAN

Semenjak revolusi industri di eropa, manusia telah mengandalkan mesin dalam membantu mengurangi pekerjaan *monotone* atau terus-menerus yang dilakukan manusia. Penemuan mesin pintar telah ada sejak pertengahan abad 20. Pada tahun 1959, Arthur Samuel

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

mendefinisikan bahwa pembelajaran mesin adalah bidang studi yang memberikan mesin kemampuan untuk belajar tanpa diprogram secara eksplisit. Teknologi pembelajaran mesin kemudian berkembang dengan tujuan agar mesin membantu manusia dalam menganalisa kejadian atau fenomena secara sinambung yang terekam dalam bentuk data.

Pengenalan pola adalah tindakan mengambil dataset dan bertindak berdasarkan klasifikasi data yang telah diketahui. Dalam pengenalan pola dengan tujuan klasifikasi terdapat banyak algoritma yang bisa diimplementasikan seperti C4.5, K-NN, Naïve Bayes. C4.5 adalah algoritma yang digunakan untuk membangun pohon keputusan. Algoritma ini dikembangkan oleh Ross Quinlan dalam tujuan mengembangkan sifat sistem heuristik (Quinlan, 1979). K-Nearest Neighbor adalah algoritma dengan prinsip setiap ihwal dalam dataset secara umum memiliki jarak terdekat dengan ihwal lainnya yang memiliki properti yang sama (Cover & Hart, 1967). Dalam pembelajaran mesin, pengelompokan Naïve Bayes adalah pengelompokan probabilistik sederhana berdasarkan penerapan teorema Bayes (Thomas Bayes,1701–1761) dengan asumsi independensi yang kuat (naif) antara fitur.

Dalam survei yang dilakukan Santa Clara County Office of Education tahun 2015 terhadap dari 809 responden 103 pekerja diantaranya mengeluh terhadap lingkungan kerja kantor yang bisa dijadikan alasan untuk berhenti atau pindah kerja meski kompetent dalam bidangnya. Masalah yang muncul adalah tenaga kerja profesional yang memilih untuk keluar dari perusahaan. Beberapa faktor penyebab diantaranya rendahnya gaji yang diterima, kecelakaan kerja, bidang pekerjaan, lamanya waktu kerja dan lainnya. Hal ini mengakibatkan perusahan menjadi kekurangan tenaga kerja kompeten yang dapat mempengaruhi performa dan pendapatan yang diperoleh perusahaan atau instansi. Oleh karena itu, diperlukan suatu analisa prediktif yang membantu pihak manajemen perusahaan untuk bisa mengetahui kecenderungan tenaga kerja yang memilih berhenti. Selain itu analisa ini bertujuan untuk melahirkan informed choice bagi pihak pengambil keputusan untuk mempertahankan tenaga kerja terbaik.

Dalam penelitian ini dilakukan analisis komparasi terhadap tiga algoritma klasifikasi Pembelajaran Mesin yaitu C4.5, K-NN dan Naïve Bayes terhadap dataset tenaga kerja. Tujuan komparasi metode mana yang terbaik terhadap dataset tenaga kerja. Selain itu untuk menghasilkan informasi berharga bagi pihak management untuk bisa mengambil langkah tindakan selanjutnya.

II. LANDASAN TEORI

2.1 *Machine Learning*

Samuel Arthur mendefinisikan bahwa pembelajaran mesin adalah bidang studi yang memberikan mesin kemampuan untuk belajar tanpa diprogram secara eksplisit. Dalam pengembangannya Samuel memiliki ide bahwa mengajar komputer untuk bermain game sangat bermanfaat untuk mengembangkan taktik yang sesuai. Samuel memilih dam karena relatif sederhana namun memiliki kedalaman strategi. Dalam keterbasan teknologi pada awal 50-an mengakibatkan memory yang tersedia yang relatif rendah hingga Samuel mengembangkan metode pencarian Alpha-Beta Pruning untuk mengakomodasi teknologi waktu itu. Selanjutnya Samuel melatih mesin tersebut ratusan kali hingga akhirnya mesin tersebut bisa mencapai level pemula dalam turnamen dam. Salah satu bagian dalam pembelajaran mesin adalah Supervised Learning atau pembelajaran terawasi. Dalam pembangunan basis pengetahuan pada pembelajaran mesin, Supervised Learning adalah bentuk tugas mesin untuk menyimpulkan sebuah fungsi dari data pelatihan berlabel (Mohri, Rostamizadeh, & Talwalkar, 2012). Dalam prosesnya mesin akan menerima informasi berupa ihwal yang sebelumnya telah terjadi dan membangun kemampuan nalar layaknya manusia belajar yang dari pengalaman. Menurut (Hamakonda, 1991) klasifikasi adalah pengelompokan yang sistematis dari obyek, gagasan, buku atau benda-benda lain ke dalam kelas atau golongan tertentu berdasarkan ciri-ciri yang sama.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

2.2 Metode Klasifikasi

Klasifikasi bisa dikatakan sebagai pembelajaran mesin karena memiliki kemampuan untuk mengunakan pengetahuan yang telah ada sebelumnya untuk menghasilkan penentuan objek baru. Keseluruhan klasifikasi terletak pada kemampuan sistem untuk memberi label terhadap objek sesuai dengan kasus yang telah ada tanpa mengubah sistem jika dihadapkan dengan objek baru.

2.2.1 *K-nearest neighbors* (KNN)

K-Nearest Neighbor atau K-Tetangga Terdekat adalah algoritma dengan prinsip setiap ihwal dalam dataset secara umum memiliki jarak terdekat dengan ihwal lainnya yang memiliki properti yang sama (Cover & Hart, 1967).

$$KNN(x,Sy) = arg\min(\sum_{i=0}^{k-1} Nearest(x,Sy) \quad)$$

Keterangan:

- Nearest = Fungsi kelas terdekat
- k = parameter konstan
- x = Data baru
- S_y = Himpunan data latih
- argmin = Vote data terdekat

K disini mengartikan majemuk atau majoritas, dimana untuk setiap ihwalan baru akan dilabelkan berdasarkan jumlah terbanyak voting ihwal yang telah ada. Fungsi nearest didapatkan dengan menbandingkan jarak kemiripan data baru dengan semua data dalam himpunan data latih seperti persamaan berikut:

$$Nearest(x,Sy) = \min(\sum_{i=0}^{Sn-1} d(x,y_i) \quad)$$

Keterangan:

- Nearest = Fungsi kelas terdekat
- S_n = jumlah data latih
- d = Jarak/*distance*
- x = Data baru
- y = Data pembanding
- S_y = Himpunan data latih

Jarak antar data dihitung dengan fungsi jarak Euclidean.

$$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=0}^{fn-1} (fx_i - fy_i)^2}$$

Keterangan:

- d = jarak/*distance*
- x = data baru
- y = data pembanding
- f_x = fitur data baru
- f_y = fitur data pembanding
- f_n = jumlah fitur data

2.2.2 *Naïve bayes* (NB)

Naïve Bayes merupakan sebuah metode klasifikasi probabilistik sederhana untuk menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan. Algoritma mengunakan teorema *Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Secara sederhana naïve bayes di rumuskan sebagai *conditional probability* sebagai berikut:

$$Posterior = \frac{Prior * likelihood}{evidence}$$

Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas mana yang cocok bagi sampel yang dianalisis tersebut. Oleh karena itu, persamaan metode *Naive Bayes* ditulis sebagai berikut:

$$P(C|E) = \frac{P(E|C) * P(C)}{P(E)}$$

Keterangan**:**

- E = *Evidence* atau bukti data yang ada

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

- C = Asumsi objek dengan *class* yang spesifik
- P(E|C) =Probabilitas E berdasar C (*Likelihood*)
- P(C|E) =Probabilitas C berdasarkan kondisi E (Posterior)
- P(E) =Probabilitas E tanpa diketahui C *(Evidence*)
- P(C) =Probabilitas objek *class* adalah benar (*Prior*)

Untuk menjelaskan teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah bukti untuk menentukan kelas dimana yang cocok bagi sampel yang dianalisis tersebut. Oleh karena itu, teorema *bayes* di atas disesuaikan untuk E sebagai himpunan bukti sebagai berikut:

$$P(C|F_1 \dots F_n) = \frac{P(C) * P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

Dimana variabel C merepresentasikan kelas, sementara variabel $F_1 \dots F_n$ merepresentasikan karakteristik bukti yang dibutuhkan untuk melakukan klasifikasi. Nilai probabilitas atribut F dari persamaan diatas didapatkan dengan mengunakan fungsi probabilitas *laplacian smoothing* dengan nilai K=1 untuk untuk menghindai probablitas bernilai nol, persamaan probablitas *laplacian smoothing* bisa dilihat pada persamaan berikut:

$$P(\mathrm{F}) = \frac{Count + K}{N + (K * Z)}$$

Keterangan:

- P = Probabilitas dari variabel F
- *Count* = Jumlah kemunculan F
- K = Parameter *smoothing*
- N = Jumlah data
- Z = Jumlah jenis kelas dari sampel

Dalam penggunaan metode *naïve bayes*, sebelum mengetahui hasil akhir perlu dilakukan perhitungan nilai *likelihood*, dengan menggunakan rumus sebagai berikut:

$$\mathrm{P}(E|\mathrm{C}) = P(F_1|C) \times P(F_2|C) \times \dots \times P(F_n|C)$$

Keterangan:

- P = Probabilitas
- E = *Evidance Value*
- C = *Class* objek
- $\mathrm{F_i}$ = Atribut dari E

Normalisasi hasil persamaan diatas akan menghasilkan nilai sederhana yang bisa digunakan untuk menentukan kelas objek baru. Fungsi normalisasi dari persamaan diatas bisa di lihat pada persamaan berikut:

$$P(X) = \frac{Likelihood\ prior}{Likelihood\ prior\ +\ Likelihood\ posterior}$$

Dimana:

- P = Probabilitas
- X = Objek baru
- $Likelihood\ prior$ = Kemungkinan sebelumnya
- $Likelihood\ posterior$ = Kemungkinan selanjutnya

Dengan membandingkan nilai *probabilitas* P(C) maka objek X bisa diklasifikasikan ke kelas spesifik.

2.2.3 *Decision Tree C4.5*(C4.5)

C4.5 adalah algoritma yang digunakan untuk menghasilkan pohon keputusan yang dikembangkan oleh Ross Quinlan. C4.5 merupakan perpanjangan dari algoritma ID3 Quinlan sebelumnya. Pohon keputusan yang dihasilkan oleh C4.5 dapat digunakan untuk klasifikasi, dan untuk alasan ini, C4.5 sering disebut sebagai pengklasifikasi statistik.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Kusrini, 2009):

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(A)$$

Keterangan:

- S = himpunan kasus
- A = atribut
- *n* = jumlah partisi atribut A
- $|S_i|$ = jumlah kasus pada partisi ke-i
- |S| = jumlah kasus dalam S

Sebelum mendapatkan nilai Gain adalah dengan mencari nilai Entropy. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan sebuah atribut. Persamaan untuk mendapatkan Entropy adalah:

$$Entropy(S) = \sum_{i=1}^{n} - Pi * \log_x Pi$$

Keterangan:

- S = himpunan kasus
- n = jumlah partisi S
- P_i = proporsi dari Si terhadap S
- x = Jumlah Label kasus

Dari fungsi diatas perhitungan C4.5 dimulai dengan mencari nilai entropy dari semua data. Nilai entropy digunakan sebagai dasar perhitungan gain tiap atribut, nilai gain tertinggi menjadi akar dari pohon keputusan perhitungan dilakukan hingga semua atribut terdefinisi.

2.3 Kriteria Evaluasi

Evaluasi kinerja dari klasifikasi tenaga kerja diperlukan untuk melihat performa dari sebuah *classifier*. Kriteria dibagi menjadi dua bagian yaitu kriteria komputasi atau perfoma dan kriteria utilitas atau *relevancy*.

2.3.1 Peforma

Peforma sebuah model klasifikasi akan mempengaruhi kegunaan sebagai suatu *tool* analisis. Peforma adalah sifat model klasifikasi terhadap lingkunganya yaitu lingkungan komputasi digital. Klasifikasi adalah suatu perhitungan matematika kompleks dan berulang dalam lingkungan digital, diproses sebagai *task* dalam *processor* komputer, maka akan memiliki angka pengunaan *memory* dan waktu penyelesaian tugas. Nilai *memory* dan waktu tersebut bisa menjelaskan bagaimana suatu model berkerja dalam suatu komputasi. Adapun kriteria peforma model klasifikasi yang didapatkan:

- *Learning Speed / Training time*
- *Classification Speed / Testing time*
- *Learning Memory Use / Training time*
- *Classification Memory Use / Testing Memory*

2.3.2 Relevancy

Di bidang pembelajaran mesin dan khususnya masalah klasifikasi, Confusion Matrix juga dikenal sebagai matriks kesalahan. Confusion Matrix adalah tata letak tabel yang spesifik yang memungkinkan visualisasi kinerja algoritma, biasanya pembelajaran yang diawasi (dalam Pembelajaran tanpa pengawasan biasanya disebut matriks pencocokan). Setiap kolom dari matriks mewakili contoh dalam kelas yang diprediksi sementara setiap baris mewakili instance di kelas sebenarnya (atau sebaliknya). Confusion Matrix bisa merepresentasikan utilitas suatu model klasifikasi karena bisa menunjukkan informasi hasil klasifikasi dan langsung dibandingkan

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

dengan data aktual. Menurut (Olson & Delen, 2008) dasar Confusion Matrix seperti yang terlihat pada selengkapnya pada tabel berikut :

Tabel 1. *Relevancy*

		Predicted Condition			
	Total Population	*Prediction Positive*	*Prediction Negative*	*Prevalence = Σ Condition Positive / Σ Total Population*	
True Condition	*Actual Positive*	*True Positive (TP)*	*False Negative (FN)*	*True Positive Rate (PR), Sensitivity, Recall, Probability of Detection = ΣTP / Σ Condition Positive*	*False Negative Rate (FNR), Miss Rate = ΣFN / Σ Condition Positive*
	Actual Negative	*False Positive (FP)*	*True Negative (TN)*	*False Positive Rate (FPR), Fall-out, Probability of False Alarm = ΣFP / Σ condition negative*	*True Negative Rate (TNR), Specificity (SPC) = ΣTN / Σ Condition Negative*
	Accuracy = (ΣTP +ΣTN) / Σ Total Population	*Positive Predictive Value (PPV), Precision = ΣTP / Σ Prediction Positive*	*False Omission Rate (FOR) = ΣFN / Σ Prediction Negative*	*Positive Likelihood Ratio (LR+) = TPR / FPR*	*Diagnostic Odds Ratio (DOR) = LR+ / LR−*
		False Discovery Rate (FDR) =ΣFP/Σ Prediction Positive	*Negative Predictive Value (NPV) = ΣTN / Σ prediction negative*	*Negative Likelihood Ratio (LR−) = FNR / TNR*	

2.4 Skenario Pengujian

Skenario komparasi bertujuan untuk menguji masing-masing metode terhadap beberapa skenario dengan tujuan untuk mengetahui bagaimana tiap-tiap metode menghadapi perubahan dataset. Dengan mengetahui sifat metode maka kesimpulan penelitian akan lebih presisi.

2.4.1 Skenario Supply Training Test

Skenario sedeherna untuk menguji model klassifikasi dengan membagikan porsi dataset kedalam dua bagian yaitu bagian data latih dan data uji. Tujuan dalam skenario ini adalah untuk menguji model klasifikasi dalam keadaan dataset yang berbeda. Dalam penelitian ini skenario yang akan dijalankan adalah 30% data latih dan 70% data uji, 50% data latih dan 50% data uji, dan 70% data latih dan 30% data uji.

2.4.2 Skenario Cross-Validation

Skenario untuk menilai bagaimana hasil statistik analisis akan digeneralisasi kumpulan data independen. Teknik ini utamanya digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat sebuah model prediktif ketika dijalankan dalam praktiknya. Dalam sebuah masalah prediksi, sebuah model biasanya diberikan kumpulan data (dataset) yang diketahui untuk digunakan dalam menjalankan pelatihan (dataset pelatihan), serta kumpulan data yang tidak diketahui (atau data yang pertama kali dilihat) terhadap model yang diuji (pengujian dataset). Tujuan dari *Cross Validation* adalah untuk mendefinisikan dataset untuk "menguji"

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

model dalam tahap pelatihan (yaitu, validasi data), dalam rangka untuk membatasi masalah seperti terjadinya overfitting.

III. HASIL DAN PEMBAHASAN

3.1 Hasil

Hasil sistem komparasi pembelajaran mesin yang didapati akan dibagi menjadi hasil implementasi antarmuka dan hasil implementasi algoritma. Bagian implementasi antarmuka membahas tampilan sistem dengan memaksimalkan informasi yang ditampilkan kepada penguna. Bagian implementasi algoritma akan menyajikan proses klasifikasi. Alur kinerja sistem secara umum adalah mulai dari menerima input data tenaga kerja dalam jumlah besar yang selanjutnya disebut dataset, dataset dibagi menjadi data latih dan data uji, data latih untuk pelatihan dan pembentukan model sedangkan data uji digunakan saat klasifikasi oleh model. Masing-masing metode yaitu KNN, C4.5 dan *Naïve Bayes* memiliki proses pelatihan yang berdeda dan model prediksi yang berbeda. Proses klasifikasi masing-masing model terhadap data uji akan menghasil keluaran *relevancy* dan pengunaan sumber daya, hasil klasifikasi masing-masing metode akan dibandingkan dan disusun untuk menjadi informasi *output* hasil komparasi.

3.1.2 Hasil Implementasi algoritma

Implementasi analisa komparasi *machine learning* dalam klasifikasi prediksi tenaga kerja non-aktif dibagi tiga algoritma yaitu *k-nearest neightbour, naïve bayes* dan *Decision Tree C4.5*. Dataset yang digunakan telah bersih dari *noise* yang bisa menganggu pembangunan model klasifikasi. Dengan demikian dataset tidak memerlukan manipulasi atau perubahan. Dataset bisa digunakan untuk membangun model prediksi.

A. Hasil Skenario Supply Training

Hasil supply training didapati dengan proses mengklasifikasikan dataset yang telah dibagi menjadi dua bagian yaitu data latih dan dataset uji. Skenario Supply Trainning, dimulai dengan menentukan proporsi data latih dan data uji. Pembagian dimulai dari 50% data latih dan 50% data uji, setelah itu akan ditingkatkan senilai 10% proporsi data latih dan mengurangi 10% proporsi data uji hingga akhirnya pembagian data latih mencapai 100%, dan jika data latih yang digunakan 100% maka data uji juga 100%.

B. Hasil Skenario Cross Validataion

Hasil skenario *Cross Validation* dilakukan dengan membagi data kedalam beberapa himpunana kecil yang secara bergantian akan digunakan sebagai data latih data data uji. Skenario *Cross Validation* dijalankan dengan menggunakan k-fold sama dengan nilai 3, 5, 7, dan 9. Masing-masing subbagian akan digunakan sebagai data latih dan data uji, sebagai contoh jika bagian A digunakan sebagai data uji maka bagian B dan C akan digunakan sebagai data latih. Proses pergantian pengunaan data latih dan data uji terus dilakukan sehingga seluruh bagian akan digunakan sekali sebagai data uji.

Eksekusi skenario *Cross Validation* mengeliminasi hasil algoritma dengan nilai K dengan nilai rata-rata Akurasi yang terendah sehingga mengahasilkan informasi model terbaik metode K-NN dan NB. Nilai performa didapatkan dengan menbandingkan sumber daya saat sebelum klasifikasi dan setelah klasifikasi. Untuk mendapatkan informasi secara umum dalam skenario pengujian *Cross Validation* maka hasil *relevancy* dan performa akan dirata-ratakan.

3.2 Pembahasan

Dalam penelitian ini penulis berhasil membangun sistem untuk komparasi metode klasifikasi. Dataset yang digunakan berjumlah 14.999, dengan jumlah 9 atribut, jumlah 2 kelas, jumlah 2 atribut binary, jumlah 2 atribut kontinous, jumlah 2 atribut katerogy, jumlah 2 atribut numeric. Dataset telah bersih dari noise data anomaly, null ataupun froud.

Dataset yang digunakan adalah dataset Tenaga Kerja berjenis data sekunder yang didapat dari repository onbaris https://www.kaggle.com/colara/hr-analytics. Dataset tenaga kerja terdiri

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

dari dua class target, yaitu 1 untuk berhenti dan 0 untuk tetap, dengan kurun waktu aktif bekerja kurang dari 10 tahun. Dataset memiliki 9 fitur yaitu satisfaction level, last evaluation, number project, average montly hours, time spend company, work accident, promotion last 5 years, division dan salary.

3.2.1 Pembahasan Skenario *Supply Training*

Skenario Supply Trainning, dimulai dengan mementukan proporsi data latih dan data uji. Pembagian dimulai 50%-50%, 60%-40%,70%-30%,80%-20% ,90-10% dan 100%-100%. Data yang dipilih dalam masing-masing pembagian data tidak secara acak dengan tujuan memberikan masing-masing metode untuk membangun model prediksi yang adil. Data penelitian masing-masing pembagian data dihitung kedalam persamaan rata-rata dengan tujuan menemukan informasi secara umum.

Tabel 2 Hasil Komparasi Rata-Rata *Supply Training*

METODE	K-NN	NB	C4.5
P LATIH	70.00	70.00	70.00
P TEST	30.00	30.00	30.00
T. TIME	5.40	2070.80	742.00
T. MEMORY	338385.60	16844608.00	40409515.20
C. MEMORY	127503908.80	105937872.00	121420992.00
C. TIME	15552.60	3218.40	5027.20
N LATIH	10500.00	10500.00	10500.00
N TEST	4499.00	4499.00	4499.00
TP	3284.80	2265.60	3384.80
TN	1022.60	915.80	997.80
FP	70.00	176.80	94.80
FN	121.60	1140.80	21.60
PRE	0.75	0.55	0.77
TPR	0.97	0.67	0.99
TNR	0.94	0.83	0.91
PPV	0.98	0.92	0.97
NPV	0.90	0.45	0.98
FNR	0.10	0.55	0.02
FPR	0.06	0.17	0.09
FDR	0.02	0.08	0.03
FOR	0.10	0.55	0.02
ACC	0.96	0.71	0.97
F1	0.97	0.77	0.98
MCC	0.89	0.43	0.93
BM	0.91	0.50	0.90
DOR	163.78	6.02	519.22
LR +	16.48	4.00	11.15
LR -	0.11	0.67	0.02

LEGENDA

HIJAU = TERBAIK KUNING = SEDANG MERAH = TERBURUK

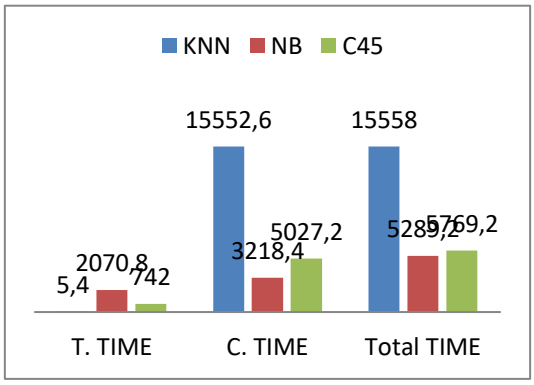
K-NN memiliki kelebihan pada fase training dengan pengunaan waktu dan memory terkecil dibanding dengan dua metode lain (T. Memory dan T. Time). Keunggulan K-NN hanya

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

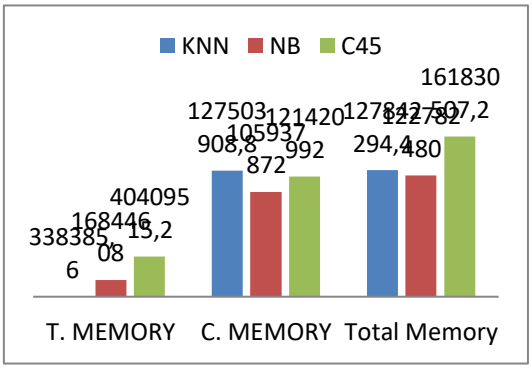
terletak pada fase pelatihan dimana pada fase klasifikasi lebih didominasi oleh NB dan C4.5. Grafik pengunaan sumber daya pada grafik gambar 1 dan 2.

Seperti pada grafik pada gambar 1 dan gambar 2 naïve bayes dengan *laplacian smoothing* menghasilkan metode yang terbaik dalam pengunaan sumber daya waktu dan memori artinya naïve bayes akan lebih optimal pada kasus yang memerlukan kemampuan model realtime yang tidak bisa diimplementasikan dengan mudah oleh metode K-NN dan C4.5.

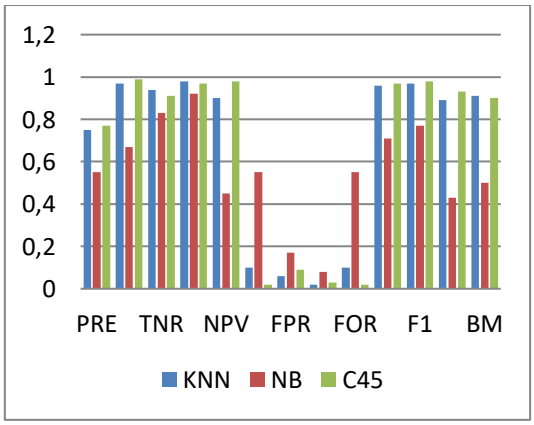
K-NN memiliki keunggulan dalam mengklasifikasikan kelas *left* dengan nilai TN, FP, dan TNR tertinggi. K-NN dengan mudah bisa memembedakan data dengan kelas stay dari data dengan kelas left dengan nilai relevancy FPR dan FDR terbaik. Walaupun tidak memiliki angka DOR tertinggi tetapi K-NN memiliki angka BM tertinggi dibanding dengan metode lainnya yang menunjukan bahwa K-NN menghasilkan klasifikasi lebih baik saat data memiliki kelas *stay* maupun *left*. LR+ tertinggi dari K-NN akibat angka FPR yang rendah artinya K-NN lebih memiliki kelebihan untuk tidak mudah salah dalam klasifikasikan kelas left.



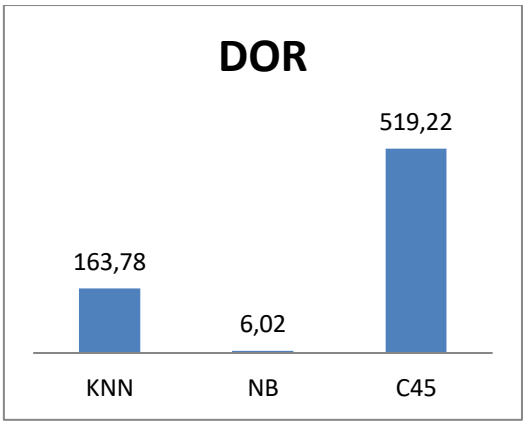
Gambar 1. Grafik Rata-rata Pengunaan Sumber Daya Waktu Supply Training



Gambar 2. Rata-rata Pengunaan Sumber Daya Memory Supply Training



Gambar 3. Rata-rata Relevancy dalam Skenario *Supply Trainning*



Gambar 4. Rata-rata Diagnostic Odd Ratio dalam Skenario *Supply Trainning*

Grafik pada gambar 3 menunjukan kelemahan naïve bayes pada atribut *relevancy* lainnya dimungkinkan asumsi independensi pada setiap atribut pada dataset, *naïve* bayes gagal mentoleransikan relasi antar atribut yang mungkin ada dalam dataset, ini mengakibatkan naïve bayes tidak bisa menghasilkan model probabilistik indepedensi dengan baik. Angka BM yang 0.5 merupakan petunjuk lain untuk keberadaan korelasi data dalam dataset lebih dari apa yang mampu model *naïve* bayes kerjakan. Nilai Akurasi 0.71 dan DOR 6.02 bukan hasil yang buruk dalam kelasifikasi untuk menyatakan metode naïve bayes suatu kegagalan besar seperti yang terlihat pada grafik Diagnostic Odd Ratio pada gambar 4.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

C4.5 menghasilkan model prediksi dengan angka DOR tertinggi. C4.5 unggul dalam menentukan klasifikasi kelas *stay* dengan angka TP dan PRE tertinggi tanpa banyak melakukan kesalahan kelasifikasi dengan FN, FNR, FOR terendah. Keunggulan model klasifikasi C4.5 juga ada pada angka akurasi dengan mengunakan sumber daya memory dan waktu lebih dari K-NN, bahkan C4.5 dapat mengklasifikasi data dengan akurasi tinggi tiga kali lipat lebih cepat dibanding dengan K-NN. BM C4.5 kalah dibandingkan dengan model prediksi K-NN tetapi MCC C4.5 lebih diunggulkan dengan angka paling dekat dengan positive satu yang artinya model C4.5 adalah model prediksi terbaik mendekati sempurna untuk klassifikasi dataset tenaga kerja seperti terlihat pada grafik gambar 3.

Dalam skenario supply training dengan pembagian data 50%-50%, 60%-40%,70%-30%,80%-20% dan 90-10%., C4.5 memang tidak bisa unggul dalam hal pengunaan sumber daya terutama pada fase pelatihan dengan pengunaan memory tertinggi tapi unggul secara umum unggul pada 12 hasil komparasi lainnya terutama pada nilai DOR dan Akurasi dibanding dengan K-NN 10 hasil komparasi dan naïve bayes dengan 2 hasil komparasi. Dengan nilai akurasi tertinggi dimiliki C4.5 maka dukungan hipotesis H0 adalah benar semakin kuat.

Skenario komparasi *Supply Trainning* secara mayoritas akurasi terbaik dimiliki metode C4.5 dengan demikan syarat pertama hipotesis H0 telah dipenuhi.

3.2.1 Pembahasan Skenario Cross Validation

Pada skenario *Cross Validation* dataset akan dibagikan menjadi k bagian yang secara bergiliran masing-masing himpunan data akan digunakan sebagai data uji dan sisanya sebagai data latih. Skenario *Cross Validation* akan memberikan informasi saat model menerima keseluruhan dataset sebagai data latih dan data uji. Hasil dari masing-masing model dengan nilai parameter "k" terbaik yang akan digunakan dalam pembahasan. Dari hasil akan dihitung nilai rata-rata dengan tujuan menemukan informasi model prediksi terbaik secara umum dengan adil tanpa bias. Hasil *relevancy*, pengunaan *memory* dan pengunaan waktu disatukan dan dijadikan acuan komparasi. Tabel voting terdiri dari 16 variable relevansi, total pengunaan memori dan total pengunaan waktu. Tabel 3 adalah tabel voting komparasi skenario *Cross Validation* dengan nilai k-fold sama dengan 3,5,7, dan 9.

Dari tabel 3 ini telah diketahui bahwa mayoritas C4.5 menang dalam *voting* nilai komparasi terbaik dengan nilai *voting* 53, disusul oleh KNN dan NB di posisi terakhir. Hasil *voting* tinggi menunjukan dominasi kecocokan metode C4.5 dalam mengklasifikasikan dataset tenaga kerja. Walaupun dengan kostuminasi parameter konstan k yang dimiliki oleh KNN dan NB masih belum bisa mengubah hasil bahwa metode C4.5 memiliki kompabilitas tinggi dengan tipe dataset tenaga kerja yang memiliki banyak atribut bilangan real dan numeric.

Tabel 3 Hasil Voting Komparasi *Cross Validation*

K fold	NB	C4.5	KNN
3	1	12	5
5	0	18	0
7	0	7	11
9	2	16	0
Total	**3**	**53**	**16**

Tabel 4 Hasil Ranking pengunaan memory skenario *Cross Validation*

K fold	NB	C4.5	KNN
3	2	1	3
5	3	1	2
7	3	2	1
9	1	2	3

Tabel 5 Hasil Ranking pengunaan memory skenario *Cross Validation*

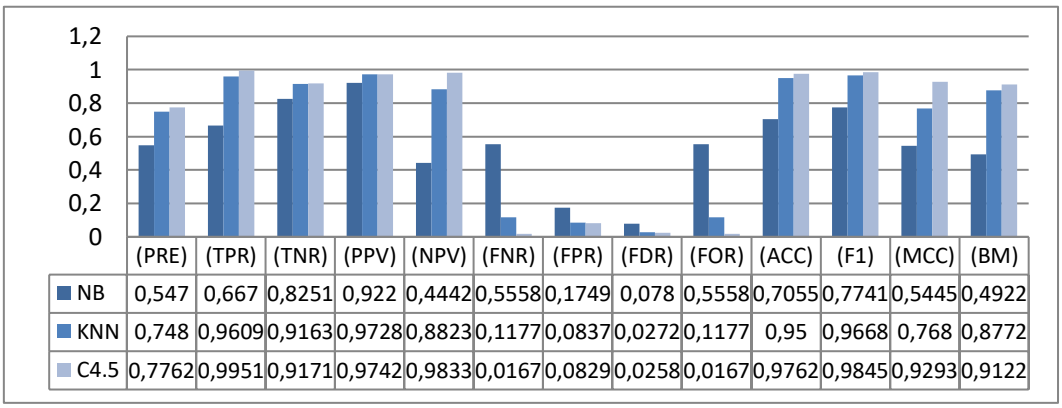
K fold	NB	C4.5	KNN
3	1	2	3
5	2	1	3
7	1	2	3
9	1	2	3

Dari tabel 4 bisa terlihat C4.5 tidak pernah mengunakan memory terbanyak, dengan kompleksitas arsitektur komputer dan optimasi java maka sangat sulit mengukur jumlah absulut pengunaan memory tetapi dengan mengunakan hitungan rata-rata dan eksperimen yang berulang maka bisa dibilang pengunaan memory C4.5 adalah terbaik dengan jumlah dataset tenaga latih yang tinggi disbanding terbalik dengan KNN atau NB.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Informasi tabel 5 menunjukan kekuatan utama NB yaitu pada pengunaan waktu. Dengan porsi data latih yang lebih tinggi dari biasanya NB bayes mengalahkan metode C4.5 dan KNN dalam pengunaan waktu. Ini adalah kelebihan NB yang bisa dijadikan bukti NB sangat kompabilitas dengan implementasi klasifikasi berbasis *realtime*.

Akurasi K-NN memang tidak menjadi yang terbaik tetapi ini dikarenakan data dengan kelas *left* tidak mudah dikelasifikasi oleh K-NN. Hasil skenario *Cross Validtation* didapatkan dengan nilai K=1 dengan kata lain hasil komparasi terbaik K-NN bersih dari noise tetapi tidak menjamin nilai K adalah optimal atau kemungkinan jumlah data kurang untuk bisa menghasilkan hasil komparasi terbaik dengan nilai K lebih dari 1. Grafik pada gambar 5 akan lebih menjelaskan komparasi relevancy K-NN terhadap dua metode lainya.



	(PRE)	(TPR)	(TNR)	(PPV)	(NPV)	(FNR)	(FPR)	(FDR)	(FOR)	(ACC)	(F1)	(MCC)	(BM)
NB	0,547	0,667	0,8251	0,922	0,4442	0,5558	0,1749	0,078	0,5558	0,7055	0,7741	0,5445	0,4922
KNN	0,748	0,9609	0,9163	0,9728	0,8823	0,1177	0,0837	0,0272	0,1177	0,95	0,9668	0,768	0,8772
C4.5	0,7762	0,9951	0,9171	0,9742	0,9833	0,0167	0,0829	0,0258	0,0167	0,9762	0,9845	0,9293	0,9122

Gambar 5 Rata-rata Relevancy dalam Skenario *Cross Validation*

Naïve bayes tertinggal jauh dibandingkan dengan dua model lainya seperti pada grafik gambar 5, tetapi tertinggal jauh tidak bisa dibilang gagal. MCC naïve bayes masih menghasilkan angka positive dan akurasinya tidak kurang dari setengah. BM naïve bayes yang dibawah 0.5 menunjukan bahwa model tidak bisa diandalkan saat penentuan kelas data, ini semakin mendukung bahwa model naïve bayes tidak cocok dengan dataset. Ketidakcocokan dataset dengan model bukan berarti model gagal, karena naïve bayes dengan informasi yang kurang mampu menghasilkan model kelasifikasi dengan akurasi 7 dari 10. Ketidakmampuan Naïve bayes dalam menghasilkan *relevancy* yang baik ini diakibatkan karena kegagalan dalam penentuan kelas stay, kemungkinan keberadaan relasi antar data dalam dataset semakin tinggi, ini bisa dijadikan petunjuk naïve bayes tidak menghasilkan komparasi yang lebih baik karena prinsip naïve bayes yang mengasumsikan tiap atribut indepen pada atribut lainnya.

C4.5 dengan data latih dan data uji yang berbeda tetap menghasilkan model klasifikasi terbaik. C4.5 juga masih memiliki karakteristic mampu mengklasifikasikan data dengan benar sehingga nilai DOR terbaik masih dipegang oleh model C4.5.

Secara umum hasil skenario *Cross Validation* menunjukan Akurasi dan DOR terbaik dihasilkan oleh metode klasifikasi C4.5 di susun K-NN dan terakhir naïve bayes. Tabel 6 menunjukan ranking akurasi dari masing-masing metode dalam penelitian ini.

Tabel 6 Hasil Ranking pengunaan memory skenario *Cross Validation*

K fold	NB	C4.5	KNN
3	3	1	2
5	3	1	2
7	3	1	2
9	3	1	2

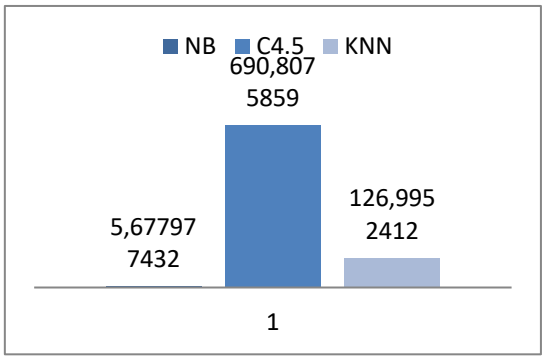
Dari tabel 6 bisa terlihat C4.5 mendominasi ranking akurasi dataset tenaga kerja. Saat fase pelatihan model C4.5 yang menghasilkan data komparasi terbaik memperkuat bahwa ada

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

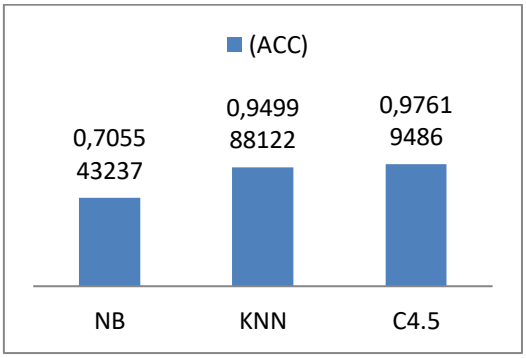
koneksi atau relasi antar atribut dalam dataset, ini belum bisa dibuktikan tetapi bisa dijadikan petunjuk bahwa dataset akan lebih baik dalam membangun model prediksi dengan metode berbasis pohon keputusan.

Naïve bayes lemah dalam dataset tenaga kerja ini karena kemungkinan relasi data tidak bisa ditoleransi saat pembangunan model, yang mengakibatkan fase pelatihan naïve bayes kalah dibanding dengan K-NN atau C4.5. Fase klasifikasi K-NN selalu berada pada waktu tunggu tertinggi ini, ini berarti bahwa metode K-NN tidak cocok untuk kasus dengan data yang memiliki skalabilitas tinggi, sedangkan naïve bayes berperforma baik saat skalabilitas dataset tinggi dengan relasi data yang masih bisa ditoleransi.

Dukungan pada dominasi keakuratan C4.5 bisa juga terlihat pada gambar 8 yang berisi jumlah rata-rata akurasi terhadapa masing-masing parameter k-fold dalam skenario *cross validation*.



Gambar 7. Rata-rata Diagnostic Odd Ratio dalam Skenario *Cross Validation*



Gambar 8. Rata-rata akurasi dalam skenario *cross validation*

Kekuatan akurasi besar C4.5 terhadap metode KNN dan NB ini bisa menjadi bukti pemenuhan syarat kedua hipotesis HO Skenario komparasi *Cross Validation* ini memiliki banyak petunjuk mengarah ke H0 adalah benar dengan metode C4.5 memiliki akurasi tertinggi disbanding K-NN dan *Naïve Bayes* sedangkan dugaan bahwa hipotesis H1 adalah tidak benar semakin kuat.

IV. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarakan penelitian analisa komparasi *machine learning* dalam klasifikasi prediksi tenaga kerja non-aktif didapati hasil skenario *supply training* dan *Cross Validation* ini didapati akurasi tertingi selalu dimiliki C4.5 ini menunjukan bahwa hipotesis H0 yaitu hasil akurasi metode C4.5 lebih tinggi dibanding K-NN dan *naïve bayes* adalah benar. Hasil kesimpulan lain yang didapatkan adalah Hasil *relevancy* dengan akurasi yang tinggi didukung BM(*Bookmaker Informedness*) dan DOR(*Diagnostic Odd Ratio*) yang tinggi menunjukan bahwa model C4.5 dalam setiap penentuan kelas bisa diandalkan dalam pengambilan keputusan dan merupakan metode terbaik dibanding dengan K-NN atau *Naïve Bayes* (NB).DOR dan BM yang rendah dimiliki naïve bayes menunjukan model yang dibangun dengan basis probabilitas akan berkinerja buruk saat asumsi ketiadaan relasi anatar atribut dalam dataset tenaga kerja tidak bisa ditoleransi *naïve* bayes.Dataset tenaga kerja lebih memiliki kompabalitas terbaik saat dibangun dengan metode berbasis pohon keputusan (C4.5) dan akan menghasilkan *relevancy* yang kurang saat model prediksi dibuat berbasis probabilitas *Naïve Bayes*.Metode *Naïve Bayes* memiliki performa baik dalam fase klasifikasi ini bisa digunakan untuk kasus klasifikasi *realtime* dengan skalabilitas tinggi sedangkan metode K-NN berperforma buruk tetapi tetap menghasilkan *relevancy* yang bisa bersaing dengan metode C4.5.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

4.2 Saran

Berdasarkan hasil penelitian komparasi *machine learning*, berikut adalah saran untuk pengembangan selanjutnya. Mengoptimasikan tahap dalam metode *Naïve Bayes* untuk mendapatkan *relevancy* yang lebih tinggi. Mencari penentuan nilai K optimal pada metode K-NN dan naïve bayes. Mengurangi pengunaan sumber daya K-NN yang tinggi tanpa mengorbankan kemampuan klasifikasi K-NN. Mengunakan metode klasifikasi berbasis pohon keputusan lain-nya dengan dataset tenaga kerja untuk komparasi *relevancy* model pohon keputusan terbaik.

DAFTAR PUSTAKA

Arnold, K., Gosling, J., & Holmes, D. (2005). *THE Java Programming Language, Fourth Edition.* Addison Wesley Professional.

Fitri, S. (2014). *Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian, Lazy-IBK , Zero-R, dan Decision Tree-J48 (2014).*

Good, I. J. (1950). *Probability and the Weighing of Evidence.* University of Wisconsin - Madison: Charles Griffin.

Hamakonda, T. P. (1991). *Pengantar klasifikasi persepuluhan dewey.* Jakarta: BPK Gunung Mulia.

Hastuti, K. (2012). *Analisis Komparisasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non-aktif.*

Kotsiantis, S. (2007). *Supervised machine learning: a review of classification techniques.*

Michie, D., Spiegelhalter, D., & Taylor, C. (2009). *Machine Learning: Neural and Statistical Classification.* Cambridge : project StatLog.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning.* Massachutes: MIT press.

Myler, H. R. (1998). *Fundamentals of Engineering Programming with C and Fortran.* Cambridge: University Press.

Quinlan, J. R. (1979). *Induction over Large Data Bases.* San Francisco: STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE.

Sanu, A. N. (2016). *Studi Perbandingan Performansi Multinomial Naïve Bayes Dan Transformed Complement Naïve Bayes Saat Klasifikasi Teks Pada Dataset Yang Tidak Seimbang.*

Sartika, D., & Sensuse, D. I. (2017). *Perbandingan Algoritma Klasifikasi Naïve Bayes, Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian.*

Słowiński, R. (1989). Rough classification in incomplete information systems. *Mathematical and Computer Modelling*, 1347-1357.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)