

NASKAH PUBLIKASI

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



Diajukan oleh

NEUTRINO SAE BELAUDIN KUSRORONG

NIM 1306081008

**ILMU KOMPUTER
FAKULTAS SAINS DAN TEKNIK
UNIVERSITAS NUSA CENDANA
KUPANG
2018**

NASKAH PUBLIKASI

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

Diajukan oleh

NEUTRINO SAE BELAUDIN KUSRORONG

NIM 1306081008

Penelitian Skripsi Ini

Telah Disetujui Oleh Pembimbing :

Pembimbing Utama



Derwin R. Sina, S.Kom, M.Cs
NIP. 19800716 200812 1 002

Pembimbing Pendamping



Nelci Dessy Rumlaklak, S.Kom., M.Kom
NIP. 19811227 201404 2 002

PERNYATAAN

Dengan ini kami selaku pembimbing skripsi mahasiswa Program Sarjana :

Nama : Neutrino Sae Belaudin Kusrorong



NIM : 1306081008

Program Studi : Ilmu Komputer

Setuju / ~~Tidak Setuju~~ *) naskah ringkasan penelitian (calon naskah berkala Penelitian Sarjana) yang disusun oleh yang bersangkutan dipublikasikan **dengan/~~tanpa~~ ***) mencantumkan nama tim pembimbing sebagai co-author.

Kemudian harap maklum.

Kupang, 7 September 2018

Nama	Status Pembimbing	Tanda Tangan
<u>Derwin R. Sina, S.Kom, M.Cs</u> NIP. 19800716 200812 1 002	Pembimbing Utama	
<u>Nelci Dessy Rumlaklak, S.Kom., M.Kom</u> NIP. 19811227 201404 2 002	Pembimbing Pendamping	

*)Coret yang tidak perlu

INTISARI

Analisa Komparasi *Machine Learning* Dalam Klasifikasi Prediksi Tenaga Kerja Non-Aktif

Oleh

Neutrino Sae Beludin Kusrorong

1306081008

Studi komparasi terhadap 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 mayoritas 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 Keputusan C4.5, dan Tenaga Kerja Non-Aktif.

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

Neutrino Sae B. Kusrorong¹, Derwin R. Sina², Nelci Dessy Rumlaklak³
^{1,2,3} Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana

INTISARI

Studi komparasi terhadap 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 mayoritas 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 Keputusan 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 training 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.

1. 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 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 perusahaan 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.

2. 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 keterbatasan 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.

2.2 Metode Klasifikasi

Klasifikasi bisa dikatakan sebagai pembelajaran mesin karena memiliki kemampuan untuk menggunakan 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).

$$KNN(x, S_y) = \operatorname{argmin} \left(\sum_{i=0}^{k-1} \operatorname{Nearest}(x, S_y) \right)$$

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 mayoritas, dimana untuk setiap ihwalan baru akan dilabelkan berdasarkan jumlah terbanyak voting ihwal yang telah ada. Fungsi nearest didapatkan dengan membandingkan jarak kemiripan data baru dengan semua data dalam himpunan data latih seperti persamaan berikut :

$$\operatorname{Nearest}(x, S_y) = \min \left(\sum_{i=0}^{S_n-1} d(x, y_i) \right)$$

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.

$$d(x, y) = \sqrt{\sum_{i=0}^{f_n-1} (f_x - f_y)^2}$$

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 menggunakan 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
- 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 menggunakan fungsi probabilitas *laplacian smoothing* dengan nilai K=1 untuk untuk menghindari probabilitas bernilai nol, persamaan probabilitas *laplacian smoothing* bisa dilihat pada persamaan berikut:

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

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:

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

Keterangan:

- P = Probabilitas
- E = *Evidence Value*
- C = *Class* objek
- 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 :

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

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.

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

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

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:

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

Keterangan:

- S = himpunan kasus
- n = jumlah partisi S
- P_i = proporsi dari S_i 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 lingkungannya 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 pgunaan *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).

Tabel 1 Confusion Matrix Sederhana

		CLASSIFIER		
		State	TRUE	FALSE
AKTUAL	TRUE	True Positive (TP)	False Negative (FN)	
	FALSE	False Positive (FP)	True Negative (TN)	

Confusion Matrix bisa merepresentasikan utilitas suatu model klasifikasi karena bisa menunjukkan informasi hasil klasifikasi dan langsung dibandingkan dengan data aktual. Menurut (Olson & Delen, 2008) dasar Confusion Matrix seperti yang terlihat pada selengkapnya pada tabel berikut :

Tabel 2 Relevancy

		Predicted Condition			
		Prediction Positive	Prediction Negative	Prevalence = Σ Condition Positive / Σ Total Population	
True Condition	Actual Positive	True Positive (TP)	False Negative (FN)	True Positive Rate (TPR), Sensitivity, Recall, Probability Of Detection = $\Sigma TP / \Sigma$ Condition Positive	False Negative Rate (FNR), Miss Rate = $\Sigma FN / \Sigma$ Condition Positive
	Actual Negative	False Positive (FP)	True Negative (TN)	False Positive Rate (FPR), Fall-out, Probability of False Alarm = $\Sigma FP / \Sigma$ condition negative	True Negative Rate (TNR), Specificity (SPC) = $\Sigma TN / \Sigma$ Condition Negative
		Positive Predictive Value (PPV), Precision = $\Sigma TP / \Sigma$ Prediction Positive	False Omission Rate (FOR) = $\Sigma FN / \Sigma$ Prediction Negative	Positive Likelihood Ratio (LR+) = TPR / FPR	Diagnostic Odds Ratio (DOR) = $LR+ / LR-$
		False Discovery Rate (FDR) = $\Sigma FP / \Sigma$ Prediction Positive	Negative Predictive Value (NPV) = $\Sigma TN / \Sigma$ 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 sederhana untuk menguji model klasifikasi 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.

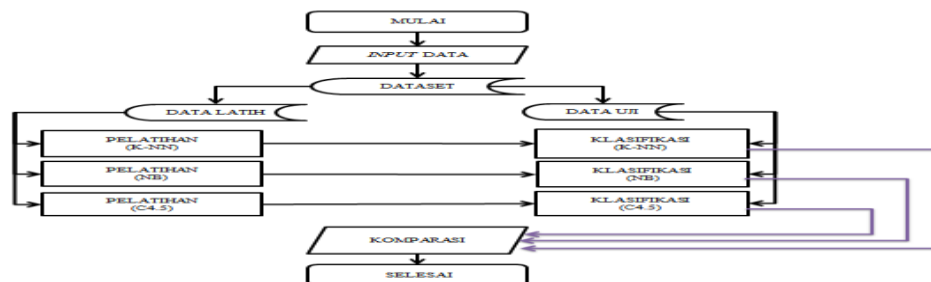
2.4.2 Skenario Cross-Validation

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

3 HASIL DAN PEMBAHASAN

3.1 Hasil

Hasil sistem komparasi pembelajaran mesin yang didapat akan dibagi menjadi hasil implementasi antarmuka dan hasil implementasi algoritma. Bagian implementasi antarmuka membahas tampilan sistem dengan memaksimalkan informasi yang ditampilkan kepada pengguna. 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 menghasilkan keluaran *relevancy* dan penggunaan sumber daya, hasil klasifikasi masing-masing metode akan dibandingkan dan disusun untuk menjadi informasi *output* hasil komparasi. *Flowchart* atau alur kinerja sistem komparasi pembelajaran mesin bisa dilihat pada gambar berikut :



Gambar 1 Flowchart Sistem Komparasi Pembelajaran Mesin

3.1.2 Hasil Implementasi algoritma

Implementasi analisa komparasi *machine learning* dalam klasifikasi prediksi tenaga kerja non-aktif dibagi tiga algoritma yaitu *k-nearest neighbour*, *naïve bayes* dan *Decision Tree C4.5*. Sebelum implementasi masing-masing algoritma berikut adalah hasil pengukuran dasar dan pengukuran statistic dataset yang digunakan dalam membangun model prediksi. Pengukuran dasar mewakili ukuran satuan kateristik dasar yang ada dalam dataset, tabel pengukuran dasar bisa dilihat pada tabel 3,

Tabel 3 Pengukuran Dasar

Jumlah Karakteristik	Jumlah
Data	14999
Atribut	9
Kelas	2
Atribut Binary	2
Atribut Kontinous	2
Atribut <i>Kategorical</i>	2
Atribut <i>Numeric</i>	3

Statistik deskriptif adalah analisis yang berkaitan dengan pengumpulan dan penyajian data sehingga dapat memberikan informasi yang berguna. Pengukuran statistic deskriptif berguna menunjukkan informasi umum nilai yang berada pada dataset tenaga kerja, informasi statistic pengukuran dasar bisa dilihat pada tabel 4

Tabel 4 Pengukuran Statistic Deskriptif

Attribut	N	Min	Q1	Q2	Q3	Max	Mean	Mode	Sum	STD
id	14999	1	3750	7500	11248	14999	7500	14974	1.12E+08	4329.98268
satisfaction	14999	0.09	0.44	0.64	0.82	1	0.61283352	0.1	9191.89	0.24863065
evaluation	14999	0.36	0.56	0.72	0.87	1	0.71610174	0.55	10740.8	0.17116911
numberproject	14999	2	3	4	5	7	3.80305354	4	57042	1.23259236
avaragehours	14999	96	156	200	245	310	201.050337	156	3015554	49.9430994
timespendcompany	14999	2	3	3	4	10	3.49823322	3	52470	1.46013623
workaccident	14999	0	0	0	0	1	0.14460964	0	2169	0.35171855
promotion	14999	0	0	0	0	1	0.02126808	0	319	0.14428146
division	14999	1	3	7	9	10	6.12087472	0	91807	2.99212032
salary	14999	1	1	2	2	3	1.59470631	1	23919	0.63718295
lefts	14999	0	0	0	0	1	0.23808254	0	3571	0.4259241

Pembahasan singkat tabel 4.2 menunjukkan tabel pengukuran statistic ini menunjukkan dataset yang digunakan tidak memiliki data yang *null*, *anomaly* atau *froud*. Dataset yang digunakan telah bersih dari *noise* yang bisa mengganggu 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 Training, 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%. Hasil implementasi bisa dilihat pada tabel 5 hingga 7.

a) Hasil K-NN

Tabel 5 Hasil Komparasi Supply Training K-NN Dengan K Terbaik

PERCOBAAN	K-NN #1	K-NN #6	K-NN #11	K-NN #16	K-NN #21	K-NN #26
METODE	K-NN k=1	K-NN k=1	K-NN k=1	K-NN k=1	K-NN k=1	K-NN k=1
P LATIH	50	60	70	80	90	100
P TEST	50	40	30	20	10	100
T. TIME	15	4	5	1	2	7
T. MEMORY	99200	1592728	0	0	0	0
C. MEMORY	2243664	309376776	26568752	211027384	88302968	89452776
C. TIME	21148	14631	12281	13727	15976	23402
N LATIH	7500	9000	10500	12000	13500	14999
N TEST	7499	5999	4499	2999	1499	14999
TP	5488	4370	3281	2182	1103	11428
TN	1658	1368	1032	705	350	3571
FP	131	94	66	42	17	0
FN	222	167	120	70	29	0
PRE	0.749	0.744	0.744	0.742	0.747	0.762
TPR	0.961	0.963	0.965	0.969	0.974	1.000
TNR	0.927	0.936	0.940	0.944	0.954	1.000
PPV	0.977	0.979	0.980	0.981	0.985	1.000
NPV	0.882	0.891	0.896	0.910	0.923	1.000
FNR	0.118	0.109	0.104	0.090	0.077	0.000
FPR	0.073	0.064	0.060	0.056	0.046	0.000
FDR	0.023	0.021	0.020	0.019	0.015	0.000
FOR	0.118	0.109	0.104	0.090	0.077	0.000
ACC	0.953	0.956	0.959	0.963	0.969	1.000
F1	0.969	0.971	0.972	0.975	0.980	1.000
MCC	0.873	0.884	0.890	0.902	0.918	1.000
BM	0.888	0.899	0.905	0.913	0.928	1.000
DOR	103.014	128.844	144.813	180.065	262.174	∞
LR +	13.126	14.981	16.049	17.233	21.035	∞
LR -	0.127	0.116	0.111	0.096	0.080	0.000

b) Hasil NB

Tabel 6 Hasil Komparasi Supply Training naïve bayes dengan K terbaik

PERCOBAAN	NB #4	NB #6	NB #15	NB #16	NB #24	NB #30
METODE	NB K=4	NB K=1	NB K=5	NB K=1	NB K=4	NB K=5
P LATIH	50	60	70	80	90	100
P TEST	50	40	30	20	10	100
T. TIME	981	1644	1918	2612	3199	3954
T. MEMORY	12130808	11484080	15892824	22004448	22710880	22888960
C. MEMORY	14435280	40231128	18313948	94071824	67894128	138255272
C. TIME	2143	5533	1647	5506	1263	4752
N LATIH	7500	9000	10500	12000	13500	14999
N TEST	7499	5999	4499	2999	1499	14999
TP	3789	3020	2263	1490	766	7668
TN	1542	1225	897	615	300	2985
FP	247	237	201	132	67	586
FN	1921	1517	1138	762	366	3760
PRE	0.538	0.543	0.548	0.541	0.556	0.550
TPR	0.664	0.666	0.665	0.662	0.677	0.671
TNR	0.862	0.838	0.817	0.823	0.817	0.836
PPV	0.939	0.927	0.918	0.919	0.920	0.929
NPV	0.445	0.447	0.441	0.447	0.450	0.443
FNR	0.555	0.553	0.559	0.553	0.550	0.557
FPR	0.138	0.162	0.183	0.177	0.183	0.164
FDR	0.061	0.073	0.082	0.081	0.080	0.071
FOR	0.555	0.553	0.559	0.553	0.550	0.557
ACC	0.711	0.708	0.702	0.702	0.711	0.710
F1	0.778	0.775	0.772	0.769	0.780	0.779
MCC	0.449	0.434	0.416	0.421	0.428	0.434
BM	0.526	0.504	0.482	0.485	0.494	0.507
DOR	7.468	6.219	5.310	5.571	5.513	6.131
LR +	4.806	4.106	3.635	3.744	3.707	4.089
LR -	0.644	0.660	0.685	0.672	0.672	0.667

c) Hasil C4.5

Tabel 7 Hasil Komparasi Supply Training C4.5

PERCOBAAN	C4.5 #1	C4.5 #2	C4.5 #3	C4.5 #4	C4.5 #5	C4.5 #6
METODE	C4.5	C4.5	C4.5	C4.5	C4.5	C4.5
P LATIH	50	60	70	80	90	100
P TEST	50	40	30	20	10	100
T. TIME	844	559	769	575	963	568
T. MEMORY	9318056	18178088	48081536	93725696	32744200	124338864
C. MEMORY	173032392	148163352	121286424	97561856	67060936	155515864
C. TIME	7935	4144	6069	3151	3837	12927
N LATIH	7500	9000	10500	12000	13500	14999
N TEST	7499	5999	4499	2999	1499	14999
TP	5679	4496	3383	2240	1126	11398
TN	1640	1341	999	679	330	3360
FP	149	121	99	68	37	211
FN	31	41	18	12	6	30
PRE	0.777	0.770	0.774	0.770	0.776	0.774
TPR	0.995	0.991	0.995	0.995	0.995	0.997
TNR	0.917	0.917	0.910	0.909	0.899	0.941
PPV	0.974	0.974	0.972	0.971	0.968	0.982
NPV	0.981	0.970	0.982	0.983	0.982	0.991
FNR	0.019	0.030	0.018	0.017	0.018	0.009
FPR	0.083	0.083	0.090	0.091	0.101	0.059
FDR	0.026	0.026	0.028	0.029	0.032	0.018
FOR	0.019	0.030	0.018	0.017	0.018	0.009
ACC	0.976	0.973	0.974	0.973	0.971	0.984
F1	0.984	0.982	0.983	0.982	0.981	0.990
MCC	0.933	0.926	0.929	0.928	0.922	0.955
BM	0.911	0.908	0.905	0.904	0.894	0.938
DOR	590.076	370.190	567.119	571.923	496.812	1794.708
LR +	11.942	11.973	11.032	10.927	9.866	16.880
LR -	0.020	0.032	0.019	0.019	0.020	0.009

B. Hasil Skenario Cross Validataion

Hasil skenario *Cross Validation* dilakukan dengan membagi data kedalam beberapa himpunan 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 penggunaan 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 menghasilkan informasi model terbaik metode K-NN dan NB. Nilai performa didapatkan dengan membandingkan 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. Hasil skenario terhadap algoritma bisa dilihat pada Tabel 9 hingga 12 .

Tabel 8 Keterangan Legenda

HIJAU = TERBAIK	KUNING = SEDANG	MERAH = TERBURUK
-----------------	-----------------	------------------

1) Hasil *K-Fold* = 3

Tabel 9 Hasil Rata-Rata Komparasi *Cross Validation K-Fold* = 3

METODE	C4.5	KNN	NB
K Fold	3	3	3
T. TIME	844	11	123
T. MEMORY	40,231,128	146,117,960	144,352,800
C. MEMORY	31,901,421	137,252,997	92,604,835
C. TIME	7,935	11,573	6,129
JUMLAH DATA	5000	5000	5000
TP	3793	3668	2556
TN	1099	1121	981
FP	91	69	209
FN	17	141	1254
(PRE)	0.777	0.747	0.553
(TPR)	0.996	0.963	0.671
(TNR)	0.924	0.942	0.824
(PPV)	0.977	0.982	0.924
(NPV)	0.985	0.888	0.439
(FNR)	0.015	0.112	0.561
(FPR)	0.076	0.058	0.176
(FDR)	0.023	0.018	0.076
(FOR)	0.015	0.112	0.561
(ACC)	0.978	0.958	0.707
(F1)	0.986	0.972	0.777
(MCC)	0.919	0.495	0.905
(BM)	0.919	0.905	0.495
(LR+)	13.023	16.611	3.815
(LR-)	0.016	0.119	0.681
(DOR)	805.377	139.801	5.604
Total Time	8,779.00	11,584.00	6,252.00
Total Memory	72,132,549.00	283,370,957.00	236,957,635.00

2) Hasil *K-Fold* = 5

Tabel 10 Hasil Rata-Rata Komparasi *Cross Validation K-Fold* = 5

METODE	C4.5	KNN	NB
K Fold	5	5	5
T. TIME	559	15	145
T. MEMORY	38,439,040	30,375,456	31,537,032
C. MEMORY	121,286,424	148,163,352	173,032,392
C. TIME	4,144	8,703	7,935
JUMLAH DATA	3000	3000	3000
TP	2240	2159	1489
TN	679	637	615
FP	68	94	132
FN	12	110	763
(PRE)	0.769	0.751	0.540
(TPR)	0.995	0.952	0.661
(TNR)	0.909	0.871	0.823
(PPV)	0.971	0.958	0.919
(NPV)	0.983	0.853	0.446
(FNR)	0.017	0.147	0.554
(FPR)	0.091	0.129	0.177
(FDR)	0.029	0.042	0.081
(FOR)	0.017	0.147	0.554
(ACC)	0.973	0.932	0.701
(F1)	0.982	0.955	0.769
(MCC)	0.928	0.817	0.420
(BM)	0.904	0.823	0.484
(LR+)	10.927	7.400	3.742
(LR-)	0.019	0.169	0.673
(DOR)	571.923	43.788	5.564
Total Time	4,703.00	8,718.00	8,080.00
Total Memory	159,725,464.00	178,538,808.00	204,569,424.00

3) Hasil *K-Fold* = 7

Tabel 11 Hasil Rata-Rata Komparasi *Cross Validation K-Fold* = 7

METODE	C4.5	KNN	NB
K Fold	7	7	7
T. TIME	769	21	234
T. MEMORY	18,178,088	9,318,056	32,783,440
C. MEMORY	155,515,864	67,060,936	97,561,856
C. TIME	6,069	9,393	3,151
JUMLAH DATA	2143	2143	2143
TP	1610	1577	1078
TN	472	500	438
FP	53	24	85
FN	9	41	543
(PRE)	0.776	0.747	0.542
(TPR)	0.995	0.974	0.665
(TNR)	0.899	0.954	0.838
(PPV)	0.968	0.985	0.927
(NPV)	0.982	0.923	0.446
(FNR)	0.018	0.077	0.554
(FPR)	0.101	0.046	0.162
(FDR)	0.032	0.015	0.073
(FOR)	0.018	0.077	0.554
(ACC)	0.971	0.969	0.707
(F1)	0.981	0.980	0.774
(MCC)	0.922	0.918	0.433
(BM)	0.894	0.928	0.503
(LR+)	9.866	21.035	4.102
(LR-)	0.020	0.080	0.661
(DOR)	496.812	262.174	6.207
Total Time	6,838.00	9,414.00	3,385.00
Total Memory	173,693,952.00	76,378,992.00	130,345,296.00

4) Hasil *K-Fold* = 9

Tabel 12 Hasil Rata-Rata Komparasi *Cross Validation K-Fold* = 9

METODE	C4.5	KNN	NB
K Fold	9	9	9
T. TIME	575	22	299
T. MEMORY	93,725,696	93,725,696	48,081,536
C. MEMORY	20,703,784	44,779,856	30,220,624
C. TIME	3,151	8,442	3,337
JUMLAH DATA	1667	1667	1667
TP	1281	1203	845
TN	356	365	332
FP	24	41	75
FN	6	57	414
(PRE)	0.783	0.747	0.552
(TPR)	0.995	0.955	0.671
(TNR)	0.937	0.898	0.815
(PPV)	0.982	0.967	0.918
(NPV)	0.983	0.865	0.445
(FNR)	0.017	0.135	0.555
(FPR)	0.063	0.102	0.185
(FDR)	0.018	0.033	0.082
(FOR)	0.017	0.135	0.555
(ACC)	0.982	0.941	0.706
(F1)	0.988	0.961	0.775
(MCC)	0.948	0.842	0.420
(BM)	0.932	0.853	0.486
(LR+)	15.746	9.360	3.632
(LR-)	0.018	0.150	0.680
(DOR)	889.119	62.218	5.337
Total Time	3,726.00	8,464.00	3,636.00
Total Memory	114,429,480.00	138,505,552.00	78,302,160.00

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 kontinuous, 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 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 Training*, dimulai dengan menentukan 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. Pada tabel 4.9 bisa diperhatikan untuk hasil masing-masing model dengan pembagian data 50%-50%, 60%-40%,70%-30%,80%-20% dan 90-10%.

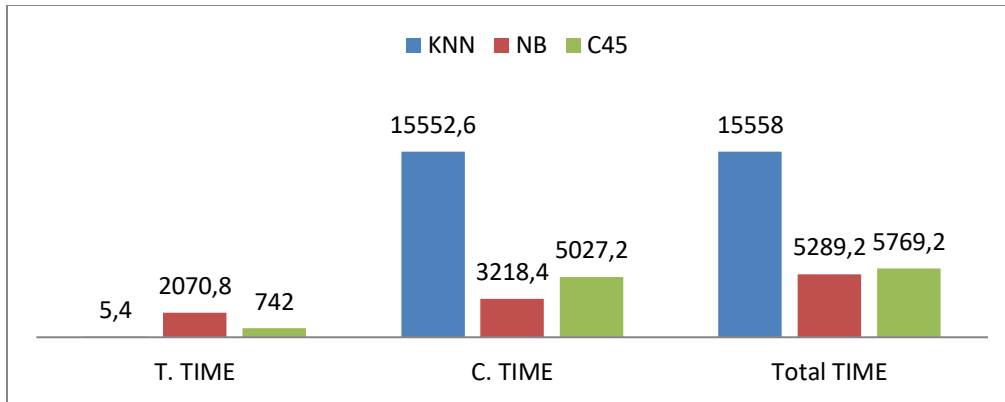
Tabel 13 Hasil Komparasi Rata-Rata *Supply Training*

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

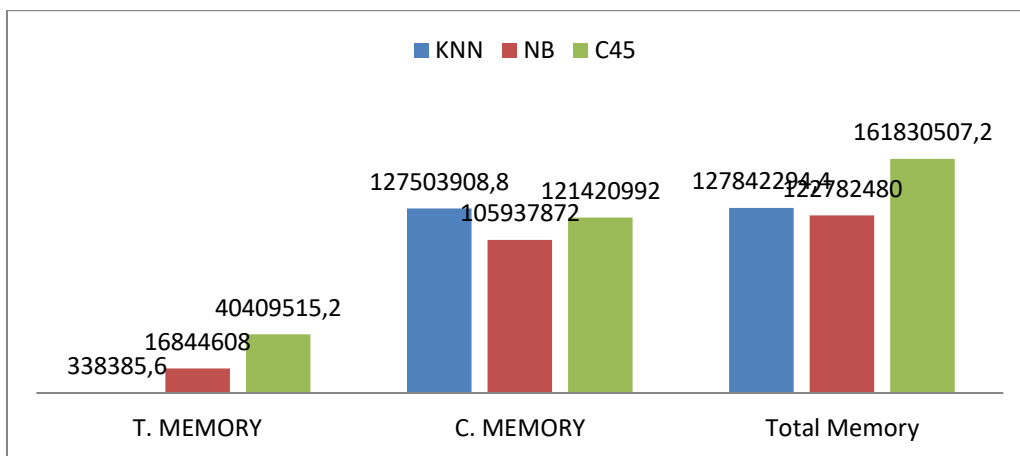
LEGENDA

HIJAU = TERBAIK	KUNING = SEDANG	MERAH = TERBURUK
-----------------	-----------------	------------------

K-NN memiliki kelebihan pada fase training dengan penggunaan waktu dan memory terkecil dibanding dengan dua metode lain (T. Memory dan T. Time). Keunggulan K-NN hanya terletak pada fase pelatihan dimana pada fase klasifikasi lebih didominasi oleh NB dan C4.5. Grafik penggunaan sumber daya pada grafik gambar 2.



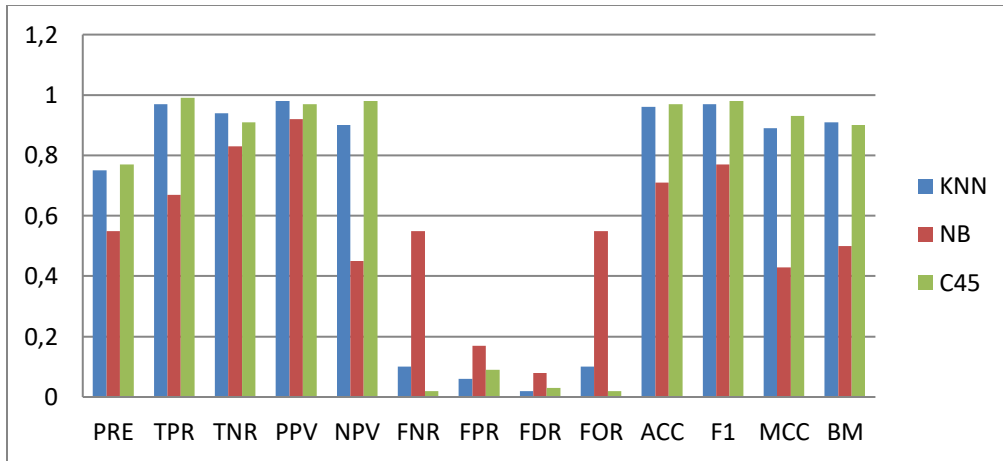
Gambar 2 Grafik Rata-rata Penggunaan Sumber Daya Waktu Supply Training



Gambar 3 Rata-rata Penggunaan Sumber Daya Memory Supply Training

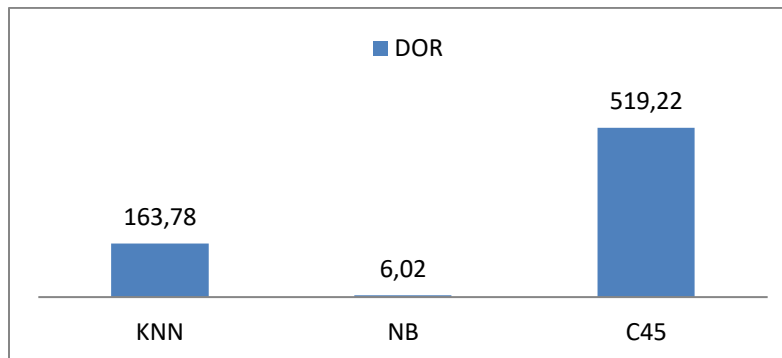
Seperti pada grafik pada gambar 3 dan gambar 4 naïve bayes dengan *laplacian smoothing* menghasilkan metode yang terbaik dalam penggunaan sumber daya waktu dan memori artinya naïve bayes akan lebih optimal pada kasus yang memerlukan kemampuan model realtime yang tidak bisa diimplementasikan dengan mudah 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 membedakan 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 menunjukkan 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*. Grafik pada gambar 4 menunjukkan nilai komparasi pada masing-masing algoritma.



Gambar 4 Rata-rata Relevancy dalam Skenario *Supply Training*

Grafik pada gambar 4 menunjukkan 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 independensi 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 klasifikasi untuk menyatakan metode naïve bayes suatu kegagalan besar seperti yang terlihat pada grafik Diagnostic Odd Ratio pada gambar 4.25.



Gambar 5 Rata-rata Diagnostic Odd Ratio dalam Skenario *Supply Training*

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 klasifikasi dengan FN, FNR, FOR terendah. Keunggulan model klasifikasi C4.5 juga ada pada angka akurasi dengan menggunakan sumber daya memori 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 klasifikasi dataset tenaga kerja seperti terlihat pada grafik gambar 4.23.

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 penggunaan sumber daya terutama pada fase pelatihan dengan penggunaan memori 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.

Dari data pada tabel 4.10 diatas bisa diketahui K-NN hampir unggul dalam seluruh komparasi hasil, ini dikarenakan K-NN yang memiliki seluruh jenis data dengan K = 1 tidak lebih dari pada metode pencocokan data, untuk setiap kasus dalam dunia nyata tidak akan pernah mungkin untuk merekam keseluruhan kemungkinan data.

Skenario komparasi *Supply Training* secara mayoritas akurasi terbaik dimiliki metode C4.5 dengan demikian 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*, penggunaan *memory* dan penggunaan waktu disatukan dan dijadikan acuan komparasi. Menggunakan informasi yang ada dalam tabel 9 sampai dengan tabel 12 maka bisa mendapatkan informasi baru berupa tabel *voting*. Tabel voting terdiri dari 16 variable relevansi, total penggunaan memori dan total penggunaan waktu. Tabel 14 adalah tabel voting komparasi skenario *Cross Validation* dengan nilai k-fold sama dengan 3,5,7, dan 9.

Tabel 14 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

Dari tabel 14 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 menunjukkan 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 komparabilitas tinggi dengan tipe dataset tenaga kerja yang memiliki banyak atribut bilangan real dan numeric.

Dengan menggunakan tabel 4.6 sampai dengan tabel 4.9 bisa juga membentuk tabel baru yaitu tabel ranking penggunaan memory secara minimal yang bisa dilihat pada tabel 4.13

Tabel 15 Hasil Ranking penggunaan 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

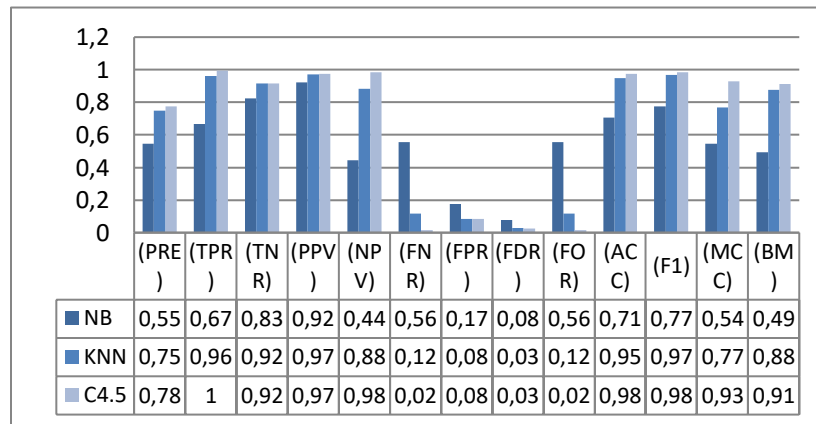
Dari tabel 15 bisa terlihat C4.5 tidak pernah menggunakan memory terbanyak, dengan kompleksitas arsitektur komputer dan optimasi java maka sangat sulit mengukur jumlah absolut penggunaan memory tetapi dengan menggunakan hitungan rata-rata dan eksperimen yang berulang maka bisa dibilang penggunaan memory C4.5 adalah terbaik dengan jumlah dataset tenaga latih yang tinggi disbanding terbalik dengan KNN atau NB.

Tabel 16 Hasil Ranking penggunaan 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

Informasi tabel 16 menunjukkan kekuatan utama NB yaitu pada penggunaan waktu. Dengan porsi data latih yang lebih tinggi dari biasanya NB bayes mengalahkan metode C4.5 dan KNN dalam penggunaan waktu. Ini adalah kelebihan NB yang bisa dijadikan bukti NB sangat kompatibility dengan implementasi klasifikasi berbasis *realtime*.

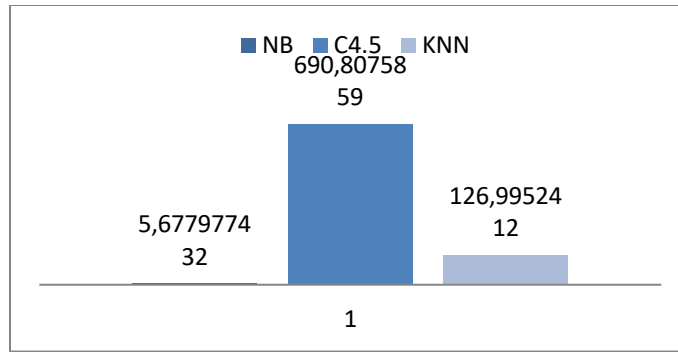
Akurasi K-NN memang tidak menjadi yang terbaik tetapi ini dikarenakan data dengan kelas *left* tidak mudah diklasifikasi oleh K-NN. Hasil skenario *Cross Validation* didapatkan dengan nilai K=1 dengan kata lain hasil komparasi terbaik K-NN bersih dari noise tetapi tidak menjamin nilai K adalah optimal atau kemungkinan jumlah data kurang untuk bisa menghasilkan hasil komparasi terbaik dengan nilai K lebih dari 1. Grafik pada gambar 4.28 akan lebih menjelaskan komparasi relevancy K-NN terhadap dua metode lainnya.



Gambar 6 Rata-rata Relevancy dalam Skenario *Cross Validation*

Naïve bayes tertinggal jauh dibandingkan dengan dua model lainnya seperti pada grafik gambar 6, tetapi tertinggal jauh tidak bisa dibalang gagal. MCC naïve bayes masih menghasilkan angka positive dan akurasi tidak kurang dari setengah. BM naïve bayes yang dibawah 0.5 menunjukkan 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 klasifikasi 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 karakteristik mampu mengklasifikasikan data dengan benar sehingga nilai DOR terbaik masih dipegang oleh model C4.5 seperti terlihat pada grafik gambar 7.



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

Secara umum hasil skenario *Cross Validation* menunjukkan Akurasi dan DOR terbaik dihasilkan oleh metode klasifikasi C4.5 di susun K-NN dan terakhir naïve bayes. Tabel 4.15 menunjukkan ranking akurasi dari masing-masing metode dalam penelitian ini.

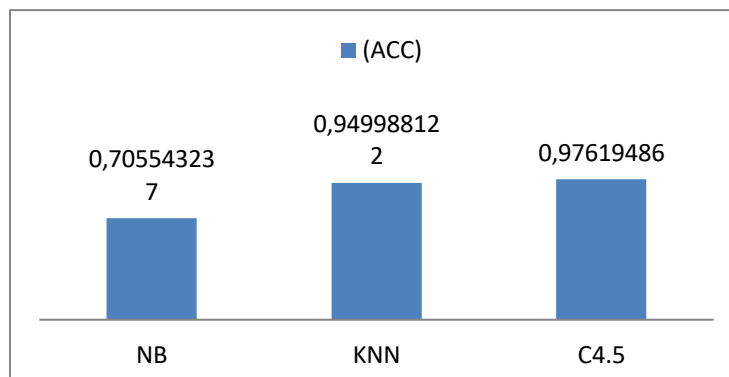
Tabel 17 Hasil Ranking penggunaan memory skenario *Cross Validation*

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

Dari tabel 17 bisa terlihat C4.5 mendominasi ranking akurasi dataset tenaga kerja. Saat fase pelatihan model C4.5 yang menghasilkan data komparasi terbaik memperkuat bahwa ada 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 4.28 yang berisi jumlah rata-rata akurasi terhadap masing-masing parameter k-fold 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 H0 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.

4 KESIMPULAN DAN SARAN

4.1 Kesimpulan

Berdasarkan penelitian analisa komparasi *machine learning* dalam klasifikasi prediksi tenaga kerja non-aktif didapati hasil skenario *supply training* dan *Cross Validation* ini didapati akurasi tertinggi selalu dimiliki C4.5 ini menunjukkan 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 :

- a) Hasil *relevancy* dengan akurasi yang tinggi didukung BM(*Bookmaker Informedness*) dan DOR(*Diagnostic Odd Ratio*) yang tinggi menunjukkan 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).
- b) DOR dan BM yang rendah dimiliki *naïve bayes* menunjukkan model yang dibangun dengan basis probabilitas akan berkinerja buruk saat asumsi ketiadaan relasi anatar atribut dalam dataset tenaga kerja tidak bisa ditoleransi *naïve bayes*.
- c) Dataset tenaga kerja lebih memiliki kompabilitas terbaik saat dibangun dengan metode berbasis pohon keputusan (C4.5) dan akan menghasilkan *relevancy* yang kurang saat model prediksi dibuat berbasis probabilitas *Naïve Bayes*.
- d) 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.
- e) Hasil klasifikasi K-NN dan C4.5 akan semakin baik jika semakin banyak informasi atau data yang digunakan, data tanpa *noise* akan menghasilkan performa dan *relevancy* terbaik dengan nilai K yang rendah yang artinya menghemat penggunaan sumber daya waktu memori dan waktu.
- f) Penggunaan sumber daya metode C4.5 dan *Naïve Bayes* akan efisien jika data latih lebih sedikit dibanding data uji, ini berbanding terbalik dengan K-NN yang akan boros sumber daya jika data latih berjumlah tinggi.
- g) Hasil performa dan *relevancy* dengan 20 atribut mampu menghasilkan wawasan lebih mendalam saat menganalisa hasil klasifikasi karena mampu memberi petunjuk kelemahan dan kelebihan metode masing-masing model yang bisa digunakan untuk membangun model prediksi berikutnya lebih baik.
- h) K-NN lemah dibidang performa tetapi memiliki hasil *relevancy* yang bisa bersaing dengan C4.5. Kinerja buruk C4.5 bisa terjadi jika seluruh data digunakan sebagai data latih dan data uji.

4.2 Saran

Berdasarkan hasil penelitian komparasi *machine learning*, berikut adalah saran untuk pengembangan selanjutnya :

- a) Mengoptimasikan tahap dalam metode *Naïve Bayes* untuk mendapatkan *relevancy* yang lebih tinggi.
- b) Mencari penentuan nilai K optimal pada metode K-NN dan *naïve bayes*.
- c) Mengurangi penggunaan sumber daya K-NN yang tinggi tanpa mengorbankan kemampuan klasifikasi K-NN.
- d) Menggunakan 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.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1999). *Classification and Regression Trees*. New York: CRC Press.
- Cover, T., & Hart, P. (1967). IEEE transactions on information theory. *Nearest neighbor pattern classification*.
- 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.
- Nilsson, N. J. (1965). *The mathematical foundations of learning machines*. California: San Mateo.
- Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques, Springer, 1st edition*. Springer.
- 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.

