JURNAL KHATULISTIWA INFORMATIKA, VOL. VII, NO. 1 JUNI 2019
p-ISSN: 2339-1928 & e-ISSN: 2579-633X

Analisa Komparasi Menggunakan 5 Metode Data Mining dalam Klasifikasi Persentase Wanita Sudah menikah di Usia 15-49 yang Memakai Alat KB (Keluarga Berencana)

Melisa Winda Pertiwi [1], Miftah Farid Adiwisastra [2], Deddy Supriadi [3]
Program Studi Sistem Informasi, STMIK Nusa Mandiri[1]
Jl. Kramat Raya No. 18, Jakarta Pusat
Program Studi Sistem Informasi Kampus Kota Tasikmalaya, Universitas BSI [2] [3]
Jl. Tanuwijaya No.4, Empangsari, Tawang, Tasikmalaya
Email : melisa.mwp@bsi.ac.id [1], miftah.mow@bsi.ac.id [2], deddy.dys@bsi.ac.id [3]

ABSTRAKSI

Program Keluarga Berencana (KB) adalah salahsatu program pemerintah yang sekarang ini banyak sekali digunakan untuk wanita untuk berbagai kebutuhannya. Badan Pengolahan Statistik (BPS) membuat suatu data persentase untuk wanita yang sudah menikah dan sedang menggunakan alat KB setiap tahunnya (*Upated* terakhir: 21 Februari 2018). Penelitian ini bertujuan untuk mendapatkan model algoritma yang baik untuk penerapan persentasenya berdasarkan keterangan bahwa setiap tahun dapat mengalami peningkatan/penurunan. Metode data mining yang digunakan adalah klasifikasi, terdiri dari 5 model algoritma yaitu *Decision Tree* (C4.5), *k-Nearest Neighbor* (k-NN), *Logistis Regression*, *Naïve Bayes*, dan *Gradient Boosted Tree*, setelah dilakukan uji hasil maka didapat bahwa dari komparasi kelima algoritma tersebut yang menunjukan baik dan akurasinya lebih besar adalah model algoritma C45 dengan nilai *accuracy* 87.50%.

Kata kunci: BPS, *Decision Tree* (C4.5), *Gradient Boosted Tree* klasifikasi, *k-Nearest Neighbor* (k-NN), *Rule Induction*, *Naïve Bayes*.

ABSTRACT

Program Keluarga Berencana (KB) is one of the government programs that are now widely used for women for various needs. Badan Pusat Statistik (BPS) makes a percentage data for women who are married and are using family planning devices every year (Updated: 21 February 2018). This study aims to obtain a good algorithm model for the application of the percentage based on information that each year can experience an increase / decrease. Data mining methods used are classification, consisting of 5 algorithm models namely Decision Tree (C4.5), k-Nearest Neighbor (k-NN), Logistic Regression, Naïve Bayes, and Gradient Boosted Tree, after the results of the test are obtained that from the comparison of the five algorithms which show good and greater accuracy is the C45 algorithm with an accuracy value of 87.50%.

Keywords: BPS, Decision Tree (C4.5), Gradient Boosted Tree classification, k-Nearest Neighbor (k-NN), Rule Induction, Naïve Bayes.

1. PENDAHULUAN

Indonesia merupakan negara yang termasuk memiliki jumlah penduduk terbanyak di dunia. Hal ini disebabkan karena negara Indonesia memiliki tingkat kelahiran yang begitu tinggi sehingga terjadilah kepadatan penduduk.Badan Kependudukan Keluarga Berencana Nasional (BKKBN) merupakan lembaga pemerintah non kementerian yang berada di bawah dan bertanggungjawab kepada Presiden melalui Menteri yang bertanggungjawab dibidang Kesehatan

Menurut Entjang (1986) dalam Garis-garis Besar Haluan Negara (GBHN) sebagai Ketetapan Majelis Permusyawarataan Rakyat (MPR) No.IV/MPR/1987 disebutkan bahwa program KB

bertujuan untuk meningkatkan kesejahteraan ibu dan anak dalam rangka mewujudkan keluarga bahagia yang menjadi dasar bagi terwujudnya masyarakat yang sejahtera dengan mengendalikan kelahiran sekaligus dalam rangka menjamin terkendalinya pertumbuhan penduduk Indonesia. Ada banyak faktor yang mempengaruhi masyarakat dalam menggunakan KB. Dalam penelitian ini diambil data wanita berusia 15-49 tahun yang berstatus menikah dan sedang menggunakan KB (data diperoleh dari www.bps.go.id update 15 Maret 2019), beberapa tahun terakhir 2000 – 2017 bisa diklasifikasikan apakah mengalami pengingkatan atau tidak dengan berdasarkan data yang terakhir dipublikasikan.

Tujuan penelitian ini adalah untuk melakukan pengujian terhadap komparasi algoritma klasifikasi manakah yang paling baik berdasarkan data yang diambil. Ada 5 algoritma klasfikasi yang digunakan yaitu : Decision Tree (C4.5), k-Nearest Neighbor (k-NN), Rule Induction, Naïve Bayes, dan Gradient Boosted Trees. Semua algoritma yang dipakai untuk klasifikasi meningkat atau menurunnya pengguna KB pada wanita berusia 15-49 tahun.

Isi dari penelitian ini terdiri dari 7 bagian, bagian 1 yaitu abstrak, yang di dalamnya terdapat isi mengenai objek, masalah, metode yang digunakan dan hasil. Bagian 2 yaitu pendahuluan berisi latar belakang masalah penelitian dan struktur paper. Bagian 3 mengenai *related works* (penelitian sebelumnya yang berkaitan dengan tema penelitian ini), bagian 4 memberikan deskripsi singkat dari algoritma klasifikasi yang digunakan dalam penelitian ini kemudian bagian 4 berisi teorical foundation yaitu definisi dari kriteeria kinerja yang digunakan untuk evaluasi model terhadap dataset, bagian 5 berisi motode yang diusulkan, bagian 6 membahas hasil ekperimen dan bagian 7 adalah kesimpulan dari penelitian ini.

2. TINJAUAN PUSTAKA

Tinjauan pustaka digunakan untuk mengetahui landasan dari judul yang diangkat dalam penelitian.

2.1. *Data Mining*

Data mining merupakan proses untuk memanipulasi data dengan mengekstraksi informasi yang sebelumnya tidak diketahui dari dataset yang berukuran besar (Vijayakumar & Nedunchezhian, 2012). Beberapa aplikasi *data mining* fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang meggambarkan apa yang terjadi di masa lalu".

Data Mining atau sering juga disebut Knowledge Discovery in Database (KDD) adalah sebuah bidang ilmu yang banyak membahas tentang pola sebuah data.Serangkaian proses guna mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining (Witten & Frank, 2005).

2.2. *Rule Induction* (LOG)

Rule induction adalah ide – ide dari regresi linear berganda dengan situasi di mana variable dependen, y, diskrit dalam *Rule Induction*(K.S & R.V, 2014). Tidak ada asumsi yang dibuat tentang distribusi variable independen. Rule Induction mencoba untuk memperkirakan probabilitas posterios dari sampel x.

$$p(y=k \mid x)=\frac{\exp(-(w_{k_0}+w_k{}^{*}x))}{1+\sum_{l=1}^{K-1}\exp(-(w_{l_0}+w_l{}^{\mathrm{T}}x))},\ k=1,\ldots,K-1,$$

and

$$p(y=K \mid x)=\frac{1}{1+\sum_{l=1}^{K-1}\exp(-(w_{l_0}+w_l{}^{\mathrm{T}}x))}.$$

Gambar 1. Algoritma *Rule Induction*

2.3. Naïve Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang 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 (Purwanto & Darmadi, 2018)

Klasifikasi bayesian memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*."

Bayes rule digunakan untuk menghitung probabilitas suatu class. Algoritma Naive Bayes memberikan suatu cara mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula yang dapat digunakan

untuk menghitung peluang dari tiap kemungkinan yang terjadi. Bentuk umum dari teorema bayes seperti dibawah ini.
Dimana:
X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik.
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H) : Probabilitas hipotesis H (prior probability)
P(X|H) : Probabilitas X berdasar kondisi pada hipotesis H
P(X) : Probabilitas dari X

2.4. *K-Nearest Neighbour*

Algoritma K-NN adalah suatu metode yang menggunakan algoritma supervised. Perbedaan antara supervised learning dengan unsupervised learning adalah pada supervised learning bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada unsupervised learning, data belum memiliki pola apapun, dan tujuan unsupervised learning untuk menemukan pola dalam sebuah data. Tujuan dari algoritma k-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada k-NN. Pada proses pengklasifikasian, algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Algoritma k-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru. Jarak yang digunakan adalah jarak Euclidean Distance. Jarak Euclidean adalah jarak yang paling umum digunakan pada data numerik. Euclidean distance didefinisikan sebagai berikut
Keterangan :
d(xi,xj) : Jarak Euclidean (Eucledean Distance).
(xi) : record ke- i
(xj) : record ke- j
a(r) : data ke-r
i,j :1,2,3,…n
Algoritma k-NN adalah algoritma yang menentukan nilai jarak pada pengujian data testing dengan data training berdasarkan nilai terkecil dari nilai ketetanggaan terdekat [...] didefinisikan sebagi berikut:

$$D_{nn}(C_1, C_2) = min_{1 \le i \le r, 1 \le j \le s} d(y_i, z_j)$$

2.5. *Pohon Keputusan (DT) Classifier*

Pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menrapkan serangkaian aturan keputusan(Rahayuningsih & Maulana, 2018). Sebuah pohon keputusan mengklasifikasikan sampel secara *top-down*, mulai dari simpul akar dan menjaga bergerak sesuai dengan hasil dari tes di internal node, sampai simpul daun tercapai dan label kelas ditugaskan".

2.6. *Rule Induction*

Rule induction adalah salah satu teknik yang paling penting dari machine learning. Data sering disajikan dalam rules, rule induction adalah salah satu alat dasar data mining.

Beberapa sistem rule induction akan menginduksi aturan yang lebih kompleks, di mana nilai-nilai atribut dapat dinyatakan dengan negasi dari beberapa nilai atau nilai subset dari domain atribut.

2.7. *Gradient Boosted Trees*

Gradient Boosted Trees adalah teknik machine learning untuk masalah regresi dan klasifikasi, yang menghasilkan model prediksi dalam bentuk sebuah ensemble dari model prediksi yang lemah, biasanya pohon keputusan. Metode *Gradient Boosted Trees* mengasumsikan sebuah nilai real bernilai y.

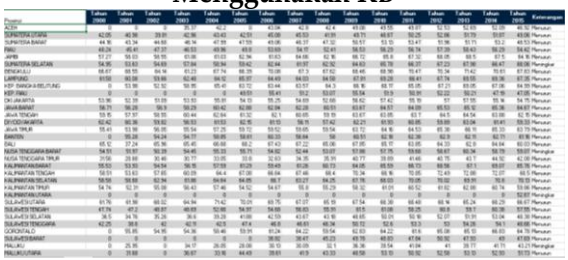
3. METODOLOGI

Kerangka yang diusulkan terdiri dari 1) dataset 2) klasifikasi algoritma, 3) validasi model 4) evaluasi model yang dan 5) model perbandingan.

3.1 *Dataset*

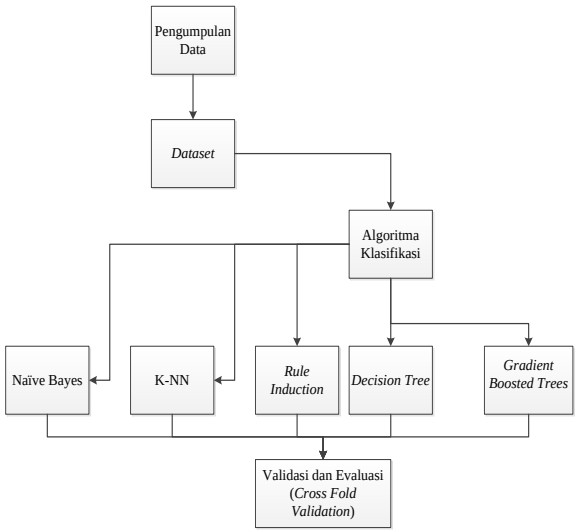
Dataset yang digunakan yaitu dataset Persantese Wanita usia 15 – 49 Tahun dan sudah Menikah yang sedang memakai/menggunakan Alat KB.

Tabel 1. Dataset Wanita Berusia15 – 49 tahun yang Berstatus Menikah Dan Sedang Menggunakan KB



(sumber: hasil survei pada halaman **www.bps.go.id**, ***update* terakhir: 18 February 2018)**

Berikut kerangka model metologi yang diusulkan pada penelitian ini:



Gambar 2. Kerangka Usulan Metodologi Analisa Komparasi Klasifikasi Persentase Wanita Usia 15-49 Tahun dan Sudah Menikah Menggunakan /Memakai Alat KB (Keluarga Berencana)

3.2 Algoritma Klasifikasi

Kerangka klasifikasi yang diusulkan bertujuan untuk membandingkan Persentase Wanita usia 15 – 49 Tahun dan sudah Menikah yang sedang memakai/menggunakan Alat KB tahun 2014-2015. Untuk tujuan penelitian ini, 5 algoritma untuk pengklasifikasi yang dipilih, Naive Bayes, *Rule Induction*, *Decision Tree*, *Gradient BoosteTrees* dan k-NN, stelah dihitung masing-masing algoritma yang nantinya akan didapatkan nilai akurasi untuk setiap *performance.*

3.3 Model Validasi

Penelitian ini menggunakan pengujian dengan 10-fold *cross-validation* untuk *data learning* dan *testing*. Artinya bahwa membagi data training menjadi 10 bagian yang sama dan kemudian melakukan data learning 10 kali. Hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa 10-*fold crossvalidation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat.

Tabel 2. *Stratified 10 Fold Cross Validation*

n-validation	Dataset's Partition									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

3.4 Model Evaluasi

Pada penelitian ini menerapkan *Area Under Curve* (AUC) sebagai akurasi Indikator dalam percobaan untuk mengevaluasi kinerja pengklasifikasi. AUC menjelaskan daerah di bawah kurva ROC menganjurkan penggunaan AUC untuk meningkatkan *crossstudy* komparatif.. AUC memiliki potensi untuk secara signifikan meningkatkan konvergensi pada eksperimen, karena memisahkan kinerja prediktif dari operasi kondisi, dan merupakan ukuran umum predictiveness.

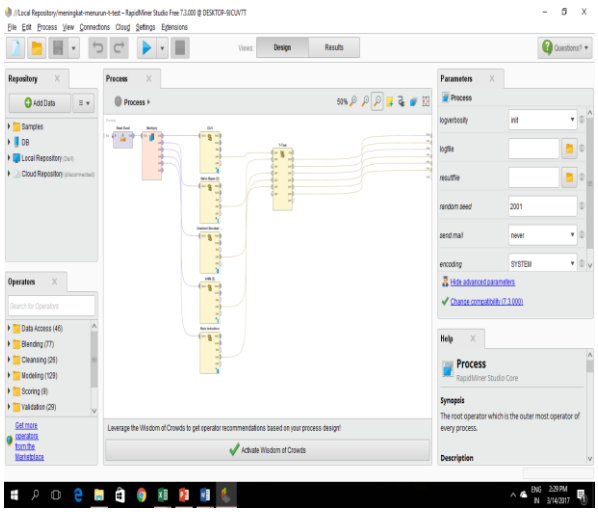
Evaluasi dalam penelitian ini adalah menggunakan uji t (t-test).

4. HASIL DAN PEMBAHASAN

Berikut adalah hasil dan pembahasan penelitian menggunakan algoritma klafisikasi terhadap Analisa Komparasi Klasifikasi Persentase Wanita Usia 15-49 Tahun dan Sudah Menikah Menggunakan/Memakai Alat KB (Keluarga Berencana).

4.1 Model Validasi

Model komparasi yang dilakukan dengan menggunakan metode parametrik. Metode parametrik itu sendiri menggunakan pengujian t-test untuk mendapatkan model terbaik dari pengujian yang dilakukan terhadap beberapa model klasifikasi tersebut.



Gambar 3. Model Komparasi menggunakan 5 Algoritma

Penelitian ini menggunakan platform Intel Celeron, 2 GB RAM dan Microsoft Windows 10 64-bit. Pengujiannya menggunakan RapidMiner 7.0.

Tabel 3. Hasil Eksperimen

	C4.5	K-NN	NB	RI	GB
Accuracy	87.50%	67.50%	25.83%	79.17%	25.83%
AUC	0.5	0.5	0.01	0.5	0.01

Hasil eksperimen yang disajikan adalah accuracy paling besar adalah algoritma C4.5 87.50 % dan AUC terbesarnya algorinya C4.5, K-NN, RI yaitu 0.5.

Tabel 4. Hasil Uji T-test

A	B	C	D	E	F
	0.875 +/- 0.155	0.792 +/- 0.180	0.675 +/- 0.228	0.875 +/- 0.155	0.258 +/- 0.216
0.875 +/- 0.155		0.281	0.034	1.000	0.000
0.792 +/- 0.180			0.220	0.281	0.000
0.675 +/- 0.228				0.034	0.001
0.875 +/- 0.155					0.000
0.258 +/- 0.216					

5. KESIMPULAN

Komparasi klasifikasi terhadap dataset tidaklah mudah untuk pemilihan algoritmanya, karena tidak semua tipe data dapat mendukung model algoritma meskipun model tersebut termasuk ke dalam klasifikasi.

Lima Model algoritma yang digunakan untuk komparasai diantaranya : *Decision Tree* (C4.5), *k-Nearest Neighbor* (k-NN), *Logistis Regression*, *Naïve Bayes*, dan *Gradient Boosted Trees*.

Nilai *accuracy* yang paling baik adalah algoritma C4.5 persentase 87.50%, nilai AUC sebesar 0.5, dapat disimpulkan untuk komparasi semua pengujian yang dilakukan terhadap *dataset*, maka model algoritma C4.5 bisa dikatakan lebih baik dibanding 4 model algoritma lainnya yang telah diuji coba.

Penelitian ini dapat berlanjut untuk algoritma klasifikasi lainnya ataupun ketika *dataset* sudah diperbaharui.

REEFERENSI

Awwalu, J., Ghazvini, A., & Abu Bakar, A. (2014). Performance Comparison of Data Mining Algorithms: A Case Study on Car Evaluation Dataset. *International Journal of Computer Trends and Technology*, *13*(2), 78–82. https://doi.org/10.14445/22312803/ijctt-v13p117

Kesehatan, Badan Pusat Statistik https://www.bps.go.id/subject/30/kesehatan.html (15 Maret 2019)

K.S, M., & R.V, K. (2014). Application of Data Mining Tools for Selected Scripts of Stock Market. *International Journal of Data Mining & Knowledge Management Process*, *4*(4), 55–63. https://doi.org/10.5121/ijdkp.2014.4405

Purwanto, A., & Darmadi, E. A. (2018). Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma. *Jurnal IKRAITH-INFORMATIKA*, *2*(1), 43–47.

Rahayuningsih, P. A., & Maulana, R. (2018). *Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner*. *VI*(1), 20–28.

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval*,

1(3), 153–172. https://doi.org/10.1007/s13735-012-0016-2

Witten, I. H., & Frank, E. (2005). Data Mining:Practical Machine Learning Tools and Techniques second edition. In *Morgan Kaufmann Publishers*. https://doi.org/0120884070, 9780120884070