Laman web jurnal: http://ejournal.stikom-db.ac.id/index.php/processor

Processor: Jurnal Ilmiah Sistem Informasi, Teknologi Informasi dan Sistem Komputer

P-ISSN: 1907-6738 | E-ISSN: 2528-0082

Perbandingan Algoritma Klasifikasi Data Kesejahteraan Sosial Kabupaten Bantul

Erfin Nur Rohma Khakim

Magister Teknologi Informasi Universitas Teknologi Yogyakarta, Bantul, Bantul, 55712, Indonesia

ABSTRACT

Today, the main problem in the economy sector in Indonesia is related to poverty alleviation. In Bantul Regency, poverty alleviation relies on the poverty data. The Social Welfare Integrated Data set by the Ministry of Social Affairs of the Republic of Indonesia is currently can not describe the classification of the poverty. It is either the causes of the delay in handling the poverty problem in Bantul Regency. Mapping of the poverty data has a major impact to the accuracy of targets for poverty reduction programs. One of the mappings that can be do for the problems of poverty data is by using the classification method. With this classification, the poverty data can be classified into several groups according to their circumstances. These groups include the very poor, the poor, and the vulnerable near poor. In this study, two classification methods will be tried, called Naive Bayes and K-nearest neighboor (KNN) to compare the best results based on a measurement method. Based on the experiments, the results showed that the Naïve Bayes performance was better than the Decision Tree algorithm with a significant difference, namely the alpha t-test that below 0.005. The Naïve Bayes algorithm has an accuracy of 65.55%, precision 44.16% and recall 51.39%. Meanwhile, for Decision Tree, accuracy is 70.01%, precision is 58.00% and recall is 50.02%.

Keywords: Poverty, data mining, classification, naïve bayes, k-nearest neighboor

ABSTRAK

Permasalahan utama dalam dunia ekonomi negara Indonesia saat ini adalah terkait dengan penanggulangan kemiskinan. Di Kabupaten Bantul, penanggulangan kemiskinan bergantung pada data warga miskin. Data Terpadu Kesejahteraan Sosial yang ditetapkan oleh Kementerian Sosial Republik Indonesia saat ini belum mampu menggambarkan pengkategorian dari warga miskin. Hal ini menjadi salah satu penyebab terlambatnya penanganan fakir miskin di Kabupaten Bantul. Pemetaan warga miskin berpengaruh besar terhadap ketepatan target program-program penanggulangan kemiskinan. Pemetaan yang dapat dilakukan terhadap data warga miskin ini salah satunya adalah dengan melakukan metode klasifikasi. Dengan klasifikasi ini, warga miskin dapat digolongkan ke dalam beberapa golongan sesuai dengan keadaan mereka. Golongan tersebut diantaranya golongan sangat miskin, miskin, dan hampir miskin. Dalam penelitian ini akan dicoba dua metode klasifikasi yaitu menggunakan Naive Bayes dan Decision Treeuntuk membandingkan hasil terbaik berdasarkan suatu metode pengukuran. Berdasarkan eksperimen yang dilakukan, didapatkan hasil performa Naïve Bayes lebih baik daripada algoritma Decision Tree dengan perbedaan yang cukup siginifikan, yaitu alpha t-tes dibawah 0,005. Algoritma Naïve Bayes memiliki accuracy 65,55%, precicion 44,16% dan recall 51,39%. Sedangkan untuk Decision Tree didapatkan accuracy 70,01%, precicion 58,00% dan recall 50,02%.

Kata Kunci: Kemiskinan, penggalian data, klasifikasi, naïve bayes, k-nearest neighboor

1. PENDAHULUAN

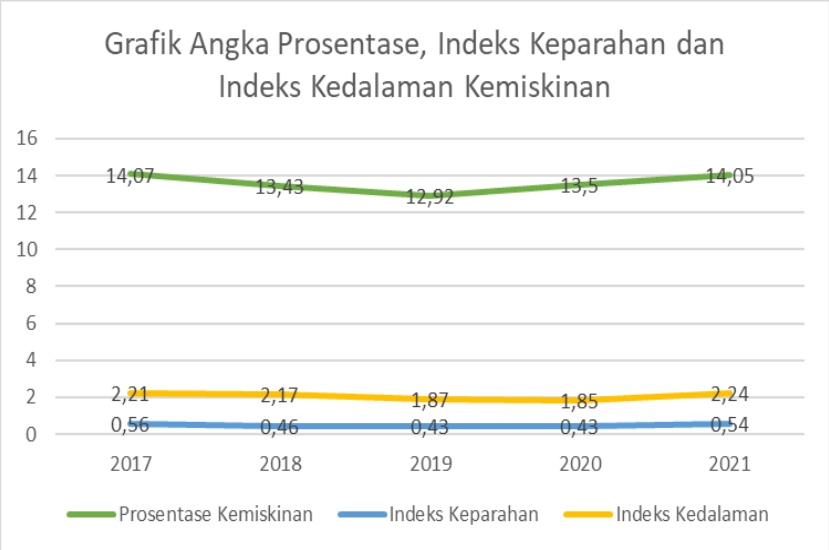
Kemiskinan merupakan permasalahan utama dalam bidang ekonomi di negara Indonesia [1]. Ketimpangan ekonomi di setiap daerah menjadi pemicu adanya warga miskin dan warga sejahtera. Pemerintah Indonesia telah melakukan berbagai cara untuk mengentaskan kemiskinan di Indonesia. Melalui Kementerian Sosial Republik Indonesia sebagai leading sector pengentasan kemiskinan telah meluncurkan berbagai program baik yang berupa bantuan sosial maupun yang bersifat pemberdayaan sosial demi menyelesaiakan masalah kemiskinan.

Masalah klasik dari langkah-langkah pengentasan kemiskinan ini justru terletak pada basis data warga miskin itu sendiri. Kekacauan data warga miskin menjadi penghambat utama program-program penanggulangan kemsikinan. Tidak sedikit warga miskin yang seharusnya masuk ke dalam data warga miskin justru tidak masuk, begitu juga sebaliknya, banyak warga yang sudah sejahtera dan tergolong tidak miskin justru masih terdaftar sebagai warga miskin. Hal ini tentu saja sangat berpengaruh terhadap ketepatan sasaran program-program kemiskinan. Anggaran besar yang dikeluarkan untuk program-program penanggulangan kemiskinan menjadi sia-sia tatkala program tersebut diberikan kepada orang yang salah. Atau sebaliknya, orang yang seharusnya mendapat program pengentasan kemiskinan justru tidak pernah mendapat manfaat dari program tersebut.

Kabupaten Bantul sebagai salah satu kabupaten di negara Indonesia juga tidak luput dari masalah kemiskinan. Selain program-program pengentasan kemiskinan yang dikeluarkan oleh Kementerian Sosial melalui Anggaran Pendapatan dan Belanja Negara, Kabupaten Bantul juga mengeluarkan program-program pengentasan kemiskinan melalui Anggaran Pendapatan dan Belanja Daerah Kabupaten. Selain digunakan sebagai pemberian program bantuan dan pemberdayaan kepada warga miskin, anggaran tersebut melalui Dinas Sosial Kabupaten Bantul juga digunakan untuk menangani permasalah data kemiskinan. Bahkan saat ini, Dinas Sosial Kabupaten Bantul sedang fokus menangani pada permasalahan data warga miskin yang ada di Kabupaten Bantul.

Jumlah total penduduk di Kabupaten Bantul adalah 956.513 pada Semester II Tahun 2021 [2]. Sedangkan prosentase angka kemiskinan Kabupaten Bantul tahun 2021 menurut Badan Pusat Statistika Kabupaten Bantul berada di angka 14,05 %, naik sebesar

0.55 % dari tahun 2020 [2]. Selain angka kemiskinan, indeks keparahan dan kedalam kemiskinan dari Kabupaten Bantul masih cukup tinggi. Adapun grafik angka kemiskinan dari tahun ke tahun digambarkan dalam fisualisasi pada gambar 1 berikut.

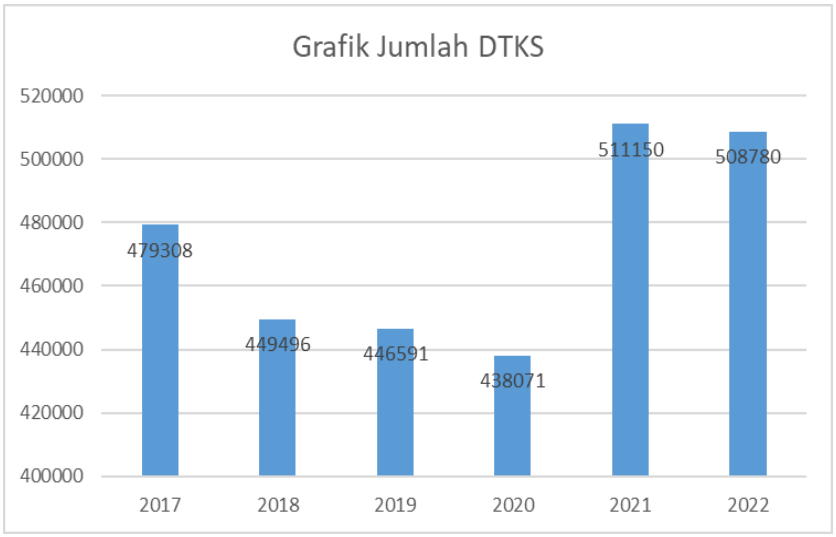


Gambar 1. Grafik Angka, Indeks Keparahan dan Kedalaman Kemiskinan [3]

Gambar 1 untuk sumbu X menunjukkan tahun sedangkan sumbu Y menunjukkan angka indeks. Dari gambar 1 tersebut, dapat dilihat dari tahun ke tahun tidak ada perubahan yang berarti dari angka, indeks keparahan dan indeks kedalaman kemiskinan di Kabupaten Bantul. Grafik pada gambar 1 di atas juga menunjukkan bahwa pada beberapa tahun terakhir tingkat ketimpangan kemiskinan di Kabupaten Bantul masih tinggi. Dibandingkan dengan kabupaten atau kota lain di wilayah Daerah Istimewa Yogayakarta.

Sebagai langkah utama program-program penanggulangan kemiskinan adalah mengetahui target calon penerima. Disinilah data kemiskinan mengambil andil paling penting. Bagaimana data kemiskinan yang ada dalam suatu daerah dijadikan sebagai pendukung keputusan untuk kebijakan-kebijakan pengentasan kemiskinan. Bagaimana suatu daerah dapat mengolah dan menyajikan data kemiskinan menjadi salah satu sistem rekomendasi kebijakan.

Data kemiskinan menurut terakhir Data Terpadu Kesejahteraan Sosial (DTKS) Kabupaten Bantul pada tahun 2022 adalah 508.780. Setiap bulan, data tersebut mengalami perubahan sesuai dengan yang ditetapkan oleh Kementerian Sosial Republik Indonesia. Adapun grafik naik turun DTKS Kabupaten Bantul tersebut difisualisasikan pada gambar 2 berikut.



Gambar 2. Grafik Jumlah DTKS

Gambar 2 untuk sumbu X menunjukkan tahun sedangkan untuk sumbu Y menunjukkan jumlah DTKS. Saat DTKS muncul pertama kali pada tahun 2015, struktur DTKS masih belum benar dan tidak sesuai dengan kaidah data yang baik. Namun mulai tahun 2017 DTKS mulai berbentuk lebih rapi dan sesuai dengan kaidah data yang benar. Dari garfik pada gamba 2, sejak 2017 jumlah DTKS berada pada kisaran 400 ribuan, namun mulai tahun 2021 DTKS bertambah menjadi sekitaran 500 ribuan. Sejak tahun 2018 sampai dengan tahun 2020, DTKS memiliki banyak atribut atau indikator sekaligus satu atribut label yang menentukan kategori warga miskin. Terdapat satu atribut bernama persentil yang berisi peringkat warga miskin berdasarkan perhitungan menggunakan suatu metode bernama Proxy Mean Test (PMT) terhdap indicator yang dimiliki dari warga miskin tersebut. Namun mulai tahun 2021 tepatnya setelah berganti kebijakan pada Kementerian Sosial RI, DTKS tidak lagi memiliki perangkingan. Atribut persentil hasil perhitungan metode PMT telah dihilangkan dan setiap daerah diminta untuk memiliki indikator masing-masing sesuai dengan angka garis kemiskinan di wilayah masing-masing.

Berdasarkan kebijakan baru mengenai DTKS itulah saat ini Kabupaten Bantul mulai bergerak untuk menyusun indikator kesejahteraan sosial sesuai dengan keadaan di wilayah Kabupaten Bantul sendiri. Kabupaten Bantul melalui Dinas Sosial Kabupaten Bantul bergerak untuk membuat label pengkategorian warga miskin dari DTKS yang telah ditetapkan oleh Kementerian Sosial.
DTKS terdiri dari atribut yang bias diperhitungkan untuk menentukan kriteria warga miskin, diantaranya adalah kepemilikan bangunan, luas lantai, jenis dinding, jenis atap, pendidikan tertinggi, jumlah tanggungan, kepemilikan asset, kepemilikan hewan ternak dan lain-lain. Adapun label dari kriteria tersebut akan dibagi ke dalam dua kelas untuk mempermudah penargetan program-program bantuan, yaitu sangat miskin, miskin dan hampir miskin.

Dari permasalah diatas dibutuhkan suatu analisa yang tepat untuk menentukan suatu keputusan. Oleh sebab itu, dalam penelitian ini akan melakukan analisa terhadap data yang diperoleh dengan indikator yang telah disebutkan diatas dengan menggunakan algoritma data mining metode klasifikasi. Metode klasifikasi dalam data mining banyak contohnya, *decision/classification trees*, *bayesian classifiers/naïve bayes classifiers*, *neural networks*, analisa statistik, algoritma genetika, *rough sets*, *k-nearest neighbor*, metode *rule based*, *memory based reasoning*, dan *support vector machines*. Untuk menentukan hasil terbaik, penelitian ini akan membandingkan dua dari banyak algoritma klasifikasi. Pengukuran terhadap tingkat akurasi dan presisi akan menentukan algoritma mana yang lebih baik dalam pengklasifikasian DTKS di Kabupaten Bantul.

2. TINJAUAN PUSTAKA

Data Terpadu Kesejahteraan Sosial adalah Data Terpadu Kesejahteraan Sosial adalah data induk yang berisi data pemerlu pelayanan kesejahteraan sosial, penerima bantuan dan pemberdayaan sosial, serta potensi dan sumber kesejahteraan sosial [4].

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui [5]. Algoritma klasifikasi yang banyak digunakan secara luas, yaitu *decision/classification trees*, *bayesian classifiers/ naïve bayes classifiers*, *neural networks*, analisa statistik, algoritma genetika, *rough sets*, *k-nearest neighbor*, metode *rule based*, *memory based reasoning*, dan *support vector machines* [6].

Metode klasifikasi Naïve Bayes adalah algoritma untuk memprediksi probabilitas keanggotaan suatu kelas [7]. Algoritma ini berdasar pada teorema Bayesian. Naïve bayes bekerja dengan menghitung keputusan dari setiap kelas berdasarkan probabilitasnya dengan syarat kelas keputusan adalah benar. Metode Bayes merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi [8]. Metode Bayes merupakan pendekatan statistik untuk melakukan inferensi induksi pada persoalan klasifikasi. Secara sederhana, Naive Bayes mengasumsikan bahwa kehadiran fitur tertentu di kelas tidak terkait dengan kehadiran fitur lainnya. Algoritma ini mudah dibuat dan sangat berguna apabila dihadapkan dengan dataset yang besar [9]. Kegunaan algoritma Naïve Bayes adalah untuk prediksi secara realtime, multi kelas, dapat mengklasifikasikan teks dan untuk sistem rekomendasi. Secara umum, rumus probabilitas Naïve Bayes dituliskan sebagai berikut

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)} \quad \text{……..(1)}$$

Metode klasifikasi lain selain Naive Bayes adalah Decision Tree. Decision Tree didefinisikan sebagai cara guna melakukan *forecasting*, memprediksi atau mengklarifikasi yang sangat kuat. Pada banyak penelitian, Decision Tree mampu memberikan hasil akurasi yang cukup tinggi sehingga algoritma ini lebih sering digunakan. Kelebihan lain dari metode ini adalah mampu mengeliminasi perhitungan atau data-data yang tidak diperlukan karena sampel yang ada biasanya hanya diuji berdasarkan kriteria atau kelas tertentu [10]. Secara umum, algoritma ini memilih atribut dengan *gain* tertinggi. Decision Tree ini dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama [11]. Adapun rumus *gain* dan *entrophy* ditunjukkan oleh persamaan 2 sebagai berikut

$$Gaint\ (S,A) = Entrophy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}$$

$$Entrophy\ (s) = \sum_{i=1}^{n} -pi * log2pi \quad \text{……...……..……(2)}$$

Rapidminer adalah perangkat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi [12]. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Sebelumnya, Rapidminer bernama YALE (Yet Another Learning Environment), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi.

Preprocessing yaitu teknik awal penggalian data untuk mengubah data mentah atau biasa dikenal dengan raw data yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya. Adapun tujuan utama dalam preprocessing data ini ialah sebagai berikut [13] : Pembersihan Data, yaitu mengisi nilai yang hilang, menghaluskan noise data, mengidentifikasi dan menghapus outlier serta menyelesaikan inkonsistensi. Selanjutnya, Integrasi Data, integrasi beberapa database, kubus data, atau file. Selain ituTransformasi Data, normalisasi dan agregasi. Selain itu juga ada Pengurangan Data,

memperoleh penurunan representasi dalam volume tetapi menghasilkan hasil analitis yang sama atau serupa dan yang terakhir Diskretisasi Data, pengurangan data namun sangat penting, terutama untuk data numerik.

Dalam penelitian ini digunakan dua algoritma untuk menentukan tingkat akurasi yang lebih tinggi. Pengukuran dilakukan dengan menghitung pada masing-masing algoritma yaitu *accuracy*, *recall* dan *precision*. Adapun rumus perhitungan *accuracy*, *recall* dan *precision* ditunjukkan pada persamaan 3 sebagai berikut

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP} \qquad \text{……..……..……(3)}$$

3. METODOLOGI PENELITIAN

3.1. Sumber Data penelitian

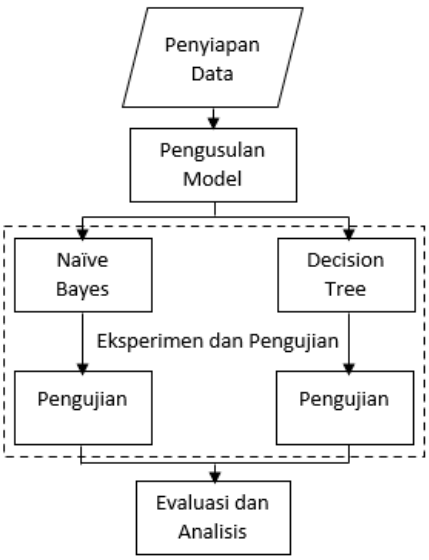
Sumber data yang digunakan dalam penelitian ini adalah data primer dari Dinas Sosial Kabupaten Bantul. Data ini berasal dari Data Terpadu Kesejahteraan Sosial (DTKS) Tahun 2020 yang ditetapkan oleh Kementerian Sosial Republik Indonesia dan diberikan kepada seluruh kabupaten/kota di wilayah Indonesia sesuai dengan wilayah masing-masing termasuk Kabupten Bantul. DTKS yang diambil bukanlah penetapan terbaru dikarenakan sejak tahun 2021 DTKS sudah tidak menyertakan indikator. Adapun metode pengumpulan data yang digunakan adalah wawancara dan meminta dataset langsung kepada Dinas Sosial Kabupaten Bantul.

3.2. Metodologi Penelitian

Penelitian ini menggunakan metode penelitian terapan deskriptif. Metode ini bertujuan untuk melakukan penerapan, melakukan pengujian serta melakukan evaluasi terhadap suatu teori algoritma dalam proses memecahkan suatu permasalahan. Penelitian ini adalah untuk melihat bagaimana pengkategorian dari data warga miskin yang dalam hal ini menggunakan DTKS. Hasil akhir dari penelitian ini diharapkan dapat menggambarkan klasifikasi dari ratusan ribu data yang ada dan metode klasifikasi seperti apa yang memiliki tingkat presisi dan akurasi paling tinggi.

3.3. Tahapan Penelitian

Tahapan penelitian terdiri dari penyiapan data, pengusulan model, eksperimen dan pengujian serta evaluasi dan analisis. Tahapan tersebut digambarkan dalam sebuah diagram yang ditunjukkan pada gambar 3 berikut



Gambar 1. Diagram Tahapan Penelitian

Tahapan pertama yang digunakan dalam penelitian ini adalah Penyiapan Data, data yang akan diuji adalah Data Terpadu Kesejateraan Sosial yang merupakan data induk yang berisi data pemerlu pelayanan kesejahteraan sosial, penerima bantuan dan pemberdayaan sosial, serta potensi dan sumber kesejahteraan sosial. DTKS terdiri dari atribut yang bisa diperhitungkan untuk menentukan kriteria warga miskin, atribut yang dipilih dalam penelitian ini diantaranya seperti ditunjukkan pada tabel 1 berikut.

Tabel 1. Indikator Kesejahteraan dalam DTKS

No	Indikator
1	Kepemilikan aset tidak bergerak
2	Status kepemilikan rumah
3	Jenis atap rumah
4	Jenis dinding rumah
5	Jenis lantai rumah
6	Luas lantai rumah
7	Kepemilikan rumah lain
8	Status bekerja
9	Jenis pekerjaan
10	Pendidikan tertinggi
11	Sumber air minum rumah tangga
12	Sumber penerangan rumah tangga
13	Bahan bakar memasak rumah tangga
14	Status Kesejahteraan

Pada tabel 1 diperlihatkan bahwa penelitian ini menggunakan 14 indikator dalam DTKS. Indikator perlu dipilih dari sekian banyak yang ada dalam DTKS dikarenakan ada beberapa indikator yang memiliki kemiripan dan cukup untuk diambil salah satu indikator saja.

DTKS yang selanjutnya disebut data warga miskin ini terbagi ke dalam data warga yang menerima bantuan sosial dan tidak. Sebelum dilakukan ujicoba menggunakan metode, data warga miskin yang memiliki total data 270.878 ini diolah terlebih dahulu sehingga didapatkan sejumlah 270.003. Pengolahan dilakukan untuk menghilangkan atau mengisi *missing value* agar tidak banyak data yang kosong dan juga menghilangkan data ganda sehingga proses ujicoba menggunakan metode menjadi lebih valid. Selain itu pengolah data yang biasa disebut *preprocessing data* ini juga berfungsi untuk memilih atribut yang akan digunakan.

Tahapan kedua adalah Pengusulan Model. Metode yang diusulkan dalam penelitian ini adalah metode klasifikasi dengan memilih dua algoritma pengujian yaitu Naive Bayes dan Decision Tree. Dua algoritma ini dipilih karena menurut banyak penelitian, kedua algoritma ini memiliki nilai akurasi yang lebih tingggi. Dari kedua algoritma tersebut akan dibandingkan algoritma mana yang lebih baik dalam mengkategorikan data warga miskin menjadi kelas layak dan tidak dengan suatu metode pengukuran akurasi. Naive Bayes adalah algoritma untuk memprediksi probabilitas keanggotaan suatu kelas. Sedangkan Decision Tree secara umum didefinisikan sebagai algoritma yang memilih atribut dengan gain tertinggi.

Tahapan selanjutnya adalah ekserimen dan pengujian. Eksperimen dilakukan terhadap data warga miskin dengan menggunakan dua metode yang diusulkan. Eksperimen ini menggunakan perangkat lunak Rapidminer yang merupakan perangkat lunak untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi. Rapidminer sangat mudah digunakan karena hanya perlu melakukan *drag and drop* operator yang telah disediakan. Sebelum dilakukan eksperimen menggunakan kedua model yang diusulkan, terlebih dahulu dilakukan preprocessing data, yaitu mengubah data mentah atau biasa dikenal dengan *raw data* yang dikumpulkan dari berbagai sumber menjadi informasi yang lebih bersih dan bisa digunakan untuk pengolahan selanjutnya. Untuk mendapatkan hasil terbaik dengan performa tertinggi maka perlu digunakan K-fold agar diuji coba setiap variasi jumlah *data training* dan *data testing*. K-Fold *Cross Validation* adalah salah satu dari jenis pengujian *cross validation* yang berfungsi untuk menilai kinerja proses sebuah metode algoritma dengan membagi sampel data secara acak dan mengelompokkan data tersebut sebanyak nilai K [14]. Setelah dilakukan eksperimen terhadap data tersebut di atas, Langkah selanjutnya adalah menguji masing-masing hasil dari kedua algoritma tersebut di atas. Pengujian dilakukan untuk mengukur *accuracy*, *recall* dan *precision*.

Tahapan terakhir dari penelitian ini adalah Evaluasi dan Analisis. Pada tahap ini akan dievaluasi bagaimana perbandingan hasil akurasi dari kedua algoritma yang digunakan. Setelah itu akan didapat analisis algoritma mana yang lebih tinggi dan algoritma mana yang lebih baik untuk diterapkan dalam mengklasifikasi data warga miskin. Adapun evaluasi pada penelitian ini menggunakan T-Test untuk menilai signifikansi perbedaan performa atau pengujian dari kedua algoritma yang digunakan.

4. HASIL DAN PEMBAHASAN

4.1. Data Preprocessing

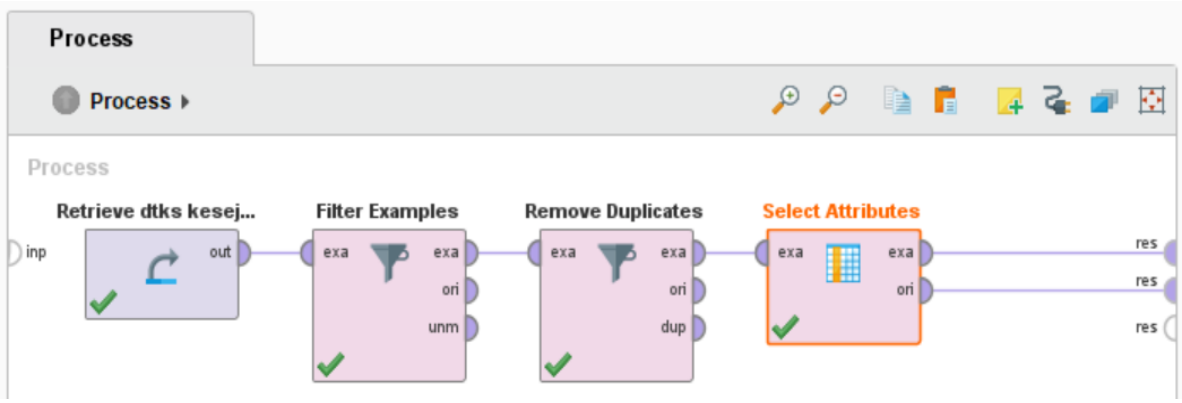
Dataset yang akan digunakan dalam penelitian ini adalah Data Terpadu Kesejahteraan Sosial yang ditetapkam oleh Kementerian Sosial Republik Indonesia pada bulan Februari 2022 dengan mengacu pada indikator Data Terpadu Kesejahteraan Sosial yang ditetapka pada bulan Oktober 2020. Adapun jumlah keseluruhan baris adalah 509.780 baris dengan 238.902 baris yang memiliki indikator kosong, sehingga total data yang bisa diimplementasikan ke dalam penelitian adalah 270.878 baris. Sedangkan untuk dimensi kolom dari dataset terdapat 39 kolom regular dan 1 kolom label dengan 7 kolom berisi data induk dan selebihnya adalah data indikator. Dimensi dataset tersebut ditampilkan ke dalam fisualisasi seperti pada gambar 4 berikut.

Row No.	Sta_kesejah...	NO URUT	ID DTKS	KECAMATAN	DESA/KELU...	DUSUN	RT	NOKK
1	Sangat Miskin	1	7D8B8805-4...	BANTUL	PALBAPANG	BALAI JANGGO	?	1304052111...
2	Sangat Miskin	2	6D962C0F-1...	JETIS	CANDEN	JL.BALAI BAR...	1	1371091602...
3	Sangat Miskin	3	0B01374B-8...	BANTUL	PALBAPANG	PELANDUK	4	1406060107...
4	Sangat Miskin	4	B119853A-1...	BANTUL	PALBAPANG	PELANDUK	4	1406060107...
5	Hampir Miskin	5	6EFF5FD0-9...	IMOGIRI	SELOPAMIO...	SUNGAI SIRIH	11	1409032203...
6	Sangat Miskin	6	2F2EED17-F...	IMOGIRI	SELOPAMIO...	SUNGAI SIRIH	11	1409032203...
7	Sangat Miskin	7	DE737937-3...	IMOGIRI	SELOPAMIO...	SUNGAI SIRIH	11	1409032203...
8	Sangat Miskin	8	DA604841-A...	DLINGO	MANGUNAN	JL.HIBRIDA	19	1502080306...
9	Sangat Miskin	9	F9DF8141-8...	SRANDAKAN	TRIMURTI	TAMBANG E...	2	1502202510...
10	Hampir Miskin	10	1DD91831-0...	DLINGO	MANGUNAN	MUARA KUM...	16	1505061111...
11	Sangat Miskin	11	56993A1E-4...	DLINGO	MANGUNAN	MUARA KUM...	16	1505061111...
12	Hampir Miskin	12	425BB637-4...	SANDEN	SRIGADING	KOMP MEGA...	4	1607100711...
13	Hampir Miskin	13	9FBD9D15-7...	PIYUNGAN	SRIMULYO	JL. URIP SU...	14	1671061802...
14	Sangat Miskin	14	2FC13CB3-9...	DLINGO	MANGUNAN	JATI BARU	5	1801050911...
15	Sangat Miskin	15	80073217-1F...	BANGUNTAP...	POTORONO	BUMI JAYA	15	1802212010...

ExampleSet (270,878 examples, 1 special attribute, 39 regular attributes)

Gambar 4. Tabel Data Sebelum *Preprocessing*

Perangkat lunak yang akan digunakan dalam penelitian ini adalah rapidminer. Rapidminer adalah perangkat lunak yang dapat digunakan untuk pemrosesan dan analisis data besar. Tersedia banyak operator dalam rapidminer yang dapat digunakan seperti diantaranya metode klasifikasi dan klasterisasi. Dengan rapidminer, penelitian ini akan menggunakan operator classification naïve bayes dan decision tree. Sebelum dilakukan pemrosesan data menggunakan metode klasifikasi, terlebih dahulu dilakukan *preprocessing* data. *Preprocessing* data yang akan dilakukan adalah pembersihan (*data cleaning*) dan transformasi (*data transformation*). Pembersihan data mencakup penghapusan data kosong dan penghapusan data duplikat. Sedangkan transformasi data dilakukan dengan menormalkan data (*normalize*). Gambar 4 tersebut menunjukkan bahwa seblum dilakukan preprocessing data, jumlah baris adalah 270.878 dengan 40 atribut. *Preprocessing* yang dilakukan menggunakan rapidminer seperti terdapat pada gambar 5 berikut.



Gambar 5. Operator *Preprocessing*

Dari gambar 5 di atas dapat dilihat bahwa *preprocessing* dilakukan dengan *filter examples* untuk mendapatkan nilai yang *not missing*, *remove duplicates* untuk menghilangkan jika ada baris ganda dan select attributes untuk memilih 14 atribut dari 40 atribut yang tersedia. Dari *preprocessing* tersebut didapatkan bahwa data berkurang karena *preprocessing*, ditunjukkan pada gambar 6.

Row No.	Sta_kesejah...	sta_bangun...	luas_lantai	lantai	dinding	atap	sumber_air...	sumber_pe...
1	Sangat Miskin	1	1	1	1	1	1	1
2	Sangat Miskin	1	1	1	1	1	1	1
3	Sangat Miskin	1	1	1	1	1	1	1
4	Hampir Miskin	1	1	1	1	1	1	1
5	Sangat Miskin	1	1	1	1	1	1	1
6	Sangat Miskin	1	1	1	1	1	1	1
7	Sangat Miskin	1	1	1	1	1	1	1
8	Sangat Miskin	1	1	1	1	1	1	1
9	Hampir Miskin	1	1	1	1	1	1	1
10	Sangat Miskin	1	1	1	1	1	1	1
11	Hampir Miskin	1	1	1	1	1	1	1
12	Hampir Miskin	1	1	1	1	1	1	1
13	Sangat Miskin	1	1	1	1	1	1	1
14	Sangat Miskin	1	1	1	1	1	1	1
15	Sangat Miskin	1	1	1	1	1	1	1

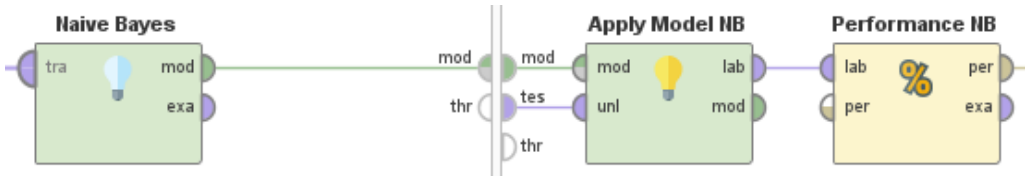
ExampleSet (270,003 examples, 1 special attribute, 13 regular attributes)

Gambar 6. Tabel Data Setelah *Preprocessing*

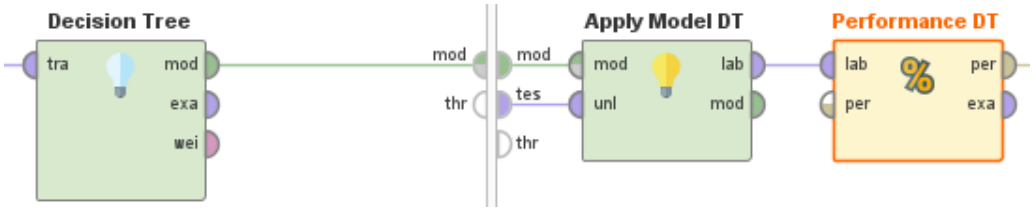
Dari preprocessing yang dilakukan, didapatkan hasil seperti pada gambar 6. Jumlah berkurang dari 270.878 menjadi ganya 270.003. Sedangkan select attributes mengurangi jumlah atribut dari 40 menjadi hanya 14 atibut dengan 13 atribut reguler dan 1 atibut kelas.

4.2. Data Preprocessing

Setelah dilakukan penyiapan data. Maka langkah selanjutnya adalah masuk ke pemrosesan data. Pemrosesan data dilakukan dengan dua model, yaitu Naive Bayes dan Decision Tree seperti ditunjukkan pada gambar 7 dan 8.



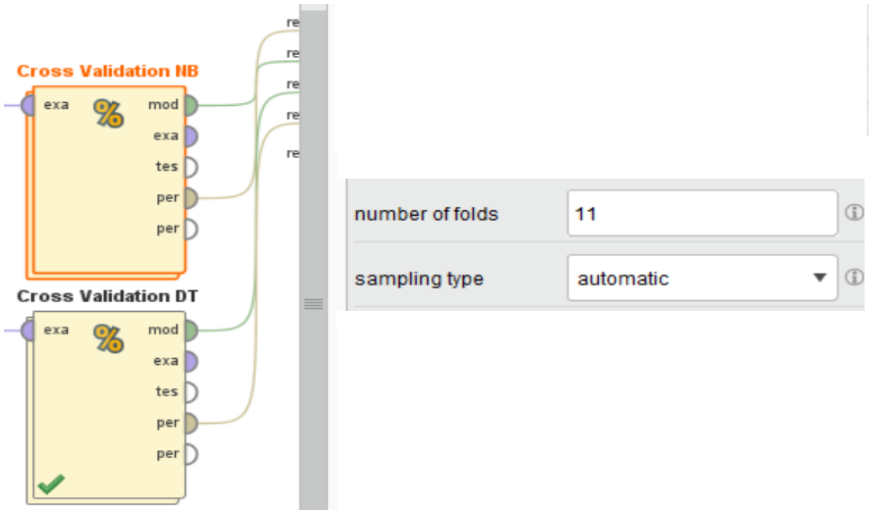
Gambar 7. Operator Algoritma Naïve Bayes



Gambar 8. Operator Algoritma Decision Tree

Dari gambar 7 dan 8 dapat kita lihat bahwa kondisi data dibuat sama agar seimbang ketika dibandingkan hasilnya. *Preprocessing data* pun juga dibuat sama, satu data hasil *preprocesing* digunakan untuk dua algoritma dengan operator *multiply*.

Untuk mendapatkan hasil yang terbaik, operator *cross validation* dibutuhkan untuk membuat sejumlah K-fold. K-fold ini membagi sampel data secara acak dan mengelompokkan data tersebut sebanyak nilai K seperti ditunjukkan pada gambar 9.



Gambar 9. Operator Cross Validation untuk K-Fold

Pada gambar 9 tersebut, contohnya adalah kedua model ditentukan 11 kali iterasi dengan kelompok data berbeda-beda. Jumlah K (iterasi) dibuat sama diantara dua model untuk membuat kondisi awal yang sama sebelum pengujian. Diharapkan dengan eksperimen beberapa jumlah K ini maka bisa didapatkan hasil yang terbaik.

4.3. Pengujian Performa

Pengujian performa dari kedua algoritma yang diterapkan juga menggunakan operator dari Rapidminer yaitu operator *performance*. Pengujian terhadap algoritma Naive Bayes ditunjukkan pada gambar 10.

accuracy: 65.55% +/- 0.17% (micro average: 65.55%)

	true Sangat Miskin	true Hampir Miskin	true Miskin	class precision
pred. Sangat Miskin	172541	34281	38215	70.41%
pred. Hampir Miskin	16471	4449	4046	17.82%
pred. Miskin	0	0	0	0.00%
class recall	91.29%	11.49%	0.00%	

Gambar 10. Hasil Performa Algoritma Naïve Bayes

Dari gambar 10 tersebut, dapat dilihat bahwa hasil K=11 dari algoritma Naive Bayes sebesar didapatkan hasil *accuracy* sebesar 65,55%, *precicion* 44,16% dan *recall* 51,39%. Selanjutnya, untuk pengujian terhadap algoritma Decision Tree ditunjukkan pada gambar 11.

accuracy: 70.01% +/- 0.01% (micro average: 70.01%)

	true Sangat Miskin	true Hampir Miskin	true Miskin	class precision
pred. Sangat Miskin	189002	38716	42260	70.01%
pred. Hampir Miskin	10	14	1	56.00%
pred. Miskin	0	0	0	0.00%
class recall	99.99%	0.04%	0.00%	

Gambar 11. Hasil Performa Algoritma Decision Tree

Dari gambar 11 tersebut, dapat dilihat bahwa hasil hasil K=11 dari algoritma Decision Tree sebesar didapatkan hasil *accuracy* sebesar 70,01%, *precicion* 58,00% dan *recall* 50,02%. Lebih lengkapnya, perbandingan setiap K yang ditentukan ditunjukkan dalam tabel 2 berikut.

Tabel 2. Hasil Performa Setiap Jumlah K

K	Naive Bayes			Decision Tree		
	Acc	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
1	65,50	44,06	51,36	70,00	54,91	50,02
4	65,56	44,14	51,40	70,01	63,85	50,04

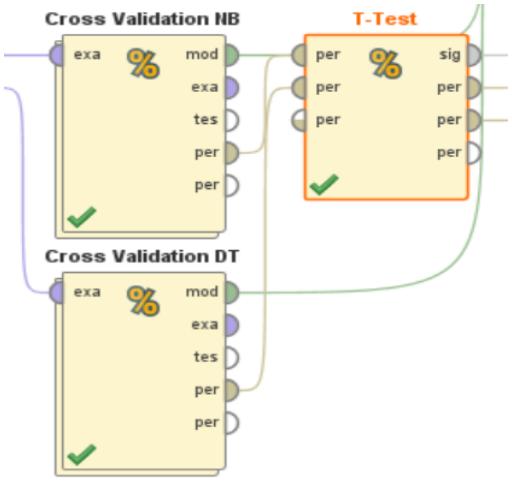
6	65,56	44,11	51,39	70,01	69,73	50,03
8	65,55	44,13	51,34	70,01	76,39	50,03
10	65,55	44,11	51,48	70,01	62,51	50,02
12	65,55	44,13	51,34	70,00	61,67	50,01

Dari tabel 2 tersebut, terlihat bahwa nilai tertinggi performa secara keseluruhan dari masing-masing algoritma baik Naive Bayes maupun Decsision Tree adalah pada saat dilakukan *cross validation* pada K=8. Sedangkan setelah K=8 maka sudah tidak ada lagi hasil performa yang lebih baik.

4.4. Evaluasi dan Analisis

Secara menyeluruh, performa dari klasifikasi menggunakan algoritma Decision Tree lebih baik dari algoritma Naive Bayes terutama pada K=9. Untuk algoritma Naive Bayes didapatkan *accuracy* 65,55%, *precicion* 44,16% dan *recall* 51,39%. Sedangkan untuk Decision Tree didapatkan *accuracy* 70,01%, *precicion* 58,00% dan *recall* 50,02%.

Untuk membandingkan performa dari kedua algoritma dalam Rapidminer digunakan operator T-Test seperti ditunjukkan pada gambar 12 berikut.



Gambar 12. Operator *Cross Validation*

Dari gambar 12 tersebut, operator T-Test dari hasil *cross validation* algoritma Naive Bayes dan Decision Tree akan membandingkan kedua performa dari algoritma secara bersamaan. Hasil dari T-Test ditunjukkan pada tabel 3.

Tabel 3. Hasil T-Test

	Naive Bayes	Decision Tree
	0,656 +/- 0,001	0,700 +/- 0,000
0,656 +/- 0,001		0,000
0,700 +/- 0,000		

Alpha yang dapat diterima dalam T-test adalah yang di bawah 0,05. Tabel 3 tersebut menunjukkan bahwa nilai signifikansi berada di bawah alpha 0,05. Hal ini menggambarkan bahwa kedua algoritma memiliki perbedaan performa yang cukup signifikan dengan decison tree yang jauh lebih baik.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Dari hasil eksperimen dapat disimpulkan bahwa performa yang paling baik dalam mengklasifikasikan Data Terpadu Kesejahteraan Sosial adalah algoritma Naïve Bayes. Hal ini dibuktikan dengan nilai *accuracy* dan *recall* dari algoritma Naïve Bayes yang lebih tinggi dari Decision Tree, meskipun nilai *precision* dari algoritma Decision Tree lebih tinggi daripada Naïve Bayes. Algoritma Naïve Bayes memiliki *accuracy* 65,55%, *precicion* 44,16% dan *recall* 51,39%. Sedangkan untuk Decision Tree didapatkan *accuracy* 70,01%, *precicion* 58,00% dan *recall* 50,02%. Hal ini juga membuktikan bahwa kinerja algoritma Decision Tree masih tetap unggul ketika pengujian dilakukan pada tipe data kategori dibandingkan dengan Naïve Bayes. Berdasarkan evaluasi menggunakan T-test

terhadap perbandingan dari kedua algoritma tersebut di atas didapatkan nilai alpha 0,000. Dimana jika nilai alpha kurang dari 0,05 maka perbedaan performa dari dua algoritma atau lebih dianggap cukup signifikan.

5.2. Saran

Saran untuk penelitian lebih lanjut adalah mencari nilai performa yang lebih tinggi dengan mencoba mengkombinasikan atribut yang dipilih, baik itu mengurangi maupun menambah atribut yang digunakan. Selain itu juga melakukan preprocesing data berupa *normalize* untuk mendapatkan nilai dengan rentang yang tidak terlalu jauh.

DAFTAR PUSTAKA

[1] F. Fajriwati, "Dampak Perekonomian Terhadap Masyarakat Miskin Di Lingkungan Kampung Nelayan Kecamatan Medan Labuhan," *Ekon. J. Ilmu Ekon. dan Stud. Pembang.*, vol. 16, no. 2, pp. 145–154, 2016, doi: 10.30596/ekonomikawan.v16i2.942.

[2] K. DKB Ditjen Dukcapil, "Statistik Penduduk DIY," *Biro Tata Pemerintahan Setda DIY*, 2020. https://kependudukan.jogjaprov.go.id/statistik/penduduk/jumlahpenduduk/14/0/12/04/.clear

[3] B. BPS, "Tabel Kemiskinan," *BPS Kab Bantul*, 2021. https://bantulkab.bps.go.id/subject/23/kemiskinan.html#subjekViewTab3

[4] R. Kemensos, "Peraturan Menteri Sosial tentang Pengelolaan Data Terpadu Kesejahteraan Sosial," *BN. 2021 No. 578, jdih.kemensos.go.id*, vol. 4, no. 1, pp. 1–2, 2021, [Online]. Available: https://peraturan.bpk.go.id/Home/Details/171535/permensos-no-3-tahun-2021

[5] Kusrini and E. T. Luthfi, *Algoritma Data Mining*. Andi Offset, 2009.

[6] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," *OJS Unisbank J.*, 2013.

[7] A. Wanto *et al.*, *Data Mining : Algoritma dan Implementasi*. Yayasam Kita Menulis, 2020.

[8] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018, doi: 10.33096/ilkom.v10i2.303.160-165.

[9] A. Gravita, "Mengenal Algoritma Naive Bayes dan Kegunaannya," *PT Semua Mahir Teknologi (SMART)*, 2022. https://codingstudio.id/algoritma-naive-bayes/

[10] A. maulana Ismail, "Cara Kerja Algoritma k-Nearest Neighboor," *Bee Solution Partners*, 2018. https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e

[11] A. Khairi, A. F. Ghozali, and A. D. N. Hidayah, "Implementasi K-Nearest Neighbor (KNN) untuk Mengklasifikasi Masyarakat Pra-Sejahtera Desa Sapikerep Kecamatan Sukapura," *TRILOGI J. Ilmu Teknol. Kesehatan, dan Hum.*, vol. 2, no. 3, pp. 319–323, 2021, doi: 10.33650/trilogi.v2i3.2878.

[12] A. Umar, "Rapidminer, Definisi dan Fitur-fiturnya," 2021. https://www.abdumar.com/2021/03/rapidminer-definisi-dan-fitur-fiturnya.html?m=1

[13] Alfarisi, "Data Preprocessing - Konsep Pembelajaran Data Mining," *Steemit*, 2017. https://steemit.com/education/@alfarisi/data-preprocessing-konsep-pembelajaran-data-mining

[14] T. & I. Hutapea, "Penerapan Algoritma Modified K-Nearest Neighbour Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3957–3961, 2018.

NOMENKLATUR

X	Data dengan class yang belum diketahui
H	Hipotesis data X merupakan suatu class spesifik
$P(H\|X)$	Probabilitas hipotesis H berdasarkan kondisi x (posteriori probabilitas)
$P(X\|H)$	Probabilitas X berdasarkan kondisi
$P(H)$	Probabilitas hipotesis H (prior probabilitas)
$P(X)$	Probabilitas dari X
$\boldsymbol{S}$	himpunan kasus
$\boldsymbol{A}$	atribut atau fitur
$\boldsymbol{n}$	jumlah partisi atribut/sampel A
$\boldsymbol{\|S_i\|}$	jumlah kasus pada pasrtisi ke-I
$\boldsymbol{\|S\|}$	jumlah kasus dalam S dan pi yang merupakan proporsi dari Si terhadap S
$\boldsymbol{x_{ij}}$	data sampel pengetahuan
$\boldsymbol{TP}$	jumlah True Positif
$\boldsymbol{TN}$	jumlah True Negatif
$\boldsymbol{FP}$	jumlah False Positif
$\boldsymbol{FN}$	jumlah False Negatif