# KLASIFIKASI STATUS KESEJAHTERAAN RUMAH TANGGA KELUARGA BINAAN SOSIAL MENGGUNAKAN ALGORITMA *NAIVE BAYES* BERBASIS SELEKSI ATRIBUT *CHI SQUARED*

**Erfan Karyadiputra[1], Edi Noersasongko[2], Aris Marjuni[3]**
[123]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

**ABSTRAK**

*Kemiskinan merupakan salah satu permasalahan yang sering dihadapi dalam upaya peningkatan kesejahteraan di hampir semua negara. Tersedianya data kemiskinan yang akurat dan berkesinambungan merupakan salah satu instrumen penting untuk mengevaluasi kebijakan pemerintah dalam mengentaskan kemiskinan dengan memfokuskan perhatian pada pendistribusian bantuan sesuai rumah tangga sasaran (RTS). Penelitian terkait klasifikasi kesejahteraan rumah tangga sering menggunakan variabel target/kelas berupa kategori miskin dan tidak miskin. Kategori tersebut jika dilihat dari aspek pendistribusian bantuan masih bersifat umum, hal tersebut karena kategori rumah tangga miskin tersebut dapat diklasifikasikan lagi kedalam status kesejahteraan rumah tangga sesuai rumah tangga sasaran (RTS) sehingga dalam pendistribusian bantuan dapat disesuaikan dengan status kesejahteraan rumah tangga sasaran (RTS). Oleh sebab itu diperlukan variabel target/kelas baru yang sesuai RTS Keluarga Binaan Sosial yaitu sangat miskin dan miskin.Dalam penelitian ini akan dilakukan pengujian menggunakan algoritma Naive Bayes berbasis seleksi atribut Chi Squared untuk mengklasifikasi status kesejahteraan rumah tangga miskin yaitu rumah tangga sangat miskin (RTSM) dan rumah tangga miskin (RTM). Hasil pengujian yang didapatkan adalah algoritma Naive Bayes menghasilkan akurasi sebesar 85.80% dan nilai AUC sebesar 0.930. kemudian Naive Bayes setelah menerapkan seleksi atribut menggunakan Chi Squared dengan nilai k sebanyak 13 atribut dapat meningkatkan akurasi menjadi 86.78% dan nilai AUC sebesar 0.944.*

*Kata kunci : Data Mining, Klasifikasi, Kemiskinan, Non-Monetary, Naive Bayes, Chi Squared, RTS, RTSM, RTM*

## 1. PENDAHULUAN

Kemiskinan adalah kondisi seseorang atau sekelompok orang, laki-laki dan perempuan, tidak mampu memenuhi hak dasarnya untuk mempertahankan dan mengembangkan kehidupan yang bermartabat sehingga dalam upaya meningkatkan kesejahteraan tersebut dapat dilakukan melalui program penanggulangan kemsikinan baik berupa bantuan sosial maupun pemberdayaan masyarakat. Banyak penelitian terkait klasifikasi kesejahteraan rumah tangga seringkali menggunakan variabel target/kelas berupa kategori miskin dan tidak miskin. Kategori tersebut jika dilihat dari aspek pendistribusian bantuan masih bersifat umum, hal tersebut karena kategori rumah tangga miskin tersebut dapat diklasifikasikan lagi ke dalam status kesejahteraan rumah tangga sesuai rumah tangga sasaran (RTS) sehingga dalam pendistribusian bantuan dapat disesuaikan dengan status kesejahteraan rumah tangga sasaran (RTS).

Oleh sebab itu diperlukan variabel target baru yang sesuai RTS Keluarga Binaan Sosial. Banyak faktor yang mempengaruhi status kesejahteraan rumah tangga tangga dari setiap data kemiskinan yang terus diperbaharui sehingga penggunaan atribut harus menyesuaikan data kemiskinan terbaru. Oleh karena itu perlu dilakukan identifikasi faktor-faktor yang paling berpengaruh dari data kemiskinan tersebut menggunakan atribut terbaru hasil PPLS 2011. Berdasarkan uraian permasalahan di atas, diperlukan

analisis proses klasifikasi status kesejahteraan rumah tangga menggunakan variabel target yang sesuai dengan kategori rumah tangga sasaran (RTS) yaitu rumah tangga miskin (RTM) dan kelompok rumah tangga sangat miskin/fakir miskin (RTSM) berdasarkan aspek *non-monetary* dengan mengunakan sejumlah atribut terbaru sesuai hasil PPLS 2011.

Penelitian ini difokuskan pada proses pengklasifikasian mengggunakan metode *Naive Bayes* yang memegang asumsi akan hubungan antarfitur atau atributnya yang independen sehingga menjadikannya lebih efektif untuk kategorisasi, sederhana, cepat, dan menghasilkan tingkat akurasi yang tinggi [1]. Kemudian untuk meningkatkan akurasi maka dilakukan seleksi atribut berbasis *Chi Squared*. Seleksi atribut dalam lingkup *data mining* bertujuan untuk mengidentifikasi variabel yang sama pentingnya dalam *dataset*, kemudian membuang variabel lain yang nilainya tidak relevan dan berlebihan [2]. Pemilihan variabel dapat mengurangi pemakaian data, hal ini memungkinkan lebih efektif dalam operasi yang lebih cepat sehingga memungkinkan dapat meningkatkan akurasi dalam pengklasifikasian data.

## 2. TINJAUAN PUSTAKA

Kesejahteraan Sosial merupakan kegiatan yang ditujukan untuk mengatasi masalah sosial melalui peningkatan pemberdayaan masyarakat dan meningkatkan akses terhadap sumber-sumber sosial yang ada di masyarakat [3]. Pembangunan bidang kesejahteraan sosial dikembangkan bersama dengan pembangunan ekonomi. Hal tersebut karena pembangunan ekonomi sangat mempengaruhi tingkat kemakmuran suatu negara sehingga penanganan masalah kesejahteraan sosial harus didekati dari berbagai sisi baik pembangunan ekonomi maupun kesejahteraan sosial.

Kemiskinan merupakan masalah yang kompleks yang dapat ditinjau dari beberapa segi, selain dari segi rendahnya pendapatan dan konsumsi pangan, juga dapat ditinjau dari segi pandangan perumahan, kesehatan, kebutuhan air bersih juga aspek non-material. Menurut Rohidi kemiskinan dipandang sebagai suatu kebudayaan atau lebih tegas lagi sebagai subkebudayaan dari kebudayaan yang lebih luas, mempunyai struktur dan sifat-sifatnya sendiri sebagai cara hidup yang diwariskan atau diwarisi antargenerasi melalui garis keluarga [4].

Badan Pusat Statistik (BPS) membuat kriteria kemiskinan, agar dapat menyusun secara lengkap pengertian kemiskinan sehingga dapat diketahui dengan pasti jumlahnya dan cara tepat menanggulanginya. Dalam penetapan keluarga miskin yang berhak menerima bantuan, pemerintah menggunakan acuan dari BPS tentang 14 (empat belas) Kriteria Kemiskinan meliputi luas lantai rumah, jenis luas lantai, jenis dinding rumah, fasilitas buang air besar, sumber air minum, penerangan yang digunakan, bahan bakar yang digunakan, frekuensi makan dalam sehari, kebiasaan membeli daging/ayam/susu, kemampuan membeli pakaian, kemampuan berobat ke puskesmas, lapangan pekerjaan kepala rumah tangga, pendidikan kepala rumah tangga dan kepemilikan aset [5].

*Data mining* merupakan teknologi terkini yang digunakan untuk membantu perusahaan menemukan sebuah informasi yang sangat penting dari data mereka. Analisis yang yang dilakukan oleh *data mining* melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan [6] Secara khusus, koleksi metode yang dikenal sebagai *data mining* menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi prediksi model [7].

Menurut Larose pendekatan *bayesian* digunakan untuk menentukan kemungkinan terhadap asumsi disekitarnya. Dalam statistik *bayesian*, parameter dipertimbangkan terhadap variabel acak dan data dipertimbangkan terhadap hasil kemungkinan [8]. *Naive Bayes* adalah salah satu metodeklasifikasi yang dapat memprediksiprobabilitas sebuah *class*, sehingga menghasilkan keputusan berdasarkan data pembelajaran dengan memberikan akurasi klasifikasi yang kompetitif dan efisiensi. Hal ini menyebabkan *Naive Bayes* banyak diterapkan dalam praktek [9].

Seleksi atribut merupakan proses yang digunakan dalam *machine learning*, dimana atribut dari subset yang tersedia dari data yang dipilih sebagai penerapan algoritma learning. *Subset* terbaik adalah berisi jumlah dimensi berkontribusi terhadap akurasi dengan menghilangkan atribut yang tidak sesuai. Seleksi

atribut terkait erat dengan pengurangan dimensi yang bertujuannya untuk mengidentifikasi tingkat kepentingan atribut dalam kumpulan data, dan membuang semua atribut lain seperti informasi yang tidak relevan dan berlebihan. maka hal ini akan memungkinkan operasi algoritma *data mining* dapat berjalan lebih efektif dan lebih cepat [2]. Metode *Chi Squared Statistic* adalah teknik statistik nonparameter yang digunakan untuk menentukan distribusi frekuensi yang diteliti dari yang diharapkan dengan menghitung bobot atribut yang berhubungan dengan kelas atribut. Semakin tinggi nilai bobot atribut maka semakin relevan.

## 3. METODE PENELITIAN

### 3.1. Pengumpulan Data

Penelitian yang dilaksanakan adalah jenis penelitian eksperimen, yaitu melakukan pengujian klasifikasi status kesejahteran RTS (Rumah Tangga Sasaran) Keluarga Binaan Sosial di Kecamatan Rantau Badauh Barito Kuala dengan metode *Naive Bayes* berbasis seleksi atribut *Chi Squared*.

Tabel 1. Penjelasan Variabel dan Kategori

| Variabel | Keterangan | Skala | Kategori |
|---|---|---|---|
| Y | Status Rumah Tangga Sasaran (RTS) | Nominal | 1 : Sangat Miskin RTSM |
| | | | 2 : Miskin RTM |
| X1 | Jenis Kelamin KRT | Nominal | 1 : Laki-Laki |
| | | | 2 : Perempuan |
| X2 | Umur KRT | Numerik | - |
| X3 | Pendidikan KRT | Nominal | 0 : Tidak Punya Ijazah |
| | | | 1 : SD/Sederajat |
| | | | 2 : SMP/Sederajat |
| | | | 3 : SMA/Sederajat |
| X4 | Lapangan Usaha KRT | Nominal | 1 : Pertanian (Padi & Palawija) |
| | | | 2 : Hortikultura |
| | | | 3 : Perkebunan |
| | | | 4 : Perikanan Tangkap |
| | | | 5 : Perikanan Budidaya |
| | | | 6 : Peternakan |
| | | | 7 : Kehutanan & Pertanian Lain |
| | | | 8 : Pertambangan / Penggalian |
| | | | 9 : Bangunan / Konstruksi |
| | | | 10 : Pedagang |

Tabel 2. Penjelasan Variabel dan Kategori (Lanjutan)

| Variabel | Keterangan | Skala | Kategori |
|---|---|---|---|
| X5 | Status Kependudukan dalam Pekerjaan Kepala Rumah Tangga | Nominal | 1 : Berusaha Sendiri |
| | | | 2 : Berusaha dibantu Buruh tidak tetap / tidak dibayar |
| | | | 3 : Berusaha dibantu Buruh tetap / dibayar |
| | | | 4 : Buruh/ Karyawan/ Pegawai Swasta |
| | | | 5 : Pekerja Bebas |
| | | | 6 : Pekerja Keluarga/Tidak dibayar |
| X6 | Status Penguasaan Bangunan Tempat Tinggal | Nominal | 1 : Milik Sendiri |
| | | | 2 : Kontrak/ Sewa |
| X7 | Jenis Atap Terluas | Nominal | 1 : Beton |
| | | | 2: Genteng |
| | | | 3 : Sirap |
| | | | 4 : Seng |
| | | | 5 : Asbes |
| | | | 6 : Ijuk/ Rumbia |
| X8 | Kualitas Atap | Nominal | 1 : Biasa/Kualitas Sedang |
| | | | 2 : Jelek/Kualitas Rendah |
| X9 | Jenis Dinding Terluas | Nominal | 1 : Tembok |
| | | | 2 : Kayu |
| | | | 3 : Bambu |
| X10 | Kualitas Dinding | Nominal | 1 : Biasa/Kualitas Sedang |
| | | | 2 : Jelek/Kualitas Rendah |
| X11 | Jenis Lantai | Nominal | 1 : Kayu / bambu |
| | | | 2 : Semen Tanpa Plester |
| X12 | Sumber Air Minum | Nominal | 1 : Air Kemasan |
| | | | 2 : Air Ledeng |
| | | | 3 : Air Terlindung |
| | | | 4 : Air Tidak Terlindung |
| X13 | Bahan Bakar Utama Memasak | Nominal | 1 : Listrik/ Gas/ Elpiji |
| | | | 2 : Minyak Tanah |
| | | | 3 : Kayu |
| X14 | Sumber Penerangan | Nominal | 1 : Listrik PLN |
| | | | 2 : Listrik Non-PLN |
| | | | 3 : Tidak Ada Listrik |
| X15 | Jumlah Keluarga | Numerik | - |
| X16 | Jumlah Individu | Numerik | - |

### 3.2. Pengujian Algoritma *Naive Bayes*

Pengujian dilakukan terhadap data *testing* dengan teknik *cross validation* dengan pengujian data mulai 2,3,4,5,6,7,8,9 dan 10 sehingga dapat di evaluasi hasilnya dengan mengukur seberapa keakuratan akurasi yang dihasilkan dari beberapa percobaan tersebut menggunakan metode *Naive Bayes.*

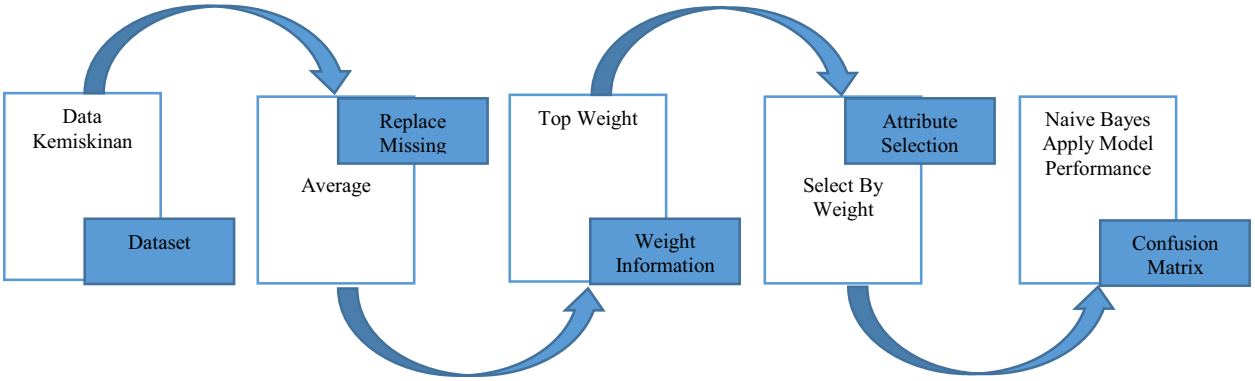
Tabel 3. Perbandingan Hasil Validasi Algoritma *Naive Bayes*

| ***Validation*** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|
| ***Accuracy*** | 85.70 | 85.80 | 86.28 | 85.50 | 85.70 | 85.80 | 86.19 | 85.60 | 85.80 |
| ***Precision*** | 82.96 | 82.44 | 83.31 | 82.12 | 82.47 | 82.96 | 83.21 | 82.63 | 82.75 |
| ***Recall*** | 87.89 | 89.33 | 88.92 | 89.12 | 88.91 | 88.31 | 89.13 | 88.71 | 88.90 |
| ***AUC*** | 0.937 | 0.935 | 0.940 | 0.927 | 0.938 | 0.935 | 0.935 | 0.938 | 0.930 |

Dari hasil percobaan seperti tabel 3.3 di atas, didapatkan pengulangan pengujian sebanyak 4 kali dengan hasil pengukuran berupa nilai rata-rata 4 kali pengujian memiliki tingkat akurasi tertinggi sebesar 86.28% dengan nilai AUC sebesar 0.940. Namun hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa penggunaan 10-*fold cross-validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat.

### 3.3. Seleksi Atribut Berbasis *Chi Squared*

Hasil akurasi yang dihasilkan sebesar 85.80 % sudah cukup baik namun untuk meningkatkan hasil akurasi yang telah dihasilkan dan untuk mendapatkan penggunaan atribut terbaik maka pengujian selanjutnya dengan melakukan seleksi atribut menggunakan *Naive Bayes* berbasis *Chi Squared.*




Gambar 5. Seleksi Atribut Berbasis *Chi Squared*

Pengujian dengan menerapkan model pengujian *Naive Bayes* dengan seleksi atribut berbasis *Chi Squared.* .Jumlah atribut keseluruhan yang digunakan sebelum dilakukan seleksi atribut adalah sebanyak 16 atribut. Jumlah tersebut kemudian diuji menggunakan seleksi atribut berbasis *Chi Squared* dengan mengurangi jumlah atribut menggunaan nilai *Top-k* yang kurang dari jumlah atribut keseluruhan. Ketika diimplementasi menghasilkan data seperti tabel 4 seperti berikut:

Tabel 4. Hasil Pengujian Seleksi Atribut *Naive Bayes* Berbasis *Chi Squared*

| K | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Accuracy*** | 84.83 | 85.12 | 86.28 | 85.70 | 85.90 | 86.19 | 86.19 | 85.90 | **86.78** | 86.68 | 85.90 |
| ***AUC*** | 0.930 | 0.931 | 0.937 | 0.935 | 0.939 | 0.941 | 0.944 | 0.944 | **0.944** | 0.944 | 0.931 |

Dari tabel 4 terlihat penggunaan seleksi atribut dengan nilai *k* sebanyak 13 atribut menghasilkan tingkat akurasi tertinggi sebesar 86.78 % dengan nilai AUC sebesar 0.944. Kemudian atribut hasil seleksi atribut yang didapatkan diantaranya seperti tabel 5 di bawah ini.

Tabel 5. Hasil Seleksi Atribut

| No | Atribut |
|---|---|
| 1 | Jenis kelamin KRT |
| 2 | Umur KRT |
| 3 | Pendidikan KRT |
| 4 | Lapangan Usaha KRT |
| 5 | SKP KRT |
| 6 | SPT KRT |
| 7 | Jenis Atap |
| 8 | Kualitas Atap |
| 9 | Kualitas Dinding |
| 10 | Jenis Lantai |
| 11 | Sumber Penerangan |
| 12 | Bahan Bakar Memasak |
| 13 | Jumlah Individu |

Dari hasil percobaan tersebut diatas maka penerapan seleksi atribut *Chi Squared* dapat meningkatkan akurasi *Naive Bayes* dengan penggunaan 13 atribut terpilih dari 16 atribut keseluruhan dan menyeleksi 3 atribut lainya yaitu jenis dinding, sumber air minum dan jumlah keluarga. Hasil tersebut kemudian digunakan untuk membandingkan seberapa besar kenaikan tingkat akurasi *Naive Bayes* yang dihasilkan dengan melakukan seleksi atribut berbasis *Chi Squared.*

### 3.4. Hasil Eksperimen

Hasil *Naive Bayes* dievaluasi dan dibandingkan dengan seleksi atribut *Naive Bayes* berbasis *Chi Squared* seperti tabel 6 berikut ini.

Tabel 6. Perbandingan Akurasi dan AUC

| Metode | *Naive Bayes* | *Naive Bayes* Berbasis *Chi Squared* |
|---|---|---|
| ***Accuracy*** | 85.80 | 86.78 |
| **AUC** | 0.930 | 0.944 |

Pada tabel 6 terlihat bahwa penggunaan seleksi atribut baik menggunakan *Chi Squared* dapat meningkatkan hasil akurasi *Naive Bayes* dari 85.80 % menjadi 86.78 %. Meskipun kenaikan yang dihasilkan tidak terlalu besar yaitu hanya sebesar 0,98% namun secara umum penerapan seleksi atribut menggunakan *Naive Bayes* berbasis *Chi Squared* lebih baik daripada *Naive Bayes* tanpa menggunakan seleksi atribut.

## 4. KESIMPULAN

Dari hasil penelitian yang dilakukan dari tahap awal hingga pengujian, dan hasil perbandingan dapat disimpulkan bahwa model yang terbentuk dengan algoritma *Naive Bayes* sendiri sudah memiliki akurasi yang cukup baik yaitu sebesar 85.80 % dan dengan proses seleksi atribut berbasis *Chi Squared*, model yang terbentuk dapat ditingkatkan lagi menjadi 86.78% dalam mengklasifikasikan status kesejahteraan rumah tangga miskin dari keluarga binaan sosial. Berdasarkan kehandalan dalam klasifikasi berupa nilai AUC yang didapat dari algoritma *Naive Bayes* sebelum seleksi atribut adalah 0.930 sehingga sudah tergolong sebagai *Excellent Classification* dan setelah dilakukan seleksi atribut nilai AUC meningkat menjadi 0.944. Dengan menerapkan seleksi atribut dapat memperbaiki *performanceNaive Bayes* dengan meningkatkan akurasi *Naive Bayes* menjadi lebih baik dari 85.80 % menjadi 86.78 %, dan menghasilkan 13 (tigabelas) atribut terpilih dari 16 (enambelas) atribut keseluruhan dengan menghilangkan 3 (tiga) atribut lainya yaitu jenis dinding, sumber air minum dan jumlah keluarga. Meskipun kenaikan yang dihasilkan tidak terlalu besar yaitu 0,98% namun secara umum hasil penerapan *Naive Bayes* dengan seleksi atribut menggunakan *Chi Squared* masih lebih baik daripada *Naive Bayes* tanpa menggunakan seleksi atribut berdasarkan kenaikan nilai akurasi tersebut dan keunggulan dalam hal kehandalan dalam klasifikasi karena memiliki nilai AUC yang sedikit lebih baik.

## DAFTAR PUSTAKA

[1] Huang, Y. & Li, L., 2011. *Naive bayes classification algorithm based on small sample set.* s.l.: IEEE Cloud Computing and Intelligence Systems.
[2] Maimon, O., 2010. *Data Mining And Knowledge Discovery Handbook.* London: Springer.
[3] Kemensos, 2014. *petunjuk pelaksanaan kelompok usaha bersama (KUBE).* jakarta: direktorat penanggulangan kemiskinan pedesaan.
[4] Rohidi, 2000. *Ekspresi Seni Orang Miskin.* Bandung: Nuansa.
[5] BPS, 2004. *Indikator Kemiskinan; Konsep dan Penghitungan.* Jakarta: Badan Pusat Statistik.
[6] Moertini, V., 2002. *Data Mining Sebagai Solusi Bisnis.* s.l.:Integral Vol. 7 No. 1.
[7] Bellazzi, R. & Zupanb, B., 2008. *Predictive Data Mining In Clinical Medicine: Current Issues And And Guidelines.* s.l.: International Journal Of Medical Informatics.
[8] Larose, 2006. *Data Mining Methods And Models.* Canada: John Wiley & Sons, Inc.
[9] Sammut, C. & Web, G., 2011. *Encyclopedia of machine learning.* New York: Springer.