UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 2 Tahun 2023 Halaman 97-104
ISSN(*Online*): 2985-475X

Application of Random Forest to Identify for Poor Households in West Sumatera Province

Febri Ramayanti, Dodi Vionanda*, Dony Permana, dan Zilrahmi
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 02 Januari 2023
Revised : 24 Januari 2023
Accepted : 13 Februari 2023

ABSTRACT

Poverty is a socioeconomic problem in Indonesia. The number of people who were living in poverty in West Sumatera increases for 26.44 thousand from 2020 to 2021. The government has created programs to cope with poverty by taking into account the criteria for the poor households. These criteria have been developed by using the data obtained through The National Socioeconomic Survey (Susenas). However, instead of showing the actual condition of poor household, the existing data only interprets the number of poor household. Thus make the program less effective. This could be overcome by classification analysis of random forest (RF). RF is collection of many decision trees. Before fitting RF, one has to determine the values if three tuning parameters, mtry, ntree and node size. The result are the smallest OOB's error rate (%) and Variable Importance Measure (VIM). The classification by RF in this research results in OOB's error rate was 5.65% or accuracy rate was 94.35% with tuning parameter using mtry=5 and ntree=500. Based on the VIM, the poor household's criteria include sources of drinking water such as protected or unprotected spring water and surface water, lighting tools such as non-PLN electricity or no usage of electricity, fuel for cooking such as charcoal and firewood, and the head of the household being self-employed, a family worker, or unpaid with at least a junior high degree.

Keywords: *OOB Error Rate, Poverty, Poverty Criteria, Random Forest, Variable Importance*

I. PENDAHULUAN

Kemiskinan merupakan permasalahan sosial ekonomi yang terjadi di beberapa Negara, termasuk Indonesia. Kemiskinan merupakan kondisi seseorang yang tidak mampu memenuhi kebutuhan pokoknya serta kondisi kebijakan yang menyebabkan ia lebih miskin dibandingkan sekitarnya (BAPPENAS, 2018). Pemerintah selalu mengupayakan kebijakan terkait program penanggulangan kemiskinan yang membutuhkan informasi berupa data nama dan alamat rumah tangga sasaran, dimana pengumpulan datanya harus dilakukan secara sensus (Kominfo, 2011). Kemiskinan yang terjadi di Sumatera Barat tahun 2021 meningkat sebesar 26,44 ribu penduduk dibandingkan tahun 2020. Pengukuran kemiskinan saat ini menggunakan konsep kemampuan memenuhi kebutuhan dasar, dimana penduduk dinyatakan miskin jika rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan (BPS, 2022) yang hanya mempresentasikan data kemiskinan makro berupa taksiran jumlah rumah tangga miskin (RTM) tetapi tidak dapat menentukan kondisi RTM tersebut (Sudirman, 2014). Hal ini menyebabkan penyaluran program penanggulangan kemiskinan menjadi belum sepenuhnya efektif di Sumatera Barat.

Kriteria kemiskinan membutuhkan data mikro yang memberikan informasi rumah tangga layak untuk menerima program. Salah satu informasinya yaitu menyediakan kriteria rumah tangga sasaran sehingga dapat menyusun perencanaan, implementasi dan pengendalian program untuk memperbaiki kinerja indikator kemiskinan makro (Sudirman, 2014). Identifikasi kriteria RTM perlu dilakukan agar target rumah tangga sebagai penerima program penanggulangan kemiskinan dapat berjalan lebih efektif dan tepat sasaran. Selain itu, identifikasi kriteria RTM juga dilakukan dengan tujuan penanggulangan agar kemiskinan yang ada di Sumatera Barat semakin menurun, serta tujuan lainnya yaitu sebagai pencegahan agar kriteria RTM yang terdapat pada saat ini, tidak terjadi kembali di masa depan. Hal ini perlu dilakukan agar program penanggulangan kemiskinan di Sumatera Barat dapat terprogram secara efektif. Identifikasi kriteria kemiskinan dapat diketahui dengan melakukan klasifikasi terhadap rumah tangga. Klasifikasi merupakan proses mengatur suatu entittas berdasarkan seperangkat sprinsip kedalam suatu kelas/grup (Jacob, 2004). Hasil klasifikasi terhadap rumah tangga berdasarkan indikator data Susenas memberikan informasi mengenai identifikasi kriteria RTM, sehingga dapat ditetapkan sebagai target dari program penanggulangan

DOI: 10.24036/ujsds/vol1-iss2/31

ISSN(*Online*): 2985-475X

kemiskinan oleh pemerintah serta sebagai pedoman dalam menetapkan kebijakan terkait kemiskinan. Salah satu metode klasifikasi yaitu metode *random forest* (RF) yang menggunakan suatu kontruksi *tree* (pohon) atau *decision tree* (pohon keputusan). D*ecision tree* terdiri dari tiga komponen utama yaitu *root node*, *internal node* serta *terminal node* (Zhang, 2010). Metode RF terbentuk berdasarkan kombinasi dari beberapa prediksi *tree* berbeda yang memberikan akurasi lebih tinggi dalam klasifikasi dibandingkan metode lainnya. Hal ini dibuktikan dengan salah satu penelitian yang dilakukan oleh Ansari S. dan Dhar M (2022) yang melakukan perbandingan metode regresi logistik, *decision tree*, *random forest*, *neural netword* dan *naïve bayes* untuk kalsifikasi kemiskinan rumah tangga di India. Hasil membuktikan bahwa diantara semua algoritma yang dilakukan menunjukkan *random forest* yang memiliki nilai akurasi tertinggi dalam mengklasifikasi kemiskinan rumah tangga dibandingkan metode lainnya.

RF membentuk hasil *forest* berdasarkan kombinasi tuning parameter *mtry* dan *ntree* yang menghasilkan nilai laju galat OOB dan *Variable Importance Measure* (VIM). Informasi yang diberikan oleh VIM digunakan sebagai pedoman untuk menentukan kriteria RTM di Sumatera Barat. Sehingga hasil yang diperoleh dapat berguna dan membantu pemerintah untuk menentukan sasaran serta kebijakan penanggulangan kemiskinan yang efektif dan terprogram. Rumusan masalah penelitian ini yaitu, bagaimana penerapan dan tingkat akurasi hasil metode *random forest* serta apa saja kriteria kemiskinan yang terdapat didaerah Sumatera Barat tahun 2021. Tujuan penelitian yaitu, mengetahui hasil dan tingkat akurasi metode *random forest* serta mengetahui kriteria rumah tangga miskin yang terdapat di daerah Sumatera Barat pada tahun 2021.

II. METODE PENELITIAN

A. *Classification and Regression Trees* (CART)

Metode CART merupakan suatu metode pohon keputusan (*decision tree*) yang bersifat *recursive partitioning*. Satu *tree* terdiri atas tiga komponen utama yaitu *root node*, *internal node* dan *terminal node*. Pada metode CART simpul akar (*root node*) dipartisi menjadi dua simpul anak (*internal node*), masing-masing simpul anak kemudian dipartisi menjadi dua simpul anak yang baru hingga menjadi *terminal node* yang bersifat homogen sebagai interpretasi dari *tree* (Zhang, 2010). CART membentuk *tree* dengan dua langkah yaitu, pembentukan maksimal dari *decision tree* berdasarkan proses *splitting* (pemilahan) dan pemangkasan (*pruning*) dengan mempertimbangkan *tree* dan cabang pohon yang terbentuk. Proses *splitting* variabel pada percabangan *node* pada *tree* dilihat dari variabel yang memiliki nilai *goodness of split* maksimal. Nilai ini dilihat berdasarkan perubahan *gini impurity/gini index* pada node *t* dan percabangan nodenya menurut Breiman (1984) dengan rumus sebagai berikut.

$$\text{Node kiri (L)}: \ imp(t_L) = \sum_{l=1}^{2} p_{t_L}^{(l)}(1 - p_{t_L}^{(l)}) \quad (1)$$

$$\text{Node kanan (R)}: imp(t_R) = \sum_{l=1}^{2} p_{t_R}^{(l)}(1 - p_{t_R}^{(l)}) \quad (2)$$

$$\text{Node } t: \ imp(t) = \sum_{k=1}^{2} p_t^{(k)}(1 - p_t^{(k)}) \quad (3)$$

Keterangan: $p_t^{(k)} = \frac{n_t^{(k)}}{n_t}$ $p_t^{(l)} = \frac{n_t^{(l)}}{n_t}$

$p_t^{(k)}, p_t^{(l)}$: Proporsi objek kelas klasifikasi ke-*k* atau ke-*l* pada node *t*

$n_t^{(k)}, n_t^{(l)}$: Jumlah observasi kelas klasifikasi ke-*k* atau ke-*l* pada node *t*

n_t : Jumlah seluruh observasi pada note *t*

Gini Impurity berfungsi untuk menentukan seberapa banyak pemisah yang akan dibentuk *decision tree*. Sementara dalam pemilihan variabel s yang digunakan untuk memilah ditentukan oleh nilai *Goodness of Split* sebagai berikut.

$$\Delta imp(s,t) = imp(t) - p_{t_L} imp(t_L) - p_{t_R} imp(t_R) \quad (4)$$

Keterangan : $p_{t_L} = \frac{n_{t_L}}{n_t}$ $p_{t_R} = \frac{n_{t_R}}{n_t}$

$p_{t_{(L\ atau\ R)}}$: Proporsi objek pada node *t* yang memilah pada node t_L atau node t_R

$n_{t_{(L\ atau\ R)}}$: Jumlah observasi pada node *t* yang memilah pada node t_L atau node t_R

n_t : Jumlah seluruh observasi pada note *t*

Variabel pemilah *s* yang memiliki *goodness of split* maksimal merupakan variabel yang lebih baik digunakan untuk melakukan proses *splitting*. Serta apabila *terminal node* yang terbentuk dari *internal node* memiliki nilai *gini index* lebih besar maka sebaiknya proses *splitting* dihentikan pada *internal node* sehingga menjadi *terminal node*.

DOI: 10.24036/ujsds/vol1-iss2/31

ISSN(*Online*): 2985-475X

B. *Random Forest* (RF)

Random Forest (RF) merupakan pengembangan metode CART. RF merupakan kumpulan banyak *decision tree* untuk membangun satu *forest* dan melihat *vote* klasifikasi dari *tree* yang menghasilkan prediktif lebih akurat (Zhang, 2010 dan Genuer, 2008). *Tree* di RF dibentuk tidak menggunakan seluruh sampel melainkan menggunakan sampel *bootstrap* dan tidak melakukan *pruning*. *Bootstrap* merupakan metode berbasis resampling data dengan syarat pengembalian dalam menyelesaikan suatu permasalahan (James, 2013). Pada RF sampel *bootstrap* yang digunakan ialah 2/3 data original dengan pengembalian sehingga membentuk sampel *bootstrap* yang memiliki jumlah sama dengan data original sedangkan 1/3 data original lainnya disebut sampel *out of bag* (OOB) yang digunakan untuk pengujian prediksi *tree* yang sudah terbentuk dari sampel *bootstrap* (Breiman, 2001).

Terdapat tiga tuning parameter yang digunakan metode RF yaitu *mtry* (banyak input variabel secara acak terpilih dalam satu *node* pemilahan) yang secara default $mtry = \sqrt{p}$ untuk kasus klasifikasi, *ntree* (jumlah banyaknya *tree* dalam *forest*) yang secara default $ntree = 500$, penelitian ini menggunakan ntree berjumlah 100, 250, 500 dan 1000, serta *node size* (minimum nomor observasi dalam sebuah *node*) yang secara default 1 untuk klasifikasi (Probst, 2019). Pembentukan *tree* pada RF dilakukan dengan cara membentuk sampel *bootstrap*, lalu melakukan teknik *recursive partitioning* pada sampel *bootstrap* sehingga menghasilkan sebuah *tree*, dimana dalam proses *splitting tree* atribut diambil berdasarkan banyaknya variabel yang terpilih melalui *mtry*. Selanjutnya melakukan kembali pembentukan sampel *bootstrap* dan metode *recursive partitioning* untuk membentuk *tree* lainnya sehingga terbentuk beberapa *tree* berdasarkan *ntree* dalam membangun satu *forest* untuk melihat vote klasifikasi dari seluruh *tree* yang terbentuk.

C. Laju Galat *Out Of Bag* (OOB)

OOB sampel berfungsi sebagai percobaan prediksi *tree* yang terbentuk dikarenakan setiap tree memiliki sampel *bootstrap* yang berbeda, sehingga setiap amatan dapat menjadi sampel OOB dan perlu diprediksi menggunakan beberapa *tree* yang dibangun tidak menggunakan sampel tersebut. Estimasi *error* pada hasil prediksi RF dapat diduga dengan menggunakan laju galat OOB *(OOB error rate)* yang dihitung dari hasil proporsi kesalahan prediksi klasifikasi setiap amatan dari hasil RF (Janitza, 2018). Penggunaan *mtry* untuk melihat hasil dari OOB *error* diharapkan tidak terlalu rendah, dikarenakan apabila terlalu rendah, maka hasil OOB *error* akan semakin tinggi yang menghasilkan RF memiliki kinerja yang buruk. OOB *error rate* diharapkan memiliki nilai terkecil (minimum). Berikut perhitungan laju galat OOB dalam klasifikasi (Genuer R, 2020).

$$Laju\ Galat\ OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_{i \neq \hat{Y}_i}} \tag{5}$$

Keterangan, $\sum_{i=1}^{n} 1_{Y_{i \neq \hat{Y}_i}}$: jumlah data hasil prediksi yang salah (misklasifikasi)

Y_i : hasil amatan sebenarnya ke-I

$\hat{Y}_i$: hasil amatan yang diprediksi ke-i

n : jumlah OOB ke-i menjadi sampel OOB.

OOB *error rate* digunakan untuk memprediksi observasi ke-*i* dari X_i, dimana prediksi hanya berlaku untuk suatu *tree* yang sampel *bootstrap*nya tidak mengandung (X_i, Y_i) dan memberikan prediksi $\hat{Y}_i$ untuk output dari *i* observasi. Selanjutnya dihitung rataan tingkat misklasifikasi OOB dengan perhitungan berikut.

$$Laju\ Galat\ OOB = \frac{\sum OOB\ error\ rate}{k} \times 100\% \tag{6}$$

dimana *k* merupakan jumlah banyak pohon terbentuk berdasarkan ketentuan parameter *ntree*. Semakin kecil *OOB error rate* yang dihasilkan maka prediksi pada *forest* akan semakin akurat dan dapat dipercaya.

D. *Variable Importance Measure* (VIM)

Penggunaan analisis dalam RF secara umum sangat sulit untuk melakukan interpretasi dalam memperoleh informasi. Salah satu solusi untuk mempermudah memperoleh informasi dalam RF ialah dengan mengidentifikasi *Variable ImportanceMeasure* (VIM) untuk variabel prediktor. Apabila variabel *importance* dapat diidentifikasi, maka hasil RF akan diperoleh metode penyeleksian variabel yang berpengaruh penting terhadap pembentukan *tree* dalam RF. Estimasi pemilihan variabel *importance* dalam *random forest* dapat dilakukan dengan melihat berapa banyak kenaikan prediksi *error* (OOB) data untuk variabel terpilih sementara yang lainnya tidak berubah (Liaw, 2002).

Metode *representatif* dari perhitungan pengukuran variabel *importance* adalah *Mean Decrease Impurity* (MDI) atau disebut juga dengan *Mean Decrease Gini* (MDG) yang diusulkan oleh Breiman pada tahun 2001. Suatu *p* peubah penjelas dengan *h* = (1,2,…,*p*) maka rumus mengukur tingkat kepentingan peubah penjelas X_h dengan cara berikut (Xiao Li. dkk, 2019).

$$MDG(x_h) = \frac{1}{k}\sum_{t=1}^{k} MDG\,(X_{h,} x^t) \tag{7}$$

DOI: 10.24036/ujsds/vol1-iss2/31

Keterangan :

$MDG(X_{h,}x^t) = \sum_{t\in(T),v(t)=h} \frac{N_n(t)}{n} \Delta x^{(t)}$

$\Delta x^{(t)}$: indeks gini untuk peubah penjelas X_h pada pohon ke-k

$N_n(t)$: jumlah sampel keseluruhan pada *leaf*

k : banyaknya pohon dalam random forest

Selain itu, perhitungan VIM dapat juga dengan menggunakan perhitungan *Mean Decrease Accuracy* (MDA) atau *Permutation Importance* yang menggunakan OOB untuk membagi data sampelnya, dimana OOB memperkirakan nilai prediksi dengan menghitung nilai akurasi OOB sebelum dan sesudah permutasi X_h dan menghitung perbedaannya, dengan rumus sebagai berikut (Strobl. C. dkk, 2008).

$$MDA(x_h) = \frac{1}{k}\sum_{t=1}^{k} \frac{\sum_{i\in OOB^{(t)}} I(y_i=\hat{y}_i^{(t)}) - \sum_{i\in OOB^{(t)}} I(y_i=\hat{y}_{i,h}^{(t)})}{|OOB^{(t)}|} \quad (8)$$

dimana $OOB^{(t)}$ adalah sampel OOB untuk satu *tree* ke-*t*, dengan t ∈{1, 2, 3,…, k}, tingkat kepentingan variabel X_h dalam *tree* ke-*t* adalah nilai rata-rata dari perbedaan antara kelas prediksi sebelum permutasi X_h, yaitu $\hat{y}_i^{(t)} = f^{(t)}(x_i)$ dan kelas prediksi setelah permutasi X_h, yaitu $\hat{y}_{i,h}^{(t)} = f^{(t)}(x_{i,h})$ dalam *i* observasi tertentu.

E. Jenis Penelitian dan Sumber Data

Penelitian ini merupakan penelitian kuantitatif yang mengembangkan dan menggunakan perhitungan matematis serta menggunakan data berupa angka untuk menganalisis permasalahan sehingga dapat menarik kesimpulan. Penelitian ini menggunakan data sekunder berupa variabel rumah tangga yang diperoleh dari survei sosial ekonomi nasional (Susenas) yang dikumpulkan oleh Badan Pusat Statistik Sumatera Barat pada tahun 2021.

Variabel yang digunakan terdiri dari satu variabel respon (Y) yaitu kelompok rumah tangga berdasarkan pengeluaran per kapita yang sudah dikategorikan berdasarkan garis kemiskinan dan 26 variabel prediktor (Xi) yaitu jumlah anggota rumah tangga (ART), ijazah tertinggti KRT (kepala rumah tangga), pekerjaan/usaha KRT, kedudukan pekerjaan KRT, umur KRT, jenis kelamin KRT, status perkawinan KRT, status penguasaan tempat tinggal, jenis dinding tempat tinggal, luas lantai tempat tingga, jenis lantai tempat tinggal, kepemilikan fasilitas sanitasi, sumber penerangan, sumber air minum, bahan bakar untuk memasak, kepemilikan rekening tabungan ART (anggota rumah tangga), kepemilikan telepon seluler (HP) ART, kepemilikan komputer/laptop ART, kepemilikan rumah lain selain rumah yang ditempati, kepmilikan tanah/lahan ART, kepemilikan mobil ART, sumber pendapatan rumah tangga, pengalaman menerima program kartu keluarga sejahtera (KKS), pengalaman menerima program keluarga harapan (PKH), pengalaman menerima bantuan pangan non tunai (BPNT)/program sembako), serta pengalaman kehabisan makanan karena kurangnya uang.

F. Langkah Analisis Data

Data dianalisis dengan menggunakan paket *randomforests* pada *software RStudio*. Selanjutnya untuk mengetahui variabel yang memberikan pengaruh terhadap kemiskinan serta identifikasi kriteria rumah tangga miskin dilakukan dengan langkah analisis berikut.

1. Melakukan pengambilan data survei sosial ekonomi nasional (SUSENAS) dari BPS Sumatera Barat.
2. Memilah dan menyiapkan data yang ada pada data susenas sesuai variabel yang diperlukan.
3. Melakukan eksplorasi data terhadap variabel penelitian untuk menggambarkan keadaan data dalam penelitian.
4. Untuk i=1 sampai $ntree_i$=100, $ntree_i$=250, $ntree_i$=500 dan $ntree_i$=1000, melakukan langkah berikut.
   a. Pilah n amatan secara acak kedalam 2 kelompok, kelompok 1 terdiri dari 2/3 bagian data dan kelompok 2 terdiri dari 1/3 bagian data lainnya.
   b. Lakukan pengambilan sampel *bootstrap* pada kelompok 1 dengan sistem pengembalian (*replacement*) sebanyak data original untuk membangun pohon ke-i.
   c. Amatan yang termasuk kelompok 2 dijadikan sebagai sampel OOB ke-i yang akan digunakan untuk prediksi.
   d. Berdasarkan sampel *bootstrap* ke-i dibangun pohon ke-i dengan langkah berikut.
      i. Hitung nilai *mtry* yang digunakan untuk menentukan jumlah variabel prediktor dalam memilah tiap simpul. Dimana, $mtry_1=\sqrt{p} = 5$, $mtry_2=2\sqrt{p} = 10$, dan $mtry_3= \sqrt{p}/2 = 3$ dengan $mtry<p$ (jumlah variabel).
      ii. Hitung nilai *impurity* simpul anak kiri dan kanan menggunakan Rumus 1 dan 2, dan hitung nilai *impurity* simpul menggunakan Rumus 3.
      iii. Mencari pemilah variabel terbaik menggunakan *goodness of split* menggunakan Rumus 4.
   e. Memprediksi sampel OOB ke-i dengan langkah berikut.
      i. Menghitung laju galat OOB ke-i menggunakan Rumus 5 pada pohon ke-i yang terbentuk.
      ii. Menghitung laju galat OOB menggunakan Rumus 6.

DOI: 10.24036/ujsds/vol1-iss2/31

f. Selesai.

5. Menentukan penerapan yang paling optimal berdasarkan tuning parameter RF *mtry* dan *ntree* yang dikombinasikan.
6. Mengidentifikasi variabel penting yang memberikan pengaruh terhadap variabel Y berdasarkan nilai *Variable Importance Measure* (VIM) MDG dan MDA menggunakan Rumus 7 dan Rumus 8, serta mengurutkan variabel importance.
7. Menarik Kesimpulan.

III. HASIL DAN PEMBAHASAN

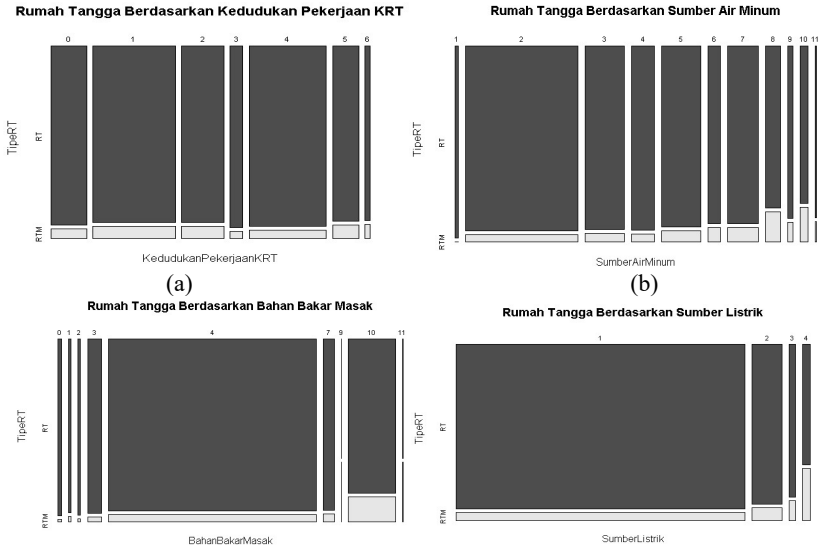
A. Ekplorasi Data

Data yang digunakan dalam penelitian ini berjumlah sebanyak 11548 amatan dengan 26 variabel prediktor dan 1 variabel respon. Berikut statistika deskriptif data untuk variabel prediktor berskala numerik.

Tabel 1 Statistik Deskriptif Variabel Numerik

Statistik Deskriptif	Rumah Tangga Miskin (RTM)			Rumah Tangga Biasa (RT)		
	Umur	Luas lantai	Jumlah ART	Umur	Luas lantai	Jumlah ART
Mean	48.55	58.71	4	49.19	76.14	4
Standar deviasi	13.62	32.43	2	13.69	50.10	2
Minimum	23	10	1	13	3	1
Maximum	89	240	13	97	640	14

Pada Tabel 1 terlihat bahwa standar deviasi pada variabel umur, luas lantai dan jumlah ART terhadap tipe rumah tangga relatif rendah yang menunjukkan bahwa sebaran data pada variabel tersebut menyebar disekitar rata-rata dan keragaman sebaran data cukup kecil. Sehingga data yang ada kurang bervariasi atau nilai pada data hampir serupa atau mendekati dengan nilai rata-rata. Serta nilai maximum dan minimum variabel umur, luas lantai dan jumlah ART yang memiliki perbedaan terhadap tipe rumah tangga yang menunjukkan terdapat hubungan antara variabel umur, luas lantai dan jumlah ART terhadap tipe rumah tangga. Selanjutnya gambaran beberapa variabel prediktor berskala kategorik yang ditampilkan dalam visualisasi pada Gambar 1.



Gambar 1. Tipe rumah tangga berdasarkan (a) kedudukan pekerjaan, (b) sumber air minum, (c) bahan bakar masak, dan (d) sumber listrik.

Keterangan kriteria :

(a) 0:Tidak bekerja, 1:Berusaha sendiri, 2:Berusaha dibantu buruh tidak tetap/buruh tidak dibayar, 3:Berusaha dibantu buruh tetap/buruh dibayar, 4:Buruh/karyawan/pegawai, 5:Pekerja bebas, 6:Pekerja keluarga/tidak dibayar.

(b) 1:Air kemasan bermrek, 2:Air isi ulang, 3:Leding meteran, 4:Leding eceran, 5:Sumur bor/pompa, 6:Sumur terlindung, 7:Sumur tak terlindung, 8:Mata air terlindung, 9:Mata air tak terlindung, 10:Air permukaan (sungai/ danau/waduk/kolam/irigasi), 11:Air hujan.

(c) 0:Tidak memasak di rumah, 1:Listrik, 2:Elpiji 12 kg, 3:Elpiji 5.5 kg, 4:Elpiji 3 kg, 5:Gas kota, 6:Biogas, 7:Minyak tanah, 8:Briket, 9:Arang, 10:Kayu bakar, 11:Lainnya.

(d) 1:PLN dengan meteran, 2:PLN tanpa meteran, 3:Non PLN, 4:Bukan Listrik.

Sumber : Susenas BPS Sumatera Barat 2021

DOI: 10.24036/ujsds/vol1-iss2/31

ISSN(*Online*): 2985-475X

Berdasarkan Gambar 1 terlihat bahwa tipe rumah tangga miskin berpeluang lebih kecil terjadi pada responden dengan pekerjaan KRT kriteria 3 (berusaha dibantu buruh tetap) dan kriteria 6 (pekerja keluarga), sumber air minum bermrek (kriteria 1), bahan bakar masak elpiji > 3 kg (kriteria 2 dan 3) dan sumber listrik PLN dengan meteran (kriteria 1). Rumah tangga miskin berpeluang lebih besar terjadi pada responden dengan kedudukan pekerjaan KRT kriteria lainnya, sumber air minum kriteria lainnya, bahan bakar masak berupa kayu bakar (kriteria 10) serta sumber penerangan ialah bukan listrik (kriteria 4) dibandingkan dengan rumah tangga biasa. Perbedaan kedudukan pekerjaan KRT, sumber air minum, bahan bakar masak dan sumber listrik yang digunakan yang terjadi pada rumah tangga menunjukkan terdapat hubungan antara variabel kedudukan pekerjaan KRT, sumber air minum, bahan bakar masak serta sumber listrik terhadap tipe rumah tangga.

B. Penerapan Metode *Random Forest*

Penerapan metode *rf* untuk identifikasi kriteria RTM di Provinsi Sumatera Barat diperoleh dari kombinasi parameter *mtry* dan *ntree*. Penerapan metode *rf* menggunakan parameter *mtry* yaitu 3, 5 dan 10 dikarenakan perhitungan parameter *mtry* $\sqrt{p}$, $2\sqrt{p}$, dan $\sqrt{p}/2$. Serta, menggunakan parameter *ntree* yaitu 100, 250, 500 dan 1000. Langkah selanjutnya yaitu menentukan hasil optimal berdasarkan nilai OOB *error rate* berdasarkan tuning paramater *mtry* dan *ntree* yang terdapat pada Tabel 2 berikut.

Tabel 2 Laju Galat OOB

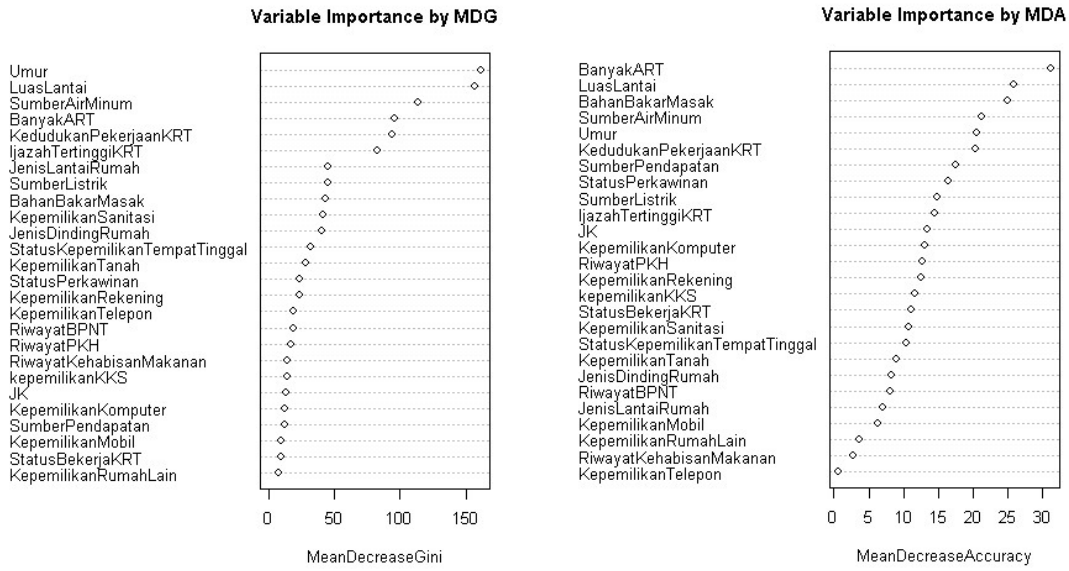
Parameter *Mtry*	Parameter *Ntree*			
	100	250	500	1000
3	5.72%	5.73%	5.72%	5.71%
5	5.77%	5.70%	**5.65%	5.66%
10	5.83%	5.70%	5.72%	5.72%

Keterangan : **) adalah hasil *forest* yang optimal

Pembentukan *forest* bernilai optimal adalah ketika laju galat OOB memiliki nilai terkecil. Pada Tabel 2 terlihat bahwa hasil *forest* optimal yang terbentuk ialah hasil *forest* dengan kombinasi *Mtry* = 5 dan *Ntree* = 500, serta memiliki nilai laju galat OOB terkecil sebesar 5.65%. Sehingga tingkat keakuratan *forest* dalam memprediksi klasifikasi tipe rumah tangga ialah sebesar 94.35%.

C. Hasil *Variable Importance Measure* (VIM)

Berdasarkan penerapan *random forest* optimal, dilanjutkan dengan melihat kepentingan variabel pada data atau *variable importance measure* (VIM) pada Gambar 2 yang memperlihatkan kepentingan masing-masing variabel prediktor berdasarkan MDA dan MDG dan variabel penting untuk klasifikasi masing-masing tipe rumah tangga didaerah Sumatera Barat berdasarkan nilai VIM pada Tabel 3 berikut.



Gambar 2. VIM data susenas Sumbar 2021

DOI: 10.24036/ujsds/vol1-iss2/31

Tabel 3. Variabel Penting Klasifikasi Rumah Tangga

No	Y = Rumah Tangga Miskin		No	Y = Rumah Tangga	
1	Sumber listrik	Jumlah ART	1	Jumlah ART	Kedudukan dalam pekerjaan
2	Bahan bakar masak	Kedudukan dalam pekerjaan	2	Luas lantai	Sumber air minum
3	Sumber air minum	Jenis lantai rumah	3	Umur KRT	Status perkawinan
4	Luas lantai	Jenis dinding rumah	4	Bahan bakar masak	Ijazah tertinggi KRT
5	Kepemilikan telepon (HP)	Kepemilikan tanah	5	Sumber pendapatan	Jenis kelamin KRT

Pada RF variabel yang memiliki kepentingan paling tinggi ialah variabel yang memiliki nilai VIM tertinggi dan diurutkan berdasarkan nilai VIM tertinggi ke terendah. Berdasarkan Gambar 2 dapat ditarik kesimpulan bahwa variabel penting pada hasil *forest* optimal ialah variabel luas lantai, sumber air minum, jumlah ART, kedudukan pekerjaan, umur KRT dan bahan bakar masak dan seterusnya Selanjutnya, Tabel 3 memperlihatkan variabel yang mencirikan kedua tipe rumah tangga, dimana untuk klasifikasi rumah tangga miskin dipengaruhi oleh variabel sumber listrik, bahan bakar masak, sumber air minum, luas lantai, kepemilikan telepon (HP), jumlah ART, jenis lantai dan lainnya. Sedangkan klasifikasi rumah tangga tidak miskin dipengaruhi oleh variabel jumlah ART, luas lantai, bahan bakar masak, sumber pendapatan, kedudukan dalam pekerjaan, sumber air minum dan lainnya.

Penerapan *random forest* yang diperoleh, didapatkan bahwa untuk identifikasi kriteria RTM di Provinsi Sumatera Barat pada tahun 2021 diperoleh dari *forest* hasil kombinasi *mtry* = 5 dan *ntree* = 500 berdasarkan nilai OOB *error rate* minimum sebesar 5.65%. Sehingga sebesar 94.35% adalah tingkat keakuratan *forest* dalam mengklasifikasi tipe rumah tangga dan memprediksi kriteria RTM yang terdapat di daerah Sumatera Barat. Informasi hasil *forest* diperoleh berdasarkan nilai *Variable Importance Measure* (VIM) yang mencirikan kedua tipe rumah tangga. Berdasarkan nilai VIM diperoleh kriteria rumah tangga miskin (RTM) yaitu sumber penerangan berupa listrik non-PLN atau tidak menggunakan listrik, bahan bakar untuk memasak berupa arang, kayu bakar, sumber air minum berupa air mata terlindung/tak terlindung atau air permukaan (sungai/danau/waduk/kolam), luas lantai rumah berkisar kurang dari 30 m^2 dan berjenis kayu/papan atau tanah serta dinding rumah berupa batang kayu, anyaman bambu dan kayu/papan, memiliki anggota keluarga berjumlah lebih dari lima orang dan tidak memiliki telepon/Hp, tanah, laptop, mobil, rekening, rumah lain selain yang ditempati, kedudukan pekerjaan KRT ialah pekerja bebas atau pekerja keluarga/tidak dibayar, umur dan ijazah yang dimiliki KRT yaitu berumur < 45 tahun dan berijazah SLTP/sederajat, SD/sederajat atau tidak memiliki ijazah, sumber pendapatan terbesar yaitu anggota keluarga yang bekerja atau kiriman uang/barang, keluarga pernah mengalami kehabisan makanan dan memiliki atau mendapatkan program BPNT, KKS dan PKH, rumah tangga tidak memiliki fasilitas sanitasi atau fasilitas sanitasi yang dimiliki berada di MCK umum, serta status penguasaan tempat tinggal ialah bebas sewa atau milik sendiri.

Selain kriteria rumah tangga miskin tersebut, berdasarkan nilai VIM juga diperoleh kriteria rumah tangga tidak miskin yaitu memiliki anggota keluarga berjumlah kurang dari lima orang, luas lantai berkisar lebih dari 30 m^2 dan berjenis marmer/granit dan keramik serta dinding rumah berupa tembok, sumber air minum berupa air kemasan bermrek dan bahan bakar untuk memasak berupa gas elpiji > 3 kg, sumber penerangan berupa listrik PLN dengan meteran, kedudukan dalam pekerjaan dan ijazah yang dimiliki KRT yaitu berusaha dibantu buruh tetap/buruh dibayar dan berijazah minimal SLTA/sederajat, sumber pendapatan terbesar yaitu investasi dan pensiunan, status penguasaan tempat tinggal ialah dinas, kontrak/sewa, terdapat anggota keluarga yang memiliki tanah, rumah lain, mobil, rekening, laptop dan telepon/Hp.

IV. KESIMPULAN

Metode RF menghasilkan *forest* yang optimal untuk identifikasi kriteria RTM di Sumatera Barat dengan kombinasi tuning parameter *mtry* = 5, *ntree* = 500 dan nilai OOB *error rate* minimum sebesar 5.65%. Tingkat keakuratan *forest* yang diperoleh ialah sebesar 94.35% dalam mengklasifikasi tipe rumah tangga dan memprediksi kriteria RTM di Sumatera Barat. Berdasarkan nilai VIM diperoleh bahwa variabel sumber air minum, bahan bakar masak, kedudukan KRT dalam pekerjaan, luas lantai rumah dan lainnya merupakan variabel penting dalam mengklasifikasi rumah tangga. Kriteria RTM yang diperoleh berdasarkan nilai VIM yaitu sumber penerangan berupa listrik non-PLN atau tidak menggunakan listrik, bahan bakar untuk memasak berupa arang dan kayu bakar, sumber air minum berupa air mata terlindung/tak terlindung dan air permukaan (sungai/ danau/ waduk/ kolam), luas lantai rumah berkisar kurang dari 30 m^2 dan berjenis kayu/papan atau tanah, kedudukan pekerjaan KRT ialah pekerja bebas atau pekerja keluarga/tidak dibayar dan kriteria lainnya. Penulis mengharapkan untuk penelitian selanjutnya dapat memperhatikan proporsi data terkait klasifikasi rumah tangga miskin dan tidak miskin agar proporsi yang diteliti dapat seimbang, selain itu diharapkan dapat menambah variabel prediktor baru terkait kemiskinan yang belum

DOI: 10.24036/ujsds/vol1-iss2/31

terfasilitasi pengukurannya agar memberikan solusi terkait penanggulangan kemiskinan yang lebih efektif bagi pemerintah dan masyarakat.

DAFTAR PUSTAKA

Ansari, S., Dhar, M. (2022). Poverty Classification Based On Unsatisfied Basic Needs Index: A Comparison Of Supervised Learning Algoritms. *SN Soc Sci 2, 69*. https://doi.org/10.1007/s43545-022-00375-y

Badan Pusat Statistik (BPS). (2022). Kemiskinan dan Ketimpangan. https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html. Diakses pada tanggal 22 Juni 2022 pukul 10.43.

Badan Pusat Statistik (BPS). (2022). Survei Sosial Ekonomi Nasional. https://www.bps.go.id/index.php/subjek/81. Diakses pada tanggal 29 September 2022, pukul 11.43.

Badan Pusat Statistik (BPS). (2022). Jumlah Penduduk Miskin Menurut Kabupaten/Kota di Sumatera Barat. https://sumbar.bps.go.id/indicator/23/125/1/jumlah-penduduk-miskin-menurut-kabupaten-kota-di-sumatera-barat-.html. Diakses pada tanggal 06 Oktober 2022, pukul 20.35.

BAPPENAS. (2018). *Analisis Wilayah dengan Kemiskinan Tinggi*. Jakarta: Kementrian PPN/Bappenas.

Breiman, L.dkk. (2001). Random Forest. UC Berkeley, Departement of Statistics. *Machine Learning*, 45, 5-32.

Denil, M. David M., dan Nando de F. (2014). Narrowing the Gap: Random Forest In Theory and In Practice. *International Conference on Machine Learning* Vol. 32

Dinas Komunikasi dan Informatika. (2011). Program Penanggulangan Kemiskinan Kabinet Indonesia Bersatu II. Jakarta : Kementerian komunikasi dan informatika RI.

Genuer, R. Jean M. P., dan Christine T. (2008). *Random Forest: Some Methodolical Insight*. France: INRIA.

Genuer, R. dan Jean M. P. (2020). *Random Forest With R*. Switzerland: Springer.

Jacob, E. 2004. Classification and Categorization: A difference that Makes a Difference.*Library Trends*, Vol.52, No.3.

James, G. dkk. (2013). *An Introduction to Statistical Learning: with Application in R*. New York: Springer.

Janitza, S. dan Roman H. (2018). On the Overestimation of Random Forest's Out of Bag Error. PloS ONE 13(8):e0201904.DOI:10.1371/journal.pone.0201904.

Liaw, A. dan Matthew W. (2002). Classification and Regression by Random Forest. ISSN 1609-3631

Louppe, G. (2014). *Understanding Random Forest*. Belgia : Universitas of Liege.

Probst, P. Marvin W., dan Anne L.B. (2019). Hyperparameters and Tuning Strategies for Random Forest. arXiv:1804.03515v2[stat.ML].

Scornet, E. (2018). Tuning Parameter In Random Forests. *Esaim Proceedings and Surveys*. Vol. 60, p.144-162.

Strobl, C. dkk. (2008). Conditional Variable Importance for Random Forests. Ludwig Maximilians Universitat Munchen.

Sudirman. (2014). Analisis kemiskinan makro dan mikro kabupaten kutai kartanegara. *Jurnal ilmu sosial MAHAKAM*. Volume 3 No 1.

Sumodiningrat, S. dan Maiwan. (1999). *Kemiskinan : Teori, Fakta dan Kebijakan*. IMPAC, Jakarta. Dalam Miftahuddin. (2011). Analisa Karakteristik Rumah Tangga Miskin dengan Metode Regresi Logistik Terbaik. *Jurnal Matematika, Statistika, Komputasi*. Vol.7.No.2.

Xiao Li, dkk. (2019). A Debiased MDI Feature Importance Measure for Random Forest. Canada : University of California Berkeley.

Zhang, H. dan Burton H.S. (2010). *Recursive Partitioning and Applications*.2nd ed. Springer. New York.

DOI: 10.24036/ujsds/vol1-iss2/31