

## ANALISIS KLASIFIKASI STATUS BEKERJA PENDUDUK DAERAH ISTIMEWA YOGYAKARTA MENGGUNAKAN METODE *RANDOM FOREST*

Eka Christy<sup>1</sup>, Kris Suryowati<sup>2</sup>

<sup>1,2)</sup> Jurusan Statistika, Fakultas Sains Terapan, IST AKPRIND Yogyakarta  
e-mail: ekachristy137.ec@gmail.com

**Abstract.** Classification is a job related to making a model that can do a mapping from each set of variables to each target, then use the model to provide target values for the new set obtained. In this study the classification analysis is used to see the prediction of the working status of residents in rural areas of Daerah Istimewa Yogyakarta. The method used to classify is the Random Forest method. Random Forest is a classification consisting of several decision trees that are built using random vectors. This study aims to get the best model/tree using Random Forest. Models/trees obtained from Random Forest will be used to classify the working status of residents in the rural areas of Daerah Istimewa Yogyakarta. Result from the obtained model with the smallest classification error is  $mtry$  4 and  $nree$  500 with classification accuracy of 80.13%. Sequentially the variable with the biggest role in classifying the working status of Daerah Istimewa Yogyakarta residents in rural areas is the Disease/Interference variable, followed by the Education Level, Gender and Certificate variables.

**Keywords:** Classification, Random Forest, Working Status

**Abstrak.** Klasifikasi merupakan pekerjaan yang berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Dalam penelitian ini analisis klasifikasi digunakan untuk melihat prediksi status bekerja penduduk di wilayah perdesaan Daerah Istimewa Yogyakarta. Metode yang digunakan untuk mengklasifikasi adalah metode *Random Forest*. *Random Forest* adalah klasifikasi yang terdiri dari beberapa pohon keputusan yang dibangun dengan menggunakan vektor acak. Penelitian ini bertujuan untuk mendapat model/pohon terbaik menggunakan *Random Forest*. Model/pohon yang diperoleh dari *Random Forest* akan digunakan untuk mengklasifikasi status bekerja penduduk di wilayah perdesaan Daerah Istimewa Yogyakarta. Hasil yang diperoleh model dengan *error* klasifikasi terkecil adalah  $mtry$  4 dan  $nree$  500 dengan tingkat akurasi ketepatan klasifikasinya sebesar 80,13%. Secara berurutan variabel yang paling besar peranannya dalam mengklasifikasi status bekerja penduduk Daerah Istimewa Yogyakarta di wilayah perdesaan adalah variabel Penyakit/Gangguan selanjutnya diikuti oleh variabel Tingkat Pendidikan, Jenis Kelamin dan Sertifikat.

**Kata kunci:** Klasifikasi, *Random Forest*, Status Bekerja

### 1. PENDAHULUAN

Pemerintah Daerah Istimewa Yogyakarta dalam melakukan pembangunan berupaya mewujudkan pertumbuhan ekonomi yang berkualitas. Hal ini ditandai dengan keinginan untuk tercapainya beberapa indikator makro seperti Indeks Pembangunan Manusia yang tinggi, rendahnya tingkat pengangguran, juga menurunnya kesenjangan pendapatan antar penduduk [1]. Tingkat Pengangguran Terbuka (TPT) di Daerah Istimewa Yogyakarta pada Februari 2018 sebesar 3,06 persen atau meningkat dibanding Februari 2017 sebesar 2,84 persen dan Agustus 2017 sebesar 3,02 persen berdasarkan data yang dimiliki oleh Badan Pusat Statistik. Pada Februari 2018 TPT perkotaan dan perdesaan tercatat masing-masing sebesar 3,93 dan 1,06 persen [2]. TPT wilayah perkotaan tercatat lebih besar daripada wilayah perdesaan. Sehingga klasifikasi akan digunakan untuk melakukan penilaian terhadap status bekerja penduduk di wilayah perdesaan untuk mengetahui klasifikasi penduduk yang bekerja. Analisis ketepatan status bekerja memiliki peran penting bagi pencari kerja karena dapat diketahui karakteristik penduduk yang bekerja.

Berdasarkan uraian diatas diperlukan analisis untuk melihat karakteristik penduduk yang berpotensi mendapatkan pekerjaan. Pada permasalahan ini dapat diterapkan teknik *data mining* untuk melakukan klasifikasi menggunakan metode *Random Forest*. *Random Forest* didasarkan pada teknik pohon keputusan sehingga mampu mengatasi masalah non-linier. Metode ini merupakan metode pohon gabungan. Untuk mengidentifikasi peubah penjelas yang relevan dengan peubah respons, *Random Forest* menghasilkan ukuran tingkat kepentingan (*variable importance*) peubah penjelas [dewi].

Penelitian ini dilakukan untuk mengetahui (1) Bagaimana gambaran umum dan karakteristik Penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan tahun 2018. (2) Bagaimana hasil analisis klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta di wilayah perdesaan tahun 2018 menggunakan metode *Random Forest*. (3) Dengan melihat tingkat akurasi berapa besar ketepatan klasifikasi dengan metode *Random Forest* untuk memprediksi Status Bekerja Penduduk Daerah Istimewa Yogyakarta di wilayah perdesaan tahun 2018. (4) Menggunakan *Variable Importance*, variabel manakah yang paling besar pengaruhnya dalam mengklasifikasi Status Bekerja Penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan tahun 2018. Dengan tujuan mengetahui gambaran umum dan karakteristik Penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan tahun 2018, mengetahui hasil analisis klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta di wilayah perdesaan tahun 2018 dengan metode *Random Forest*, mengetahui berapa besar ketepatan klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta di wilayah perdesaan tahun 2018 menggunakan metode *Random Forest* dengan melihat tingkat akurasi dan mengetahui variabel yang memiliki pengaruh paling besar dalam mengklasifikasi Status Bekerja Penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan tahun 2018.

## 2. METODOLOGI PENELITIAN

### 2.1 Data

Data yang digunakan dalam penelitian ini adalah data sekunder dari Survei Angkatan Kerja Nasional tahun 2018 yang dilaksanakan oleh Badan Pusat Statistik (BPS) Daerah Istimewa Yogyakarta dengan studi kasus wilayah perdesaan. Data yang digunakan terdiri atas sebuah peubah respons (Status Bekerja) dan empat peubah penjelas (Jenis Kelamin, Tingkat Pendidikan, Sertifikat dan Penyakit/Gangguan). *Software* yang digunakan untuk membantu analisis adalah R paket Rattle.

### 2.2 *Random Forest*

*Random Forest* adalah klasifikasi yang terdiri dari beberapa pohon keputusan yang dibangun dengan menggunakan vektor acak. Secara sederhana, algoritma pembentukan *Random Forest* dapat disebutkan sebagai berikut, andaikan gugus data *training* yang dimiliki berukuran  $n$  amatan dan terdiri atas  $p$  peubah penjelas (*predictor*). Tahapan penyusunan dan pendugaan menggunakan *Random Forest* adalah (Breiman, 2001; Breiman & Cutler, 2003):

- a) (tahapan *bootstrap*) melakukan penarikan contoh acak dengan pemulihan berukuran  $n$  dari gugus data *training*.
- b) (tahapan *random sub-setting*) dengan menggunakan gugus data *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih  $m$  peubah penjelas secara acak, dimana  $m < p$ , lalu pemilah terbaik dipilih berdasarkan  $m$  peubah penjelas tersebut.
- c) Ulangi langkah a-b sebanyak  $k$  kali sehingga diperoleh sebuah hutan yang terdiri atas  $k$  buah pohon acak.

Pemilihan  $m$  memegang peranan dalam menentukan kebaikan *Random Forest* yang dihasilkan. Terdapat tiga cara untuk mendapatkan nilai  $m$  yaitu (Breiman & Cutler, 2003):

$$m = \begin{cases} m_1 = \frac{1}{2} \sqrt{p} \\ m_2 = \sqrt{p} \\ m_3 = 2 \times \sqrt{p} \end{cases} \quad (1)$$

Dimana  $p$  adalah total variabel.

Respons suatu amatan diprediksi dengan menggabungkan (*aggregating*) hasil prediksi  $k$  pohon. Misklasifikasi *Random Forest* diduga melalui *error* OOB yang diperoleh dengan cara (Breiman, 2001; Breiman & Cutler, 2003):

- a) Lakukan prediksi terhadap setiap data OOB pada pohon yang bersesuaian. Data OOB (*Out-Of-Bag*) adalah data yang tidak termuat dalam gugus data *bootstrap*.
- b) Secara rata-rata, setiap amatan gugus data asli akan menjadi data OOB sebanyak sekitar 36% dari banyak pohon. Oleh karena itu, pada langkah 1, masing-masing amatan gugus data asli mengalami prediksi sebanyak sekitar sepertiga kali dari banyaknya pohon. Jika  $a$  adalah sebuah amatan dari gugus data asli, maka hasil prediksi *Random Forest* terhadap  $a$  adalah gabungan dari hasil prediksi setiap kali  $a$  menjadi data OOB.
- c) Error OOB dihitung dari proporsi misklasifikasi hasil prediksi *Random Forest* dari seluruh amatan gugus data asli.

Breiman dan Cutler (2005) menyarankan untuk mengamati *error* OOB saat  $k$  kecil, lalu memilih  $m$  yang menghasilkan *error* OOB terkecil. Jika *Random Forest* dilakukan dengan menghasilkan *variable importance*, disarankan untuk menggunakan banyak pohon, misalnya 1000 pohon atau lebih. Jika peubah penjelas yang dianalisis sangat banyak, nilai tersebut dapat lebih besar agar *variable importance* yang dihasilkan semakin stabil.

### 2.2.1 Variable Importance

Pendekatan naif untuk mengukur tingkat *variable importance* (tingkat kepentingan) adalah menghitung berapa kali variabel tersebut muncul dalam kelompok pohon keputusan. Semakin besar dampaknya, semakin penting variabelnya. Metode representatif dari indeks pengukuran *variable importance* adalah *Mean Decrease Impurity* (MDI) atau biasa juga disebut *Mean Decrease Gini* (MDG) dan *Mean Decrease Accuracy* (MDA) yang diusulkan oleh Breiman pada tahun 2001. Misalkan terdapat  $p$  peubah penjelas dengan  $h = (1, 2, \dots, p)$  maka MDG mengukur tingkat kepentingan peubah penjelas  $x_h$  dengan cara:

$$MDG(x_h) = \frac{1}{k} [1 - \sum_k Gini(h)^k] \quad (2)$$

Dimana:  $Gini(h)^k$  : indeks Gini untuk peubah penjelas  $X_h$  pada pohon ke- $k$   
 $k$  : banyaknya pohon dalam *Random Forest* (ukuran *random forest*)

MDA adalah metode yang menghitung tingkat kepentingan dengan permutasi dan metode ini menggunakan OOB untuk membagi data sampelnya. Dengan kata lain, OOB memperkirakan nilai prediksi yang lebih akurat dengan menghitung nilai akurasi OOB sebelum dan sesudah permutasi variabel  $x_h$  dan menghitung perbedaannya. Mengukur nilai MDA dilakukan dengan cara:

$$MDA(x_h) = \frac{1}{k} \sum_{t=1}^k \frac{\sum_{i \in OOB} I(y_i = f(x_i)) - \sum_{i \in OOB} I(y_i = f(x_i^j))}{|OOB|} \quad (3)$$

Dengan  $t \in \{1, 2, 3, \dots, k\}$ , tingkat kepentingan variabel  $x_h$  dalam pohon  $t$  adalah nilai rata-rata dari perbedaan antara kelas prediksi sebelum permutasi  $x_h$ , yaitu  $y_i = f(x_i)$ , dan setelah permutasi variabel  $x_h$ , yaitu  $y_i = f(x_i^j)$ , dalam  $i$  pengamatan tertentu.

### 2.2.2 Ketepatan Klasifikasi

Alat ukur yang akan digunakan untuk mengevaluasi hasil akhir dari algoritma klasifikasi ini adalah matriks yang biasa disebut “*Confusion or Contingency Matrix*” (Matriks Konfusi atau Matriks Kontingensi). Matriks Konfusi ini mengilustrasikan klasifikasi “*confusion*” atau dengan kata lain *classification error*, dimana setiap baris menampilkan kelas aktual dan setiap kolomnya menampilkan kelas prediksi.

**Tabel 1.** Tabel *Confusion Matrix*

| Aktual             | Prediksi           |                   |
|--------------------|--------------------|-------------------|
|                    | <i>Good Credit</i> | <i>Bad Credit</i> |
| <i>Good Credit</i> | TP                 | FN                |
| <i>Bad Credit</i>  | FP                 | TN                |

Keterangan:

TP = Jumlah prediksi yang tepat bersifat positif (*True Positive*).

TN = Jumlah prediksi yang tepat bersifat negatif (*True Negative*).

FP = Jumlah prediksi yang salah bersifat positif (*False Positive*).

FN = Jumlah prediksi yang salah bersifat negatif (*False Negative*).

*Accuracy* merupakan proporsi jumlah prediksi yang tepat/benar. Rumus untuk mendapatkan nilai akurasi adalah:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+F} \quad (4)$$

## 3. HASIL DAN PEMBAHASAN

### 3.1 Gambaran Umum

Jumlah total penduduk yang diteliti adalah 2517 penduduk. Dari 2517 penduduk yang diteliti, sebanyak 76% dari total penduduk atau 1902 penduduk yang Berkerja, sedangkan 24% penduduk lainnya atau sejumlah 615 penduduk yang Tidak Bekerja. Pada variabel Jenis Kelamin penduduk yang diteliti, 52% atau sejumlah 1305 penduduk berjenis kelamin Perempuan, ini lebih banyak dibandingkan penduduk berjenis kelamin Laki-laki yang berjumlah 1212 penduduk atau 48%. Pada variabel Tingkat Pendidikan terakhir yang ditamatkan, 589 penduduk tidak tamat sekolah itu adalah 23,40% dari total penduduk yang diteliti. 22,33% atau total 562 penduduk merupakan tamatan SD atau sederajat. 571 penduduk atau 22,69% merupakan tamatan SMP atau sederajat. 647 penduduk atau 25,70% merupakan tamatan SMA atau sederajat. 51 penduduk atau 2,03% merupakan tamatan Diploma I/II/III. 94 penduduk atau 3,73% merupakan tamatan Diploma IV/S1. Tiga penduduk atau 0,12% merupakan tamatan S1. Pada variabel kepemilikan Sertifikat pelatihan/kursus/*training* dari penduduk yang diteliti, sebanyak 2302 penduduk atau 91% dari total penduduk tidak memiliki sertifikat, hanya 9% dari total penduduk atau 215 penduduk yang memiliki sertifikat. Pada variabel Penyakit/Gangguan yang dimiliki oleh penduduk yang diteliti, 2277 penduduk atau 90,47% penduduk tidak memiliki penyakit/gangguan. 39 penduduk atau 1,55% penduduk memiliki penyakit/gangguan penglihatan. 26 penduduk atau 1,03% penduduk memiliki penyakit/gangguan pendengaran. 25 penduduk atau 0,99% penduduk memiliki penyakit/gangguan berjalan/naik tangga (mobilitas). Dua penduduk atau 0,08% penduduk memiliki penyakit/gangguan menggunakan/menggerakkan jari/tangan. Satu penduduk atau 0,04% penduduk memiliki penyakit/gangguan berbicara dan atau memahami/berkomunikasi dengan orang lain. 12 penduduk atau 0,48% penduduk memiliki penyakit/gangguan lainnya (misal seperti mengingat/konsentrasi, perilaku/emosional, mengurus diri, dan lain-lain). 135 penduduk atau 5,36% penduduk memiliki lebih dari satu penyakit/gangguan yang telah ditulis.

### 3.2 Analisis Klasifikasi *Random Forest*

Peneliti menggunakan 75% data untuk data *training* yaitu sebesar 1888, dan sisanya 25% atau sebesar 629 untuk data *testing*/uji coba. Selanjutnya adalah menentukan menentukan berapa banyak pohon yang akan terbentuk (*ntree*) dan menentukan berapa banyak *random sample* yang diambil untuk setiap percobaan (*mtry*). Penentuan nilai *mtry* dilakukan menggunakan “Persamaan (1)”, yaitu:

- 1)  $Mtry_1 = \frac{\sqrt{\text{banyak variabel}}}{2} = \frac{\sqrt{5}}{2} = 1,118 \approx 1$
- 2)  $Mtry_2 = \sqrt{\text{banyak variabel}} = \sqrt{5} = 2,236 \approx 2$
- 3)  $Mtry_3 = \sqrt{\text{banyak variabel}} \times 2 = \sqrt{5} \times 2 = 4,472 \approx 4$

Nilai *ntree* yang dicobakan adalah:

- |                 |                  |
|-----------------|------------------|
| $Ntree_1 = 25$  | $Ntree_4 = 250$  |
| $Ntree_2 = 50$  | $Ntree_5 = 500$  |
| $Ntree_3 = 100$ | $Ntree_6 = 1000$ |

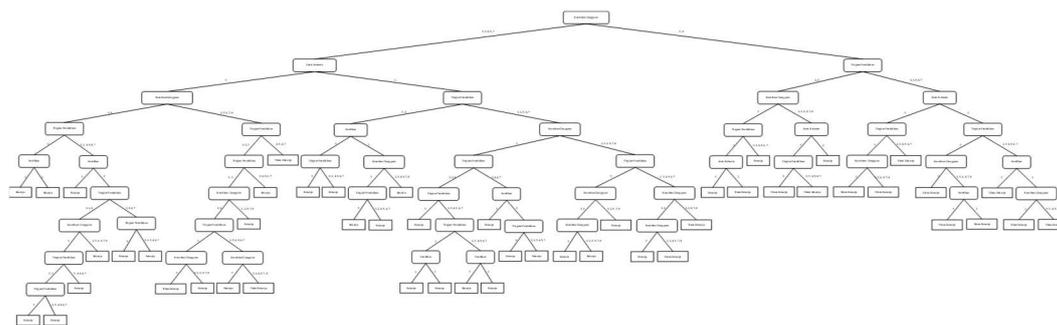
Untuk memperoleh model yang optimal dan nilai *error* OOB yang bernilai kecil, pembentukan *Random Forest* dilakukan berulang-ulang dengan kombinasi nilai *mtry* dan *ntree* yang berbeda. Hasil dari nilai *error* OOB untuk masing-masing nilai *mtry* dan *ntree* adalah sebagai berikut:

**Tabel 2.** Nilai *Error* OOB *Mtry* dan *Ntree* pada Data *Training*

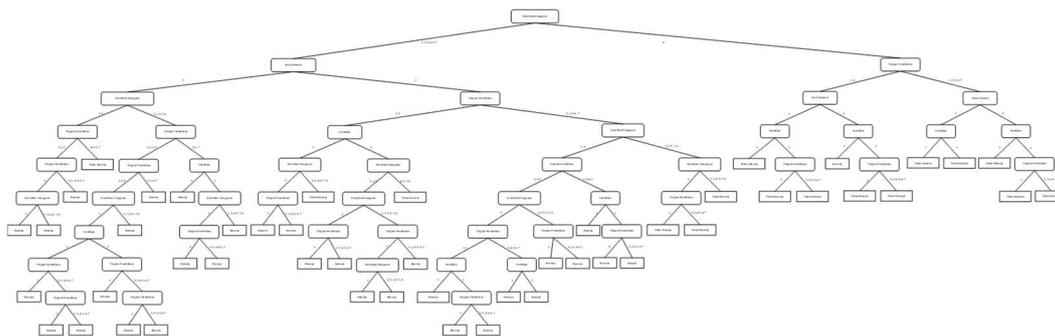
| <i>Mtry</i> | <i>Ntree</i> |        |        |        |        |        |
|-------------|--------------|--------|--------|--------|--------|--------|
|             | 25           | 50     | 100    | 250    | 500    | 1000   |
| 1           | 22.25%       | 22.62% | 22.51% | 21.82% | 21.66% | 21.61% |
| 2           | 21.98%       | 21.61% | 21.93% | 21.61% | 21.50% | 21.56% |
| 4           | 21.72%       | 21.82% | 21.56% | 21.50% | 21.45% | 21.45% |

Nilai *error* terendah adalah 21.45% dengan nilai *mtry* sebesar 4 dan nilai *ntree* sebesar: 500 dan 1000. Karena peningkatan *ntree* dari 500 pohon menjadi 1000 pohon terlihat tidak memberikan penurunan rata-rata tingkat *error* yang berarti. Dengan demikian, dapat dikatakan bahwa tingkat *error* dalam memprediksi mulai konvergen saat menggunakan 500 pohon, maka selanjutnya akan digunakan nilai *mtry* sebesar 4 dan *ntree* sebesar 500.

Diperoleh *Random Forest* untuk *mtry* 4 dan *ntree* 500 seperti pada “Gambar 1”.



Pohon



Pohon 500

**Gambar 1.** *Random Forest* yang Terbentuk dengan *Mtry* 4 dan *Ntree* 500

Nilai estimasi *error* OOB yang didapatkan menggunakan data latih (*training*) sebesar 21.45%. Perhitungan adalah dari Tabel 4.7, yaitu  $\frac{405}{1888} \times 100\% = 21.45\%$ .

**Tabel 3.** Hasil Prediksi Data Latih (*Training*) untuk *mtry* 4 dan *nree* 500

| Data <i>Training</i> | Prediksi |               | Total |
|----------------------|----------|---------------|-------|
|                      | Bekerja  | Tidak Bekerja |       |
| Bekerja              | 1374     | 36            | 1410  |
| Tidak Bekerja        | 369      | 109           | 478   |
| Total                | 1743     | 145           | 1888  |

Dapat dilihat bahwa penduduk yang diprediksi dengan tepat yaitu mereka yang Bekerja sebanyak 1374 penduduk dan mereka yang Tidak Bekerja adalah 109 penduduk. Namun, ada 369 penduduk yang sebenarnya Tidak Bekerja telah diprediksi Bekerja dan ada 36 penduduk yang sebenarnya Bekerja tetapi diprediksi Tidak Bekerja. Model *Random Forest* dengan *mtry* 4 dan *nree* 500 dapat diketahui sangat bagus dalam memprediksi, dimana hanya terdapat 405 penduduk atau 21.45% yang diprediksi tidak tepat.

### 3.3 Tingkat Akurasi Ketepatan Klasifikasi

Setelah membuat model pada metode *random forest*, langkah selanjutnya yaitu melakukan pengujian akurasi menggunakan data uji coba/data baru (*testing*) dengan menggunakan *mtry* = 4 dan *nree* = 500.

**Tabel 4.** Hasil Prediksi Data Uji Coba (*Testing*)

| Data <i>Testing</i> | Prediksi |               |
|---------------------|----------|---------------|
|                     | Bekerja  | Tidak Bekerja |
| Bekerja             | 483      | 9             |
| Tidak Bekerja       | 116      | 21            |

“Tabel 4” merupakan matriks konfusi sehingga dapat dihitung akurasinya menggunakan rumus pada “Persamaan (4)”, hasilnya sebagai berikut:

$$\text{Akurasi} = \frac{TP+T}{TP+TN+FP+FN} = \frac{483+9}{483+21+116+9} = 0.8013 = 80.13\%$$

Hasil akurasi data *testing* menunjukkan bahwa metode *Random Forest* untuk melakukan prediksi dari data *testing* atau untuk data baru dengan hasil prediksi tepat sebesar 80.13% atau bisa dikatakan dapat memprediksi dengan baik.

### 3.4 Penentuan Tingkat Kepentingan Variabel

**Tabel 5.** *Variable Importance* untuk Status Bekerja

| Variabel           | <i>Mean Decrease Accuracy</i><br>(MDA) | <i>Mean Decrease Gini</i><br>(MDG) |
|--------------------|--|------------------------------------|
| Penyakit/Gangguan  | 119.33                                 | 62.79                              |
| Tingkat Pendidikan | 10.17                                  | 19.84                              |
| Jenis Kelamin      | 4.85                                   | 18.59                              |
| Sertifikat         | -0.59                                  | 3.85                               |

“Tabel 5” menunjukkan hasil ukuran tingkat kepentingan dari *Random Forest* dengan *mtry* 4 dan *n tree* 500. Pada ukuran tingkat kepentingan MDA dan ukuran tingkat kepentingan MDG memiliki hasil yang sama, dimana variabel bebas yang memiliki tingkat kepentingan tertinggi secara berurutan adalah Penyakit/Gangguan, Tingkat Pendidikan, Jenis Kelamin dan Sertifikat.

## 4. KESIMPULAN

Berdasarkan hasil analisis yang dilakukan maka dapat disimpulkan beberapa hal sebagai berikut:

- 1) Dari data sampel penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan tahun 2018 yang berjumlah 2517 penduduk, menunjukkan bahwa sebanyak 76% dari total penduduk atau 1902 penduduk yang berkerja, sedangkan 24% penduduk lainnya atau sejumlah 615 penduduk yang tidak bekerja.
- 2) Prediksi klasifikasi melalui metode *Random Forest* dilakukan dengan berbagai kombinasi *mtry* dan *n tree*. Berdasarkan perbandingan nilai *error* OOB, kombinasi yang mampu menghasilkan *error* klasifikasi terkecil adalah dengan *mtry* 4 dan *n tree* 500. Hasil prediksi klasifikasi berdasarkan data latih (data *training*) adalah penduduk yang Bekerja sangat bagus dimana dari total 1743 penduduk, 1374 orang tepat prediksi sedangkan 369 orang lainnya tidak tepat prediksi. Untuk prediksi kelas Tidak Bekerja juga sangat bagus dimana dari total 145 penduduk, 109 orang tepat prediksi sedangkan 36 orang lainnya tidak tepat prediksi.
- 3) Dengan melihat tingkat akurasinya pada data uji (data *testing*), diperoleh besar ketepatan klasifikasi status bekerja penduduk di Daerah Istimewa Yogyakarta wilayah perdesaan pada tahun 2018 menggunakan metode *Random Forest* sebesar 80,13%. Hal ini menunjukkan bahwa ketepatan metode klasifikasi *Random Forest* untuk data ini sudah sangat bagus, dengan banyak pohon yang terbentuk (*n tree*) adalah 500 dan banyak random sample yang diambil untuk setiap percobaan (*mtry*) adalah 4.
- 4) Variabel yang paling mempengaruhi hasil prediksi menurut urutan besar pengaruhnya adalah; Penyakit/Gangguan memiliki peran yang paling besar, selanjutnya Tingkat Pendidikan, Jenis Kelamin dan Sertifikat.

## UCAPAN TERIMA KASIH

Dalam penyusunan tulisan ini, banyak pihak yang telah memberikan dukungan kepada penelitian ini. Peneliti menyampaikan terima kasih kepada Institut Sains & Teknologi AKPRIND Yogyakarta yang telah memberikan fasilitas sarana dan prasarana dalam pelaksanaan penelitian, serta kepada Bapak/Ibu Dosen Jurusan Statistika IST AKPRIND Yogyakarta atas arahan dan bimbingannya.

**DAFTAR PUSTAKA**

- [1] Anonim, 2017, *Analisis Ketimpangan Pengeluaran Penduduk Indeks Gini Daerah Istimewa Yogyakarta 2014-2016*, Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, Yogyakarta.
- [2] Anonim, 2018, *Statistik Ketenagakerjaan Provinsi Daerah Istimewa Yogyakarta 2017-2018*, Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, Yogyakarta.
- [3] Brieman L, 2001, *Random Forest Machine Learning*, Kluwer Academic Publisher, Belanda.
- [4] Dewi N K, Syafitri U D, dan Mulyadi S Y, 2011, Penerapan Metode *Random Forest* dalam *Driver Analysis*, *Forum Statistika dan Komputasi*, Vol. 16, No. 1, Hal. 35-43, Bogor, <https://journal.ipb.ac.id/index.php/statistika/article/view/5443>.
- [5] Jae-Hee H, Sun-Young I, Young-Ho P, 2017, *A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing*, *Hindawi*, Vol. 2017, Seoul, <https://doi.org/10.1155/2017/681727>.
- [6] Nidhomuddin, dan Otok B W, 2015, *Random Forest dan Multivariate Adaptive Regression Spline (MARS) Binary Response* untuk Klasifikasi Penderita HIV/AIDS Di Surabaya, *Statistika*, Vol. 1, No. 3, Hal. 50-57, Surabaya, <https://jurnal.unimus.ac.id/index.php/statistik/article/view/1439>.
- [7] Widiastuti J, 2018, Klasifikasi Pembiayaan Warung Mikro Menggunakan Metode *Random Forest* dengan Teknik *Sampling* Kelas *Imbalanced* Studi Kasus: Data Nasabah Pembiayaan Warung Mikro Bank Syariah Mandiri KC Jambi, *Skripsi*, Fakultas Matematika Dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Yogyakarta.
- [8] Williams G, 2011, *Data Mining with Rattle and R The Art of Excavating Data for Knowledge Discovery*, Springer, New York.