UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume I Nomor 2 Tahun 2023 Halaman 45-52
ISSN(*Online*): 2985-475X

Application of Random Forest for The Classification Diabetes Mellitus Disease in RSUP Dr. M. Jamil Padang

Fazhira Anisha, Dodi Vionanda*, Nonong Amalita, 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

Diabetes Mellitus is a disease in which blood sugar levels exceed normal limit (GDS>200 mg/dl). Diabetes Mellitus is defined as an insulin function disorder in the pancreatic organ. In Indonesia, Diabetes Mellitus continues to increas every year. Prevention and control of the disease are necessary to avoid complications with other organs, even causing death. Because of this, one needs to study a method to predict the occurance of this disease and to knows the variable that most affect a person suffered from it. This could be accomplished by using a classification methods. The purpose of the classification is to specify a class or category of new data based on preexisting data characteristics. One of classification methods is Random Forest. The concept of random forests was to combine multiple decision trees with the binary type, built with a sample of bootstraps. The result of this study are the smallest OOB's error rates (%) and Variable Importance Measure (VIM).The classification by a Random Forest methods on the incidence of Diabetes Mellitus in RSUP Dr. M. Jamil Padang results in OOB's error rate was 1,2% or accuracy rates was 98,8%. The most optimal model produced is the one using mtry=4 and ntree=1000. From this study, it can be concluded that Age, Polyphagia, Polyuria, HB, BMI, and Delayed Healing are identified as important variables. Hence, one might utilize these variables to classify whether a patient will be labeled as type I DM or type II DM.

Keywords: *OOB's Error Rate, Random Forest, VIM*

I. PENDAHULUAN

Diabetes Melitus (DM) didefinisikan terdapat suatu gangguan pada fungsi insulin. DM diklasifikasikan atas DM tipe 1, DM tipe 2, dan DM pada kehamilan. DM menjadi masalah kesehatan dunia karena insiden penyakit ini terus meningkat di belahan dunia manapun, termasuk Indonesia. Menurut hasil Riset Kesehatan Dasar (RisKesDas) 2018, jumlah penderita DM terus meningkat tiap tahunnya. Pada Tahun 2019, prevalensi DM di Indonesia mencapai 6,2% atau sekitar 10,7 juta penderita. Selanjutnya pada Tahun 2021 mencapai 10,8% atau sekiatr 19,46 juta penderita.

Menurut Bawono, *et al* (2021) pada Tahun 2018, Sumatera Barat memiliki prevalensi total DM sebanyak 1,6%. Ibukota Provinsi yakni Kota Padang, menjadi kota yang memiliki jumlah penderita DM terbanyak di Provinsi Sumatera Barat yaitu sebanyak 44.280 kasus. Upaya pencegahan dan pengendalian penyakit ini perlu dilakukan agar tidak menyebabkan komplikasi pada organ tubuh lainnya bahkan hingga kematian. Salah satu upaya yaitu dengan mendeteksi dari awal seseorang yang memiliki ciri-ciri atau kategori seperti pengidap DM. Tujuannya adalah sebagai pencegahan, ketika telah diketahui faktor-faktor yang kemungkinan mempengaruhi terjadi penyakit ini, seseorang dapat menghindar dan mencegah agar penyakit ini tidak masuk ke dalam tubuhnya. Tujuan lainnya adalah sebagai penyembuhan, ketika seseorang terdeteksi memiliki penyakit ini maka akan ada upaya untuk pengobatannya jauh lebih cepat dan efisien berdasarkan tipe DM yang diidap oleh penderita.

Berdasarkan hal tersebut perlu adanya suatu metode yang dapat memprediksi serta mengetahui variabel yang paling mempengaruhi seseorang terkena penyakit DM dengan tingkat analisis yang akurat. Salah satunya adalah metode *Random Forest* (RF). Metode ini mampu mengklasifikasikan data yang memiliki amatan yang tidak lengkap (*missing value*). Menurut Scornet (2015: 1716), metode ini merupakan metode *ensemble* yang mampu meningkatkan kestabilan pohon klasifikasi yang terbentuk dan keakuratan prediksi yang dihasilkan. RF merupakan pengembangan dari metode CART. RF sangat disarankan untuk digunakan dibandingkan metode CART, hal ini dikarenakan pada metode RF, tiap-tiap variabel diberikan kesempatan yang sama untuk membangun pohon dengan menggunakan sampel *bootstrap*. Hal

DOI: 10.24036/ujsds/vol1-iss2/30

ISSN(*Online*): 2985-475X

ini juga dibuktikan dengan salah satu penelitian yang dilakukan oleh Sabariah, *et al* (2014), dengan melakukan perbandingan metode RF dan CART. Hasil menunjukkan model RF mencapaikan keakurasi yang paling tinggi.

Pada penelitian ini bertujuan untuk mengetahui variabel penting dalam data dan tingkat keakurasi metode RF untuk klasifikasi penyakit DM di RSUP Dr. M. Jamil Padang. Secara umum, hasil penelitian ini berupa laju galat OOB (%) eror terkecil dan *Variable Importance Measure* (VIM) pada data.

II. METODE PENELITIAN

A. *Classification and Regression Trees* (CART)

CART merupakan metode statistika untuk membangun sebuah pohon keputusan untuk mengatasi permasalahan regresi pada amatan variabel respon numerik atau klasifikasi pada amatan variabel respon kategorik. Tujuan utama dari CART adalah kemampuan untuk memudahkan proses pengambilan keputusan yang kompleks menjadi lebih sederhana. Dalam membentuk pohon, setiap amatan akan dipilah di suatu *node*, sehingga akan menghasilkan kelas yang sehomogen mungkin (Guener dan Poggi, 2020: 9).

Pada Metode CART, *root node* dibagi menjadi dua simpul anak (*child node*) dengan masing-masing simpul kemudian dibagi lagi menjadi simpul anak yang baru. Dalam menyusun sebuah pohon keputusan, ada sebuah proses pemilahan (*splitting*), dimana atribut harus ditanyakan di suatu *node*. Sehingga atribut tersebut dapat memisahkan amatan menurut kelasnya. Sebuah *node* yang homogen mengandung amatan yang berasal dari kelas yang sama. Agar dapat menentukan kehomogenan pada amatan terhadap masing-masing kategori variabel respon, digunakanlah sebuah kriteria yang disebut dengan *goodness of split* yang berasal dari nilai *impurity*. *Gini index* merupakan salah satu nilai *impurity* yang umum digunakan.

Misalkan s pemilah partisi dalam simpul t yang dibagi menjadi 2 simpul anak (*child nodes)* yaitu simpul kanan (t_R) dan simpul kiri (t_L). Maka perhitungan nilai *impurtiy*-nya seperti sebagai berikut (Breiman, 2001).

$$imp(t_L) = 2p_{t_L}^{(1)}p_{t_L}^{(2)} \tag{1}$$

Selanjutnya perhitungan *impurity* pada simpul t_R sebagai berikut.

$$imp(t_R) = 2p_{t_R}^{(1)}p_{t_R}^{(2)} \tag{2}$$

Jika nilai *impurity* dari t_R maupun t_L bernilai 0, maka semua amatan dalam suatu variabel prediktor berasal dari kelas yang sama. Terakhir adalah perhitungan *impurity* pada simpul t sebagai berikut.

$$imp(t) = 2p_t^{(1)}p_t^{(2)} \tag{3}$$

Misalkan $\Delta impt(s,t)$ adalah nilai *goodness of split* untuk pemilah s dalam simpul t suatu atribut. Sehingga $\Delta impt(s,t)$ dapat didefinisikan sebagai berikut (Han, *et al*, 2012: 342).

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

Atribut yang memiliki $\Delta impt(s,t)$ yang paling maksimum akan dipilih sebagai pemilah *node*. Proses *splitting* dengan kriteria *goodness of split* akan terus berlanjut hingga kelas telah sehomogen mungkin (*terminal node*).

dengan keterangan:

$p_{t_L}^{(1)}, p_{t_L}^{(2)}$: proporsi dari amatan yang termasuk dalam simpul kiri t_L berdasarkan kelas kategori 1 dan 2

$p_{t_R}^{(1)}, p_{t_R}^{(2)}$: proporsi dari amatan yang termasuk dalam simpul kanan t_R berdasarkan kelas kategori 1 dan 2

$p_t^{(1)}, p_t^{(2)}$: proporsi dari amatan keseluruhan simpul t berdasarkan kelas kategori 1 dan 2

Nilai proporsi pada simpul kiri t_L diperoleh dengan rumus sebagai berikut.

$$p_{t_L}^{(l)} = \frac{n_{t_L}^{(l)}}{n_{t_L}} \tag{5}$$

Selanjutnya, nilai proporsi pada simpul kanan t_R diperoleh dengan rumus sebagai berikut.

$$p_{t_R}^{(l)} = \frac{n_{t_R}^{(l)}}{n_{t_R}} \tag{6}$$

Perhitungan nilai proporsi pada simpul t diperoleh dengan rumus sebagai berikut.

$$p_t^{(l)} = \frac{n_t^{(l)}}{n_t} \tag{7}$$

dimana:

l : kelas variabel respon tipe binary dengan kategori 1 dan 2

$n_{t_L}^{(l)}$: amatan pada simpul kiri t_L berdasarkan kelas kategori l

$n_{t_R}^{(l)}$: amatan pada simpul kiri t_R berdasarkan kelas kategori l

DOI: 10.24036/ujsds/vol1-iss2/30

ISSN(*Online*): 2985-475X

$n_t^{(l)}$: amatan pada simpul t berdasarkan kelas kategori l

B. *Random Forest* (RF

RF merupakan pengembangan dari metode CART, proses pemilahan atribut metode RF sama dengan CART hanya saja pohon yang terbentuk di RF lebih dari 1 pohon. Konsep RF adalah untuk mengkombinasikan atau menggabungkan banyak pohon keputusan dengan tipe *binary*, yang dibangun dengan menggunakan beberapa sampel *bootstrap* yang berasal dari sampel data dan dipilih secara acak pada setiap node dalam variabel prediktor (Genuer, *et al*, 2008: 3). Data sampel *bootstrap* yang digunakan adalah 2/3 dari data *original*. Sedangkan, 1/3 data lainnya dapat dijadikan sebagai sampel *out-of-bag* (OOB). Pengambilan data sampel *bootstrap* ke-*k* akan terus dilakukan hingga pohon-pohon yang terbentuk telah berjumlah *k*.

Dalam RF terdapat istilah *tuning parameter*, pada bagian ini akan dijelaskan bagaimana pengaruh parameter dalam performa prediksi variabel dan pengukuran *Variable Importance* (VI). Terdapat *mtry* yang digunakan untuk menghitung banyaknya variabel prediktor yang terpilih pada tiap-tiap pemisah (*split*), secara *default mtry*=$\sqrt{p}$. Dimana *p* merupakan banyaknya variabel prediktor. Berikutnya adalah *number of trees* (*ntree*) yang digunakan untuk menentukan banyaknya pohon yang akan dibentuk, secara *default ntree=500*. Terakhir, ada *node size* merupakan jumlah minimum amatan dalam sebuah *terminal node*, secara *default node size=1*.

Pembentukan *tree* ke-*i* pada RF menggunakan sampel *bootstrap* ke-*i*. *Tree* akan menghasilkan *nodes* yang telah melalui proses *splitting* berdasarkan nilai *impurity* dari amatan variabel prediktor ke-*h* yang telah dibagi berdasarkan kelasnya. Setelah terbentuknya pohon sebanyak *k*, maka seluruh *tree* akan dikumpulkan dalam satu *forest*. Sehingga, diperoleh hasil klasifikasi dari masing-masing *tree* menggunakan *majority vote* atau suara keputusan terbanyak.

C. Laju Galat *Out-of-Bag* (OOB)

Sampel OOB digunakan untuk memprediksi keakuratan struktur pohon yang telah dibentuk. Jika ada sampel OOB yang tidak sesuai dengan prediksi maka hal ini dapat disebut sebagai laju galat OOB. Laju galat OOB dihitung dari proporsi misklasifikasi hasil prediksi metode RF dari seluruh amatan gugus data. Berikut adalah rumus menghitung laju galat untuk sampel OOB ke-*i* (Guener dan Poggi, 2020: 43).

$$Laju\ Galat\ OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i} \tag{8}$$

dimana, $\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i}$ merupakan jumlah dari misklasifikasi suatu amatan yang terpilih menjadi sampel OOB pada *tree* sebanyak n, bernilai 1 ketika amatan aktual tidak sama dengan amatan prediksi dan bernilai 0 lainnya. Setelah mendapatkan nilai tingkat misklasifikasi dari sampel OOB ke-*i*, selanjutnya akan dihitung rataan laju galat OOB dengan rumus sebagai berikut.

$$Laju\ Galat\ OOB = \frac{\sum OOB_i\ error\ rate}{k} x100\% \tag{9}$$

dimana *k* merupakan jumlah banyaknya pohon yang terbentuk. Semakin kecil estimasi laju galat OOB yang dihasilkan, maka prediksi pada forest akan semakin akurat dan dapat dipercaya.

D. *Variable Importance Measure* (VIM)

Menurut Guener dan Poggi (2020: 4) *forest* yang dihasilkan sangatlah besar dan rumit untuk diinterpretasikan. Sehingga ada satu solusi untuk merangkum hasil dari informasi yang diperoleh dari *forest*. Jika prediktor penting dapat diidentifikasikan, maka model RF yang dihasilkan dapat juga menyajikan sebuah metode *variable feature selection*. Agar mengetahui bagaimana sebuah variabel prediktor itu penting, terdapat beberapa perhitungan dari VI yang sudah ditujukan.

Terdapat perhitungan untuk memperoleh nilai VIM yaitu *Mean Deacrease Gini* (MDG) adalah sebagai berikut (Sandri dan Zuccolotto, 2006).

$$MDG(x_h) = \frac{1}{k}(\sum_t [\Delta i(s,t) I(s,t)]) \tag{10}$$

dimana $\Delta i(s,t)$ gini index untuk prediktor x_h, *t* simpul ke-*t*, dan $I(s,t)$ fungsi indikator bernilai 1 jika x_h memilah simpul t dan bernilai 0 untuk lainnya.

Mean Decrease Accuracy (MDA) merupakan perhitungan yang dapat dilakukan selain menggunakan rumus MDG.

$$MDA(x_h) = \frac{1}{k}\sum_{t=1}^{k} \frac{\left(\sum_{i \epsilon OOB} I(y_i = f(x_i)) - \sum_{i \epsilon OOB} I(y_i = f(x_i^j))\right)}{|OOB|} \tag{11}$$

DOI: 10.24036/ujsds/vol1-iss2/30

ISSN(*Online*): 2985-475X

dengan $t \in \{1,2,3,\dots,k\}$, tingkat kepentingan variabel x_h dalam pohon t adalah nilai dari perbedaan antara kelas prediksi sebelum permutasi x_h, yaitu $y_i = f(x_i)$, dan seteleha variabel x_h, yaitu $y_i = f(x_{i_i}^{j})$ dalam i pengamatan tertentu. Semakin besar nilai MDA atau MDG yang dihasilkan maka variabel predictor tersebut semakin penting keberadaannya dalam mengklasifikasikan data.

E. Penanganan Missing Value

Alasan terjadinya *missing value* atau amatan data yang hilang adalah tidak terkumpulnya beberapa informasi secara lengkap. Sehingga hal ini dapat terjadi pada data sekunder. Terjadinya *missing value* dapat menyebabkan informasi yang dikumpulkan sulit untuk dilakukan analisis. Salah satu cara penanganan *missing value* adalah dengan menghitung nilai pengganti (*imputation*). Pada paket randomForest menyediakan suatu fungsi yang dapat mengatasi *missing value* pada data. Fungsi ini adalah *rfImpute()*, dimana fungsi ini bekerja dengan cara *imputation* atau menghitung nilai amatan yang paling mendekati pada amatan lainnya.

F. Jenis Penelitian dan Sumber Data

Penelitian ini merupakan penelitian terapan (*applied research*). Jenis data yang digunakan adalah data sekunder. yang berasal dari rekam medis dan informasi pasien RSUP Dr. M. Jamil Padang.

Variabel penelitian terdiri dari 23 variabel prediktor yaitu, Jenis Kelamin, Umur, Indeks BMI, Tekanan Darah, Kadar Kolesterol, Kadar Gula Darah Sewaktu (GDS), Kadar Gula Darah Puasa (GDP), Kadar Gula Darah 2 Jam Pasca Puasa (GD2PP), Rasa Nyeri, Riwayat Alergi, Penurunan BB, Kondisi Lemah, Tampak Kurus, Status Merokok, Riwayat Penyakit Dm Pada Keluarga, Riwayat Penyakit Terdahulu, *Junkfood*, *Polyuria*, *Polydipsia*, *Polyphagia*, Luka Sukar Sembuh, dan Sering Mengalami Gatal-gatal. Sedangkan 1 variabel respon yaitu tipe DM dengan kategori DM tipe I dan tipe II.

G. Teknik Analisis Data

Analisis data dalam penelitian ini menggunakan paket *randomForest* dalam *software RStudio*. Langkah-langkah dari masing-masing tahap dijelaskan sebagai berikut.

1. Menyiapkan data yang akan diteliti yaitu data diabetes melitus dari RSUP Dr. M. Jamil Padang.
2. Melakukan eksplorasi data.
3. Untuk *i*=1 sampai dengan *ntree*, lakukan langkah berikut.
   a. Pilah *n* amatan 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* dengan sistem pengembalian (*replacement*) pada kelompok 1 untuk membangun pohon ke-*i*.
   c. Amatan yang termasuk pada kelompok 2, dijadikan sebagai sampel OOB ke-*i* yang akan digunakan untuk prediksi.
   d. Dari sampel *bootstrap* ke-*i*, bangun pohon ke-*i* dengan langkah sebagai berikut.
      i. Hitung nilai *mtry* yang akan digunakan untuk menentukan jumlah variabel prediktor dalam memilah tiap simpul. Dimana $mtry_1 = \sqrt{p},\ mtry_2 = 2\sqrt{p},$ dan $mtry_3 = \frac{\sqrt{p}}{2}$, dengan $mtry \ll p$ (jumlah variabel).
      ii. Hitung nilai *impurity* simpul anak kiri dan kanan menggunakan Rumus 1 dan 2.
      iii. Hitung nilai *impurity* simpul menggunakan Rumus 3.
      iv. Mencari pemilah terbaik menggunakan *goodness of split* menggunakan Rumus 4.
   e. Untuk sampel OOB ke-*i,* lakukan langkah sebagai berikut.
      i. Pohon ke-*i* yang terbentuk, hitung laju galat OOB ke-*i* menggunakan Rumus 8.
      ii. Hitung laju galat OOB menggunakan Rumus 9.
   f. Selesai.
4. Tentukan model yang paling optimal diantara tuning parameter RF *mtry* dan *ntree* yang dikombinasikan.
5. Mengidentifikasi variabel penting yang memengaruhi variabel y dengan menggunakan nilai MDG dan MDA pada Rumus 10 dan Rumus 11 dan selanjutnya mengurutkan *variable importance.*
6. Kesimpulan.

III. HASIL DAN PEMBAHASAN

Eksplorasi data dilakukan untuk melihat ringkasan data secara umum mengenai variabel penelitian. Terdapat 501 amatan dengan 23 variabel prediktor dan 1 variabel respon. 5 diantara 23 variabel prediktor terdapat sejumlah amatan

DOI: 10.24036/ujsds/vol1-iss2/30

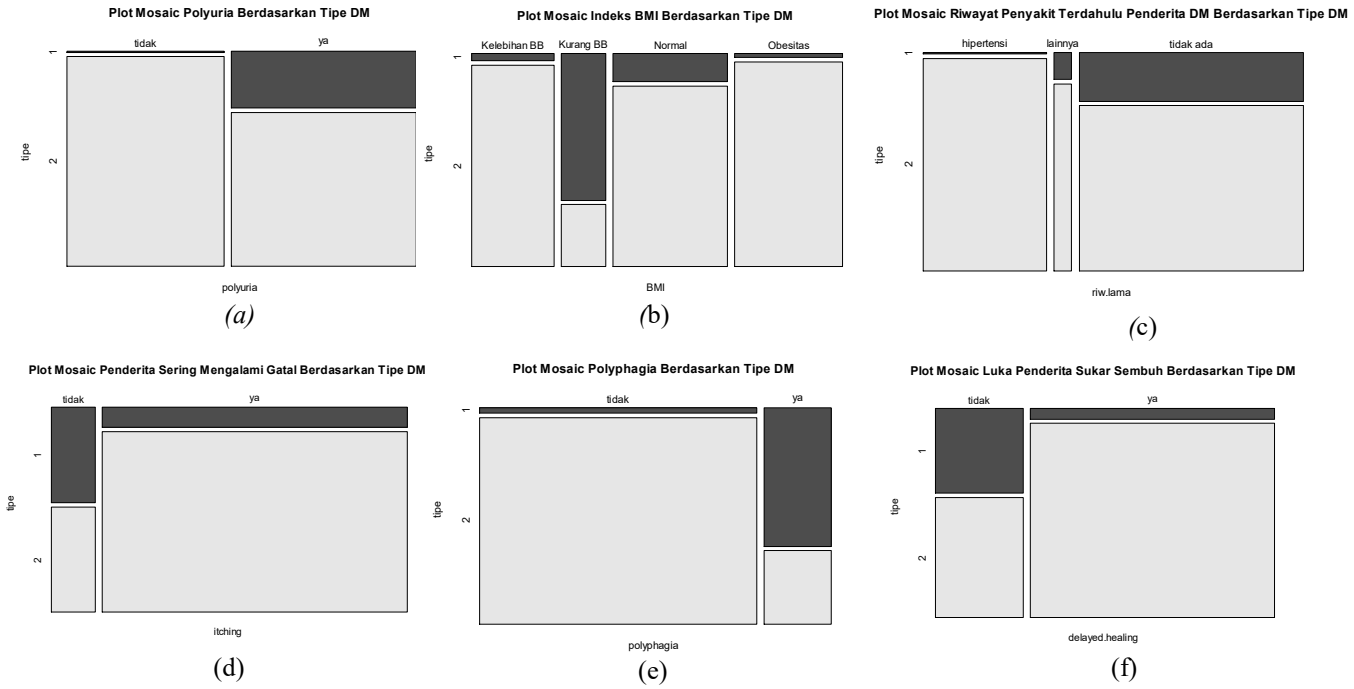
ISSN(*Online*): 2985-475X

yang hilang (*missing value*). Sehingga dengan menggunakan fungsi *rfImpute()* yang merupakan bagian dari paket *randomForest*, amatan yang hilang diganti dengan nilai rataan atau modus atau nilai yang paling mendekati amatan lainnya. Pada Tabel 1 menyajikan analisis deskriptif 6 variabel prediktor yang memiliki tipe data rasio sebagai berikut.

Tabel 1. Deskripsi Data pada Variabel Prediktor dengan Tipe Data Rasio

Variabel Respon	Variabel Prediktor	Rataan	Missing Value
DM Tipe I	Umur (Tahun)	14	0
	Hemoglobin (mmHg)	13,23	12
	Kolesterol (mg/dl)	148,9	52
	GDS (mg/dl)	383,8	9
	GDP (mg/dl)	111,5	23
	GD2PP (mg/dl)	245,4	22
DM Tipe II	Umur (Tahun)	56	0
	Hemoglobin (mmHg)	10,24	15
	Kolesterol (mg/dl)	151,8	74
	GDS (mg/dl)	292,1	14
	GDP (mg/dl)	138,9	50
	GD2PP (mg/dl)	251,7	42

Pada Tabel 1 menampilkan deskripsi data untuk variabel prediktor dengan tipe data rasio saja, hal ini dikarenakan variabel prediktor dengann tipe kategorik tidak memberikan informasi yang jelas jika dideskripsikan melalui rataan. Berkas rekam medis yang tidak lengkap atau berkas yang tidak diisi oleh petugas RS dapat menjadi salah satu faktor amatan pada variabel prediktor mengalami *missing value*. Pada kasus data DM di RSUP Dr. M. Jamil, rata-rata penderita DM tipe I berada di kalangan usia anak-anak dan penderita DM tipe II berada di usia paruh baya dan lansia. Sedangkan kadar gula penderita DM tipe I maupun tipe II berada diluar batas normal. Variabel prediktor dengan tipe data nominal maupun ordinal, tidak memiliki amatan yang hilang seperti amatan di variabel prediktor bertipe data rasio. Berikut adalah gambaran beberapa variabel prediktor yang ditampilkan dalam bentuk visualisasi yang dapat dilihat pada Gambar 1.



DOI: 10.24036/ujsds/vol1-iss2/30

ISSN(*Online*): 2985-475X

Gambar 1. (a) Plot mosaic polyuria, (b) Plot mosaic indeks BMI, (c) Plot mosaic Riwayat penyakit terdahulu penderita DM, (d) Plot mosaic penderita sering mengalami gatal, (e) Plot mosaic polyphagia, (f) Plot mosaic luka penderita sukar sembuh.

Kejadian sering mengalami buang air kecil (*polyuria*) cenderung sering dialami oleh penderita DM tipe I maupun tipe II. Pada *Body Mass Index* (BMI) yang kurang dari normal cenderung adalah penderita DM tipe I, sedangkan penderita DM tipe II cenderung memiliki BMI yang melebihi normal. Jika berdasarkan riwayat penyakit terdahulu penderita, penderita DM tipe II cenderung mengidap hipertensi sebelumnya. Kejadian sering mengalami gatal-gatal (*itching*) sering dialami oleh penderita DM tipe II. Selanjutnya, kejadian sering merasakan lapar sering dialami oleh penderita DM tipe I. terakhir, pada penderita yang mengalami luka yang sukar sembuh cenderung dirasakan oleh penderita DM tipe II. Perbedaan *polyuria*, BMI, riwayat penyakit terdahulu, *itching*, *polyphagia*, dan luka sukar sembuh yang terjadi pada tipe DM menunjukkan bahwa terdapat hubungan variabel prediktor tersebut terhadap tipe DM.

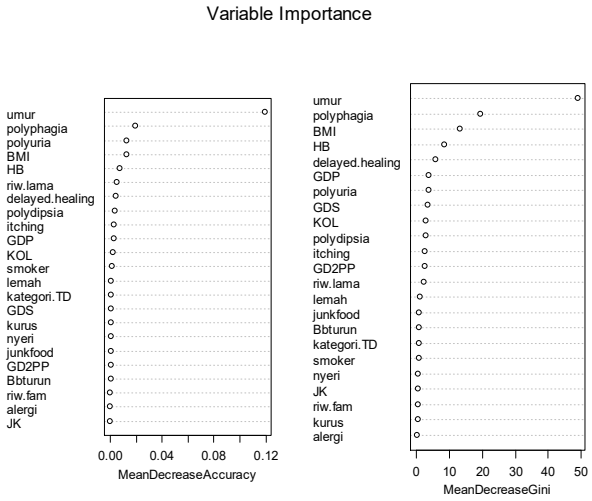
Setelah dilakukan eskplorasi data, selanjutnya dilakukan analisis RF pada data. Setiap pembentukan pohon ke-k data akan dibagi menjadi dua bagian, dimana 2/3 sampel *bootstrap* ke-i dan 1/3 lainnya adalah sampel OOB ke-i. *mtry* yang digunakan pada penelitian ini adalah $mtry_1 = 4, mtry_2 = 9$, dan $mtry_3 = 2$ dengan *ntree* yang akan digunakan adalah sebanyak 100, 250, 500, dan 1000 pohon. Nilai laju galat OOB (%) masing-masing percobaan diperoleh seperti yang disajikan pada Tabel 2 berikut:

Tabel 2. Laju Galat OOB pada Data

Mtry	Ntree			
	100	**250**	**500**	**1000**
2	1,8%	1.8%	1,6%	1,6%
4	1,6%	1,4%	1,2%	1,2%
9	1,4%	1,6%	1,4%	1,4%

Pembentukan pohon yang paling optimal adalah ketika *mtry* dan laju galat OOB bernilai kecil dan/atau *ntree* bernilai besar. Sehingga beberapa pembentukan RF yang telah dibentuk dapat disimpulkan bahwa pembentukan RF dengan *mtry* adalah 4, *ntree* adalah 1000, dan laju galat OOB adalah sebesar 1,2%. Tingkat keakuratan mencapai 98,8% dari model yang dibentuk oleh metode RF pada kasus data penyakit DM RSUP Dr. M. Jamil Padang.

Selanjutnya pada Gambar 2 menyajikan hasil nilai VIM pada variabel data. Pengukuran dengan menggunakan MDA, variabel umur, *polyphagia*, polyuria, HB dan indeks BMI menjadi VIM dalam pengklasifikasian penyakit DM. sedangkan dengan menggunakan MDG, variabel Umur, *polyphagia*, BMI, HB, dan *delayed healing* menjadi VIM dalam pengklasifikasian penyakit DM.



Gambar 2. VIM pada kasus data

Pada variabel riwayat alergi, jenis kelamin, riwayat penyakit pada keluarga penderita tidak memberikan pengaruh yang penting terhadap pengklasifikasian penyakit DM di RSUP Dr. M. Jamil Padang. Melalui nilai VIM akan diketahui variabel apa saja yang memiliki tingkat kepentingan paling tinggi berdasarkan masing-masing kategori dari variabel respon yang disajikan pada Tabel 3 sebagai berikut.

DOI: 10.24036/ujsds/vol1-iss2/30

Tabel 3. VIM pada Kategori Variabel Respon Data

No	Y= DM tipe I	Y= DM tipe II
1	Umur < 30 tahun	Umur >= 30 tahun
2	*Polyuria*	*Polyphagia*
3	*Polyphagia*	BMI = Kelebihan BB/Obesitas
4	BMI = Kurang BB	*Itching*
5	HB (rendah)	HB (rendah)

Berdasarkan Tabel 3 dapat disimpulkan bahwa penderita DM tipe I cenderung berumur dibawah 30 tahun, sering mengalami BAK dan lapar yang berlebihan, berat badan yang kurang dari batas normal, dan mempunyai HB yang rendah dari batas normal. Sedangkan, penderita DM tipe II cenderung berumur diatas 30 tahun, sering mengalami lapar dan gatal-gatal yang berlebihan, dan mempunyai HB rendah dari batas normal.

Berdasarkan penelitian-penelitian sebelumnya, RF baik digunakan untuk mengklasifikasikan data yang berukuran besar dan keakuratan prediksi yang dihasilkan sangat baik. Sehingga hal ini sangat direkomendasikan untuk mengklasifikasikan penyakit DM pada kasus data Diabetes Melitus di RSUP Dr. M. Jamil Padang.

Metode RF menghasilkan nilai prediksi yang sangat baik untuk digunakan dalam mengklasifikasikan penyakit DM di RSUP Dr. M. Jamil Padang. Hal ini disebabkan karena nilai laju galat yang dihasilkan adalah kecil. Kriteria laju galat yang digunakan dalam penelitian ini adalah laju galat dari sampel OOB. Nilai laju galat OOB yang dihasilkan memiliki nilai sebesar 1,2% atau tingkat keakurasian prediksi yang dihasilkan sebesar 98,8%.

IV. KESIMPULAN

Klasifikasi yang dilakukan oleh metode RF pada kasus data penyakit DM di RSUP Dr. M. Jamil Padang menghasilkan laju galat OOB sebesar 1,2% atau tingkat keakurasiannya mencapai 98,8%. Model yang paling optimal ini dihasilkan dengan menggunakan *mtry= 4* dan *ntree= 1000*. Penggunaan VIM, menghasilkan informasi bahwa variabel umur (< 30 tahun), *polyuria* (sering BAK), *polyphagia* (sering merasakan lapar), BMI (Kurang BB), dan kadar HB menjadi 5 variabel teratas dalam mendeteksi penyakit DM tipe I. Sedangkan variabel umur (≥30 tahun), polyphagia, BMI (Kelebihan BB atau Obesitas), *itching* (sering mengalami gatal-gatal), dan kadar HB menjadi 5 variabel teratas dalam mendeteksi penyakit DM tipe II.

DAFTAR PUSTAKA

Al-Quraishi, T., Abawajy, Jemal.H., Chowdhury, M.U., Rajasegara, S., dan Abdalrada, A.S. 2018. Breast Cancer Recurrence Prediction Using Random Forest Model. *Springer International Publishing*., 318-329, DOI: 10.1007/978-3-319-72550-5_31.

Ayyadevara, V.K. 2018. *Pro Machine Learning Algorithms*. India.

Bawono, Agus., Malini, Hema., Lenggogeni, Devia Putri., dan Rahmah, Siti. 2021. Korelasi *Illness Perception* dan *Self Care* Diabetes Melitus tipe 2 di Puskesmas Kota Padang. *Jurnal Penelitian Kesehatan Suara Forikes*., Vol 12(4), E-ISSN: 2502-7778.

Biau, G., dan Scornet, E. 2016. A Random Forest Guided Tour. DOI:10.1007/s11749-016-0481-7.

Breiman, L. 2001. Random Forest. UC Berkeley, Department of Statistics. *Machine Learning,* 45, 5-32.

Decroli, Eva. 2019. *Diabetes Melitus Tipe 2*. Padang: Pusat Penerbitan Bagian Ilmu Penyakit Dalam Fakultas Kedokteran Universitas Andalas.

Genuer, R., Poggi, J.M. 2020. *Random Forest with R*. Switzerland: Springer.

Genuer, R., Poggi, J.M., dan Tuleau, C 2008. *Random Forest: Some Methodological Insight*. France: INRIA.

Han, J., Kamber, M., dan Pei, J. 2012. *Data Mining: Concepts and Techniques*. 3rd edition, USA: Elsevier.

James, G., Witten, D., Hastie,T., dan Tibshirani, R. 2013. *An Introduction to Statistical Learning: with Application in R*. New York: Springer.

Louppe, G. 2014. "Understanding Random Forest: From Theory to Practice", *Disertasi*, 226 Hal., University of Liege. Oktober 2014.

DOI: 10.24036/ujsds/vol1-iss2/30

Makarim, Fadli Rizal. 2022. Diabetes Tipe 1. https://www.halodoc.com/kesehatan/diabetes-tipe-1. (diakses tanggal 7 Juli 2022 pukul 11.50)

Marsland, Stephen. 2015. *Machine Learning: An Algorithmic Perspective*. 2sc edition, UK: CRC Press.

P2PTM Kementerian Kesehatan RI. 2020. Tetap Produktif, Cegah, dan Atasi Diabetes Melitus. ISSN: 2442-7659.

Pahlevi, Reza, 2021. Jumlah Penderita Diabetes Indonesia Terbesar ke Lima di Dunia.https://databoks.katadata.co.id/datapublish/2021/11/22/jumlah-penderita-diabetes-indonesia-terbesar-kelima-di-dunia#:~:text=Jumlah%20Pengidap%20Diabetes%20Berdasarkan%20Negara%202021&text=Tiongkok%20menjadi%20negara%20dengan%20jumlah,Amerika%20Serikat%2032%2C22%20juta. (diakses tanggal 13 Juni 2022 pukul 20.00)

Pittara. 2021. Penyebab Diabetes Tipe 2. https://www.alodokter.com/diabetes-tipe-2%2Fpenyebab. (diakses tanggal 7 Juli 2022 pukul 11.43)

Probst, P., Wright, M., dan Boulesteix, A.L. 2019. Hyperparameters and Tuning Strategies for Random Forest. DOI: 1804.03515v2 [stat.ML].

Qanita, Eny. 2011. *Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia 2011*: Perkumpulan Endokrinologi Indonesia.

Sabariah, K.M., Hanifa, Aini., dan Siti Sa'adah. 2014. "Early detection of Type II Diabetes Mellitus with Random Forest and Classification and Regression Tree (CART)". Proceeding of IEEE 2014: International Conference of Advanced Informatics: Concept, Theory, and Application (ICAICTA), (20-21 August 2014, Bandung, Indonesia), 238-242.

Sandri, M., dan Zuccolotto, P. 2006. Variable Selection Using Random Forest, dalam Ramadhan, A. 2019. *Pemodelan Klasifikasi Random Forest untuk Mengidentifikasi Faktor Penting dalam Meningkatkan Mutu Pendidikan*. *Tesis*, 34 Hal., Institut Pertanian Bogor, Bogor, Indonesia, 01 Agustus 2019.

Scornet, E., Biau, G., dan Vert, J.P. 2015. Consistency of Random Forest. *The Annals of Statistics*. Vol. 43 (4), 1716-1741. DOI: 10.1214/15-AOS1321.

Sen, Saikat., Chakraborty, R., dan De, B. 2016. *Diabetes Melitus in 21st Century*. Singapore: Springer.

Siyoto, S., dan Sodik, M. Ali. 2015. *Dasar Metodologi Penelitian*. Yogyakarta: Literasi Media Publishing.

Srivastava, Rachit., Tiwari, A.N., dan Giri, V.K. 2019. Solar Radiation Forecasting Using MARS, CART, M5, and Random Forest Model: A Case Study for India. *Elsevier*. 1-14. https://doi.org/10.1016/j.heliyon/2019.e02692.

Syahrum, dan Salim. 2014. *Metodologi Penelitian Kuantitatif*. Bandung: Citapustaka Media.

Vrtkova, A., dan Prochazka V. 2019, "Comparing the Performance of Regression Model, Random Forest, and Neutral Networks for Stroke Patient's Outcome Prediction". Proceeding of IEEE 2019: International Conference on Information and Digital Technologies, (25-27 June 2019, Zilinia, Slovakia), University of Ostrava, 2019. 543-550.

Zhang, H., dan Singer, B.H. 2010. *Recursive Partitioning and Applications*. 2sc edition, USA: Springer.

Zhou, L., Wang, Q., dan Fujita, H. 2017. One Versus One Multi-class Classification Fusion Using Optimizing Decision Directed Acyclic Graph for Predicting Listing Status of Companies. 36(1): 80-89, dalam Ramadhan, A. 2019. *Pemodelan Klasifikasi Random Forest Untuk Mengidentifikasi Faktor Penting Dalam Meningkatkan Mutu Pendidikan*. *Tesis*, 34 Hal., Institut Pertanian Bogor, Bogor, Indonesia, 01 Agustus 2019.