UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 130-136.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Dental Caries in RSGM Baiturrahmah Using the Random Forest Method

Martia Rosada, Zilrahmi*, Syafriandi, dan Tessy Octavia Mukhti
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: zilrahmi@fmipa.unp.ac.id

Submitted : 26 Februari 2024
Revised : 22 Maret 2024
Accepted : 29 Mei 2024

ABSTRACT

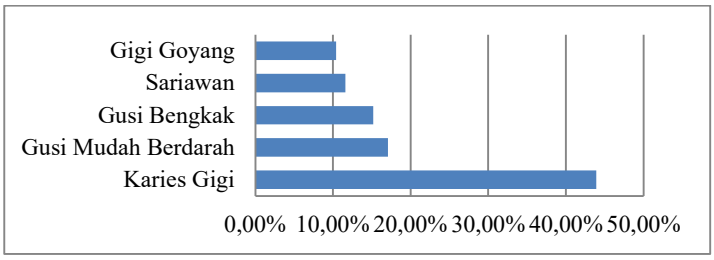
The mouth cavity is the main gate through which for germs and bacteria enter which can harm health. The dental and oral problems that many people experience are caries or cavities. The problem of dental caries in West Sumatera society has a fairly high prevalence rate. Prevention of dental caries needs to be done by making people aware of the importance of maintaining oral hygiene. Therefore, there is a need for a method that is able to classify dental caries based on its symptoms. The classification method is very useful for finding out the main factors that cause dental caries, one of which is random forest. Random forest is an ensemble method, namely the development of several decision tree methods using bootstrap sampling. The results of this research use the smallest OOB level and Variable Importance Measure (VIM). Random forest classification using dental and oral pain medical record data at Baiturrahmah Padang Hospital produced an OOB error rate of 29.52% or an accuracy rate of 71%. The optimal model is obtained using mtry=2 and ntree=500. From this research, it can be concluded that dental plaque, age, and tooth brushing habits are important variables or main factors that influence dental caries.

Keywords: *Random Forest, VIM, OOB*

I. PENDAHULUAN

Kesehatan yang perlu diperhatikan selain kesehatan tubuh secara umum, juga kesehatan gigi dan mulut. Faktor utama yang mempengaruhi kesehatan gigi dan mulut yaitu tingkat kebersihannya. Hal tersebut dapat dilihat secara klinis dari ada tidaknya sisa-sisa makanan yang menempel pada gigi, seperti pelikel, materi alba, debris, kalkulus, dan plak gigi (Sintawati, 2008). Ketika kebersihan gigi dan mulut tidak terjaga maka akan menyebabkan masalah atau penyakit gigi dan mulut seperti, periodontitis, karies gigi, abses gigi, gingivitis dan masalah kesehatan gigi dan mulut lainnya. Menurut Listrianah (2017) masalah gigi dan mulut yang banyak dialami oleh masyarakat adalah karies gigi. Karies gigi terjadi karena rusaknya jaringan keras gigi akibat mikroorganisme dalam plak yang menyebabkan demineralisasi

Permasalahan karies gigi di Indonesia mempunyai tingkat prevalensi yang cukup tinggi, dari 265 juta jiwa penduduk Indonesia sebanyak 45,3% mengalami karies gigi. Di Sumatera Barat sendiri juga memiliki prevalensi karies gigi yang cukup tinggi, dari 5.519.245 jiwa penduduk Sumatera Barat sebanyak 43,9% mengalami karies gigi (Riskesdas, 2018). Gambar 1 menyajikan persentase penyakit gigi dan mulut di Sumatera Barat.



(Sumber: Riskesdas, 2018)

Gambar 1. Persentase penyakit gigi dan mulut di Sumbar

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Berdasarkan Gambar 1 dapat terlihat bahwa masalah karies gigi merupakan penyakit yang paling banyak dialami oleh masyarakat Sumatera Barat. Sebanyak 43,90% dari 5.519.245 jiwa masyarakat Sumatera Barat mengalami karies gigi. Hal ini berarti dari 5.519.245 jiwa penduduk 2.422.948 jiwa mengalami karies gigi. Namun, banyak masyarakat yang tidak peduli dengan kesehatan giginya karena tidak menyadari risiko dari karies gigi. Maka perlunya meningkatkan kesadaran masyarakat dengan melakukan edukasi tentang faktor-faktor yang mempengaruhi terjadinya karies gigi. Dengan meningkatnya kesadaran masyarakat tentang pentingnya perawatan gigi dan mulut di Sumatera Barat maka masalah karies gigi dapat berkurang. Oleh karena itu, perlu adanya metode yang mampu mengklasifikasi karies gigi berdasarkan gejalanya. Metode klasifikasi sangat berguna untuk mengetahui faktor utama yang menjadi penyebab karies gigi. Selain itu, metode klasifikasi mampu untuk mengelola data yang komplek dan memprioritaskan faktor resiko terjasdinya karies gigi.

Random forest adalah salah satu metode dalam klasifikasi yang digunakan untuk data dengan jumlah yang besar. Metode *random forest* memiliki beberapa kelebihan antara lain, klasifikasi yang baik, menghasilkan nilai *error* yang lebih rendah, dan secara efisien dapat mengatasi data *training* dengan jumlah data yang sangat besar atau *overfitting* (Breiman, 2001). *Random forest* merupakan metode *ensemble* yaitu pengembangan dari beberapa metode yang mampu menghasilkan pohon klasifikasi lebih baik dengan tingkat keakuratan yang tinggi. Hal ini juga dibuktikan berdasarkan penelitian terdahulu yang dilakukan oleh Hanum dan Zailani (2020) tentang Penerapan Algoritma Klasifikasi *Random Forest* untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera dengan hasil akurasi sebesar 87,88%.

Tujuan dari penelitian ini untuk mengetahui variabel penting dalam data dan untuk mengetahui tingkat akurasi yang dihasilkan oleh *random forest* dalam klasifikasi jenis karies gigi di rumah sakit gigi dan mulut baiturrahmah Padang.

II. METODE PENELITIAN

A. Jenis Penelitian dan Sumber Data

Jenis penelitian yang diterapkan adalah penelitian terapan. Jenis data pada penelitian adalah data sekunder. Data yang digunakan merupakan data rekam medis "sakit gigi dan mulut" yang diperoleh dari rumah sakit gigi dan mulut baiturrahmah. Data yang diteliti dari bulan Januari-Oktober 2023 dengan 300 amatan. Variabel penelitian terdiri dari 6 variabel prediktor yaitu jenis kelamin, umur, frekuensi menyikat gigi dalam sehari, status pasien, plak gigi, dan tempat tinggal. Sedangkan 1 variabel respon yaitu jenis karies dengan kategori parah dan tidak parah.

B. Langkah-langkah Analisis

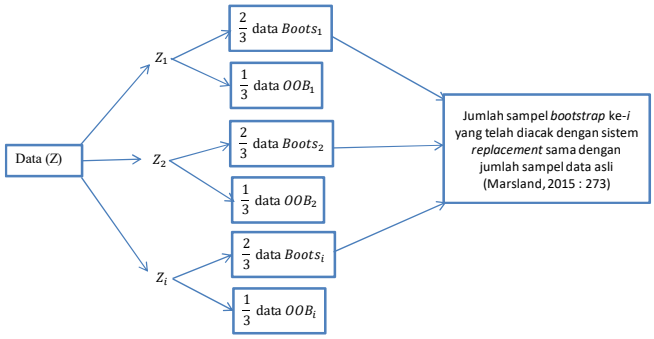
Analisis data yang digunakan pada penelitian ini adalah algoritma random forest dalam *software RStudio*. Langkah-langkah dari masing-masing tahap dijelaskan sebagai berikut:

1. Menyiapkan data yang akan diteliti yaitu data jenis penyakit gigi dan mulut dari RSGM Baiturrahmah Padang, yaitu sebanyak 300 amatan dari 5 variabel prediktor dan 1 variabel respon.
2. Melakukan pembagian data *training* dan *testing.* Data *training* yang digunakan yang digunakan 70% atau sebanyak 201 data dan untuk data *testing* 30% atau sebanyak 91 data.
3. Pengambilan sampel *bootstrap* yaitu pengambilan sampel secara acak berukuran *n* dari kumpulan data *training* dengan pengembalian. Dengan jumlah data pada *bootstrap* sama dengan data *training*.

Random forest memiliki dua konsep dasar yaitu membangun *ensamble* dari *tree* via *bagging* dengan pengembalian dan menyeleksi fitur secara acak untuk tiap *tree* yang dibangun. *Bagging* merupakan singkatan dari *bootstrap aggregating,* yaitu salah satu metode *ensemble* yang digunakan untuk mereduksi variansi variabel prediktor sehingga dapat memperbaiki kualitas prediksi dari pohon klasifikasi. Ide dasar *bagging* adalah memisahkan data *training* menjadi beberapa data *training* baru dengan pengembalian sampel secara acak dan membuat model untuk data t*raining* baru (Breiman, 1996). Pembagian sampel *training* dan *testing* pada *random forest* menggunakan 2/3 data asli untuk sampel *training* dan 1/3 dari data asli untuk *testing*. Menurut Breiman (2001: 11), penggunaan data untuk sampel *bootstrap* adalah 2/3 dari data asli. Selanjutnya, 1/3 data asli lainnya disebut sampel *Out Of Bag* (OOB). Sampel ini digunakan untuk memprediksi keakuratan struktur pohon yang terbentuk. Ilustrasi diagram untuk pembagian sampel *bootstrap* dan sampel OOB dari data sampel asli disajikan pada gambar 2.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Sumber : (Marsland, 2015 : 273)

Gambar 2. Pembagian sampel *bootstrap* dan sampel OOB dari data asli

Penentuan nilai *error* pada hasil prediksi *random forest* dapat diperoleh dengan menggunakan OOB (*Out Of Bag*) laju galat (*error rate*) yang dihitung dari hasil proporsi kesalahan prediksi klasifikasi setiap amatan gugus data asli dari hasil prediksi *random forest* (Janitza, 2018). Pada proses ini melibatkan pembentukan subset data dengan *bootstrap* sampling untuk setiap keputusan, dimana data yang tidak terpilih pada sampel *bootstrap* menjadi OOB. Nilai OOB yang telah didapatkan tadi digunakan untuk menghitung prediksi, dan kelas mayoritas prediksi OOB dibandingkan dengan kelas asli untuk menghitung proporsi kesalahan.

4. Melakukan pemilihan peubah penjelas X_p sebanyak parameter *mtry* (banyak variabel kandidat yang dipilih secara acak), pengambilan *subset* prediktor sebanyak *m* secara acak, dimana $m<p$ (jumlah variabel prediktor) untuk membentuk satu *decision tree* dengan menggunakan persamaan *entropy* dan *information gain* seperti pada Persamaan 1 dan Persamaan 2.

Analisis *random forest* mempunyai tiga parameter utama yang digunakan yaitu *mtry* (banyak input variabel secara acak terpilih dalam satu *node* pemilah), *ntree* (jumlah banyaknya *tree* dalam *forest*) dan *node size* (jumlah amatan minimum dalam sebuah *terminal node*) (Genuer R, 2008:5). Untuk perhitungan nilai mtry menggunakan $mtry_1=\sqrt{p}$, $mtry_2=2\sqrt{p}$, dan $mtry_3=\frac{\sqrt{p}}{2}$. Pembentukan jumlah pohon *(ntree)* yang digunakan adalah 500, namun terdapat empat nilai parameter *ntree* yang bisa digunakan yaitu 100, 200, 500 dan 1000.

Pembentukan pohon keputusan dalam *random forest* dimulai dengan pencarian nilai *mtry* yang tertinggi pada masing-masing variabel prediktor yang terpilih dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dari nilai *information gain*. Pencarian nilai *entropy* menggunakan rumus:

$$Entropy\ (Y) = -\sum_{i=1}^{n} p(c\,|\,Y) log_2 p(c\,|\,Y) \quad (1)$$

(Schouten, 2016)

Keterangan:
Y : Jumlah variabel prediktor terpilih
n : jumlah partisi himpunan i
$p(c\,|\,Y)$: Proporsi nilai Y terhadap kelas c (kategori dari variabel prediktor)

Selanjutnya mencari nilai dari *information gain* yang digunakan untuk mengukur efektivitas suatu atribut dalam pengklasifikasian data dapat dihitung dengan rumus:

$$Information\ Gain\ (Y, \alpha) = Entropy\ (Y) - \sum_{i=1}^{n} \frac{Y_i}{Y} Entropy\ (Y_i) \quad (2)$$

Keterangan:
Y : Himpunan amatan
α : Atribut
n : Jumlah partisi atribut i
Y_i : Jumlah kasus pada partisi ke-i

5. Setelah *tree* dan *forest* terbentuk, dilanjutkan dengan menghitung tingkat missklasifikasi menggunakan sampel OOB berdasarkan *mtry* dan *ntree* menggunakan Persamaan 3 untuk satu *tree* dan menghitung nilai laju galat OOB menggunakan Persamaan 4 dan memilih kombinasi yang memiliki nilai OOB *error rate* terkecil.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Laju galat OOB dihitung dari proporsi klasifikasi hasil prediksi metode *random forest* dari seluruh amatan gugus data. Nilai *error* klasifikasi *random forest* diperoleh dengan rataan *error* sampel OOB-I. Berikut persamaan laju galat untuk sampel OOB ke-I (Guener dan Poggi, 2020).

$$Laju\ Galat\ OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i} \quad (3)$$

Keterangan:
$\sum_{i=1}^{n} 1_{Y_i \neq \hat{Y}_i}$: jumlah data hasil prediksi yang salah (missklasifikasi)
Y_i : hasil amatan sebenarnya ke-i
$\hat{Y}_i$: hasil amatan yang diprediksi ke-i
n : jumlah OOB ke-i menjadi sampel OOB

Setelah mendapatkan nilai tingkat misklasifikasi dari sampel OOB ke-I, selanjutnya akan dihitung rataan tingkat misklasifikasi OOB dengan persamaan sebagai berikut:

$$Laju\ Galat\ OOB = \frac{\sum OOB_i\ error\ rate}{k} * 100\% \quad (4)$$

6. Tentukan model paling optimal yang dihasilkan dari *tuning* parameter *random forest mtry* dan *ntree* yang dibentuk.

Pembentukan pohon yang paling optimal ketika tingkat missklasifikasi yang dihasilkan dari OOB nilainya kecil. Semakin kecil estimasi OOB *error rate* yang dihasilkan, maka prediksi pada *forest* akan semakin akurat dan dapat dipercaya (Janitza, 2018). Terdapat nilai selang kategori keakuratan yang dapat dijadikan sebagai bahan pengukuran OOB yang ditunjukkan pada tabel berikut.

Tabel 1. Kriteria Tingkat Kesalahan Prediksi

Kategori OOB *error rate*	Penjelasan
< 10%	Kemampuan klasifikasi sangat baik
10 – 40 %	Kemampuan klasifikasi baik
50 – 75 %	Kemampuan klasifikasi cukup baik
>75 %	Kemampuan klasifikasi kurang baik

7. Mengidentifikasi variabel *importance.*

Analisis *random forest* dalam mempermudah memperoleh informasi pada penelitian adalah dengan mengidentifikasi *Variable Importance Measure* (VIM) untuk variabel prediktor. Apabila VIM dapat diidentifikasi, maka sebuah *random forest* yang dihasilkan dapat juga menghasilkan sebuah metode variabel *feature selection*. Untuk mengetahui bagaimana sebuah variabel itu penting, terdapat beberapa perhitungan dari VI yang sudah ditujukan. Menurut Zhang (2010: 84) pengukuran variabel VI menggunakan *The Gini Measure or Gini Importance.*

Berikut adalah rumus yang dapat digunakan untuk mengukur MDG (Breiman, 2001).

$$MDG(x_h) = \frac{1}{k}(\sum t[\Delta i(s,t)I(s,t)]) \quad (5)$$

dengan:
$\Delta i(s,t)$: gini index untuk peubah jelas x_h
k : banyak pohon dalam random forest
t : simpul ke- t
$I(s,t)$: fungsi indikator, bernilai 1 jika x_h memilah simpul t dan bernilai 0 untuk lainnya

8. Menarik kesimpulan.

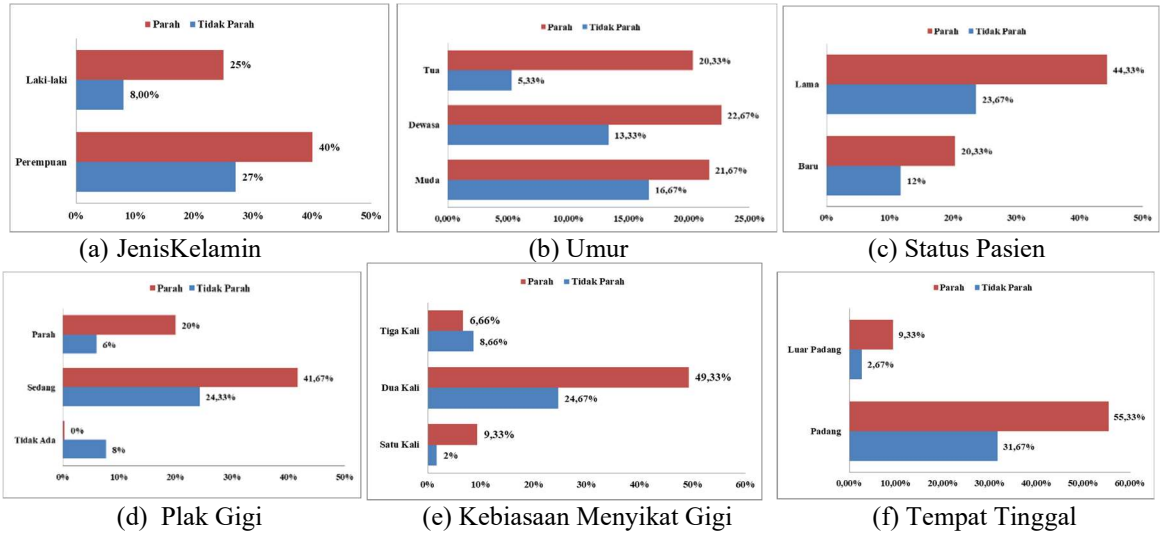
III. Hasil dan Pembahasan

A. Analisis Deskriptif

Gambaran karakteristik pasien penyakit karies gigi di Rumah Sakit Gigi dan Mulut Baiturrahmah Padang dapat dilihat pada visualisasi variabel prediktor yang ditampilkan pada Gambar 2.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



(a) JenisKelamin (b) Umur (c) Status Pasien

(d) Plak Gigi (e) Kebiasaan Menyikat Gigi (f) Tempat Tinggal

Gambar 2. Visualisasi Variabel Prediktor

Berdasarkan Gambar 2 menjelaskan mengenai variabel prediktor untuk karies gigi di rumah sakit Baiturrahmah Padang pada tahun 2023. Dari 300 pasien yang menjadi subjek penelitian, dapat diketahui bahwa pasien yang mengalami karies gigi mayoritas dialami oleh perempuan yaitu sebanyak 66,33% atau 199 pasien. Karies gigi juga banyak dialami oleh pasien kategori muda yang memiliki rentang umur 16-24 tahun. Ketika kondisi plak gigi pasien sedang dan parah maka mengalami karies giginya yang parah. Kemudian kebiasaan menyikat gigi satu kali dan dua kali sehari lebih dominan menyebabkan karies gigi parah.

B. Analisis *Random Forest*

Membangun pohon keputusan pada analisis *random forest* dilakukan pembagian data menjadi dua bagian, dimana 2/3 untuk data *training* dan 1/3 untuk data *testing*. Setelah dilakukan pembagian data *training* dan *testing,* tahap berikutnya dari data *training* yang sudah dipilih dilakukan *bootstrap sampling* yaitu pengambilan sampel secara acak dengan pengembalian bertujuan untuk memperbaiki hasil prediksi *random forest.* Selanjutnya adalah menentukan banyaknya variabel prediktor atau nilai *mtry* yang digunakan dalam penentuan klasifikasi. Penggunaan nilai *mtry* pada penelitian ini adalah $mtry_1=2$, $mtry_2=3$, $mtry_3=4$. Dari masing-masing *mtry* tersebut kemudian dilakukan percobaan sebanyak *k* pohon (*ntree)*. Banyak pohon yang digunakan adalah 100, 200, 300, dan 500.Nilai *mtry* yang digunakan pada penelitian ini adalah $mtry_1=4$, $mtry_2=3$, $mtry_3=2$ dengan *ntree* yang digunakan adalah 100, 200, 300, dan 500. Maka dihasilkan nilai laju galat OOB (%) pada setiap percobaan seperti yang terlihat pada Tabel 3 berikut:

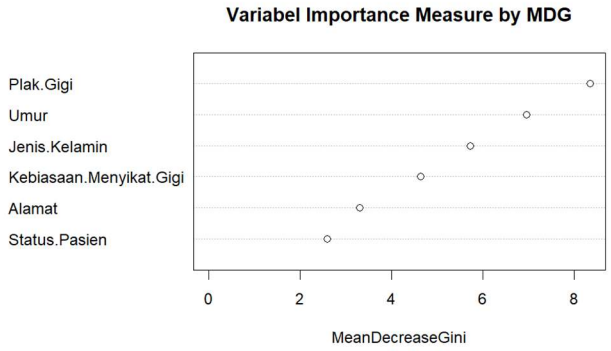
Tabel 3. Laju Galat OOB pada Data

Mtry	*Ntree*			
	100	200	300	500
2*	30,00%	31,90%	30,95%	**29,52%**
3	36,19%	37,62%	34,29%	34,76%
4	39,05%	40,00%	40%	41,43%

Pembentukan *random forest* dari beberapa pohon yang telah dicobakan diperoleh pohon optimal dengan menggunakan *mtry*=2 dan *ntree*=500 yang menghasilkan laju galat OOB atau tingkat kesalahan klasifikasi yaitu 29,52%. Sehingga dihasilkan tingkat keakuratan sebesar 71% pada kasus data karies gigi di Rumah Sakit Gigi dan Mulut Baiturrahmah. Selanjutnya akan dihitung variabel *importance* atau faktor yang mempengaruhi hasil klasifikasi terhadap terjadinya karies gigi dapat dilihat pada Gambar 3.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 3. VIM pada klasifikasi Karies Gigi

Variabel VIM yang terdapat pada Gambar 2 telah diurutkan dari nilai terbesar dan terkecil. Semakin besar nilai VIM yang dihasilkan maka variabel tersebut semakin berpengaruh terhadap karies gigi. Setelah didapatkan VI pada pembentukan *random forest* dengan menggunakan *mtry*=2 dan *ntree*=100. Berikut beberapa nilai VI yang dihasilkan terdapat pada Tabel 4.

Tabel 4. Nilai VI

VI	Nilai
Plak Gigi	8.35
Umur	5.72
Jenis Kelamin	5.72
Kebiasaan Menyikat Gigi	4.64
Alamat	3.311
Status Pasien	2.59

Nilai VI yang dihasilkan pada Tabel 2 untuk pembentukan pohon yang paling optimal yaitu dipengaruhi oleh variabel plak gigi, umur, jenis kelamin, kebiasaan menggosok gigi, dan diikuti oleh variabel lainnya. Plak gigi merupakan faktor utama yang mempengaruhi tingkat karies gigi. Semakin tinggi tingkat plak gigi seseorang maka akan semakin parah tingkat karies gigi nya.

Nilai yang diperoleh pada metode *random forest* menghasilkan prediksi yang baik dalam mengklasifikasikan jenis karies gigi di Rumah Sakit Gigi dan Mulut Baiturrahmah. Hal ini disebabkan karena *random forest* memiliki tingkat kesalahan laju galat kecil. Laju galat yang digunakan pada penelitian ini yaitu laju galat OOB. Nilai laju galat yang dihasilkan adalah 29,52% yang artinya menghasilkan tingkat akurasi sebesar 71%.

IV. Kesimpulan

Klasifikasi yang dihasilkan oleh metode *random forest* pada kasus data sakit gigi dan mulut di RSGM Baiturrahmah Padang menghasilkan laju galat OOB sebesar 29,52% yang artinya menghasilkan tingkat akurasi sebesar 71%. Model yang paling optimal dihasilkan pada penggunaan *mtry*=2 dan *ntree*=500. Menghasilkan variabel *importance* atau variabel yang penting terhadap faktor yang menyebabkan tingkat keparahan karies gigi yaitu plak gigi, umur jenis kelamin, kebiasaan menyikat gigi, dan diikuti oleh variabel lainnya. Untuk penelitian selanjutnya dapat ditambahkan beberapa atribut dan menggunakan lebih banyak data dalam mengklasifikasi pasien karies gigi. Hal ini berguna untuk menghasilkan klasifikasi yang lebih baik.

DAFTAR PUSTAKA

Breiman, L., & Friedman, J. (1984). Classification and Regression Tree. New York: Chapman and Hall.

Breiman, L. (2001). *Random Forest.* Berkeley: Statistics Department University of California, 5-32.

Genuer , R., & Poggi, J. M. (2020). *Random Forests with R.* Switzerland: Springer Nature .

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Genuer , R., & Poggi, J. M. (2008). *Random Forests: some methodological insights.* France: Inria.

Hanun, N., & Zailani, A. (2020). Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra SejahterA. *Journal Of Technology Information, 6*(1), 7-14.

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. PLoS ONE, 1-31.

Listrianah, & Putri, N. (2017). Indeks karies gigi ditinjau dari penyakit umum dan sekresi saliva pada anak di Sekolah Dasar Negeri 30 Palembang , Jurnal Kesehatan Palembang, 12(2), 136-148).

Marsland, S. (2015). *Machine Learning An Algoritmic Perspective Second Edition. UK: CRC Press.*

Riset Kesehatan Dasar (Riskesdas). 2018. Kesehatan Gigi Dan Mulut. Diakses Agustus 2023.

Schouten, K., Frasincar, F., & Dekker, R. (2016). An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis. *in Natural Language Processing and Information Systems*, 48-59.

Sintawati & Indirawati. (2008). Faktor-faktor yang mempengaruhi kebersihan gigi dan mulut masyarakat DKI Jakarta 2007. Jurnal Ekologi Kesehatan, 8 (1), 860-873.

Zhang, H., & Singer, B. H. (2010). *Recursive Partitioning and Applications Second Edition.* USA: Springer Science.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/155