Klasifikasi Jenis Persalinan pada Ibu Hamil dengan Metode Random Forest

Ayuna Armonica[1], Paulina H. Prima Rosa[2]
Prodi Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma Yogyakarta
Paingan, Maguwoharjo, Depok, Sleman, Daerah Istimewa Yogyakarta
Telp (0274) 883037, 889368, Fax (0274) 886
[1]armonicaayuna@gmail.com
[2]rosa@usd.ac.id

Abstrak – Salah satu faktor penyebab kematian ibu pada saat melahirkan adalah keterlambatan pengambilan keputusan pada saat penanganan persalinan. Untuk mengidentifikasi penanganan yang tepat dalam persalinan, dalam penelitian ini dibangun model klasifikasi jenis persalinan ibu hamil menggunakan metode random forest terhadap 302 data yang diambil dari RSUD Argamakmur. Model klasifikasi diuji dengan variasi jumlah tree 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096 pada data asli dan data yang telah dilakukan balancing. Dengan teknik cross validation, diperoleh akurasi terbaik 92,55130% pada jumlah fold 3 dan jumlah tree 64.

***Kata kunci**: jenis persalinan, kehamilan, klasifikasi, random forest.*

Abstract - One of the factors causing maternal death during childbirth is the delay in decision-making at the time of delivery. To identify the appropriate treatment in childbirth, in this study a classification model of the type of delivery for pregnant women was built using the random forest method on 302 data taken from Argamakmur Hospital. The classification model was tested with variations in the number of trees 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096 on the original data and data that had been balanced. With the cross validation technique, the best accuracy is 92.55130% on the number of folds 3 and the number of trees 64.

***Keywords**: type of childbirth, pregnancy, classification, random forest.*

I. PENDAHULUAN

Angka Kematian Ibu (AKI) dan Angka Kematian Bayi (AKB) di Indonesia masih mengkhawatirkan, berkisar 305 kasus per 100.000 kelahiran hidup menurut Ketua Komite *Ilmiah International Conference on Indonesia Family Planning and Reproductive Health* (ICIFPRH) tahun 2019. Pada tahun 2017,setiap hari sekitar 810 wanita hamil meninggal diakibatkan oleh penyebab yang sesungguhnya dapat dicegah, terkait kehamilan dan persalinan (who.int, 2018) [1].

Pada umumnya persalinan pada ibu hamil dilakukan melalui 2 (dua) proses yaitu secara normal dan *caesar*. Persalinan normal adalah proses pengeluaran bayi, plasenta dan selaput ketuban dari uterus pada usia kehamilan cukup bulan (umur kehamilan lebih dari 37 minggu) tanpa disertai penyulit (JNPK-KR, 2010) sedangkan persalinan *caesar* (*Sectio Caesarea*) didefinisikan sebagai proses lahirnya bayi dengan cara membuat sayatan pada perut dan dinding rahim [2].

Pada penelitian ini, penulis membangun sebuah model klasifikasi menggunakan metode *random forest* yang dapat mengklasifikasikan jenis persalinan (*partus*) yang akan dilakukan oleh ibu hamil. Model yang dibangun diharapkan dapat digunakan untuk membantu tenaga medis dalam proses pengambilan keputusan jenis persalinan yang akan dilakukan dalam penanganan persalinan ibu hamil.

II. METODE RANDOM FOREST

Random forest merupakan algoritma klasifikasi *supervised.* Algoritma ini menciptakan hutan dengan sejumlah pohon. Secara umum, semakin banyak tree di hutan, semakin kuat bentuk hutannya. Pada pengklasifikasi *random forest*, semakin banyak jumlah tree maka akurasi yang didapatkan semakin tinggi (Polamuri, 2017).

Random forest merupakan metode pohon gabungan yang berasal dari pengembangan metode *Classification and Regression Tree* (CART), yaitu dengan menerapkan metode *bootstrap aggregating (bagging)* dan *random feature selection* (Breiman, 2001).

Random forest sudah banyak digunakan dalam aplikasi-aplikasi dalam segala bidang kehidupan seperti bidang kesehatan, pendidikan, industri sudah banyak menerapkan metode *random forest*. Metode *random forest* dipilih karena menghasilkan kesalahan yang lebih rendah, memberikan akurasi yang bagus dalam klasifikasi, dapat menangani data *training* yang jumlahnya sangat besar, dan efektif untuk mengatasi data yang tidak lengkap (Breiman, 2001).

Bagging atau disebut juga dengan *bootstrap aggregating* merupakan metode yang dapat *memperbaiki*

hasil dari algoritma klasifikasi. *Bagging* merupakan salah satu metode yang berdasar pada *ensemble method*, yaitu metode yang menggunakan kombinasi beberapa model. *Bagging* predictor adalah metode yang digunakan untuk membangkitkan *multiple versions* dari prediktor dan menggunakannya untuk mendapatkan kumpulan prediktor. *Multiple versions* dibentuk dengan replikasi (*replacement*) *bootstrap* dari sebuah data percobaan (Breiman, 1996).

Pemodelan *random forest* dilakukan dengan cara berikut (Breiman 2001; Breiman & Cutler 2003):

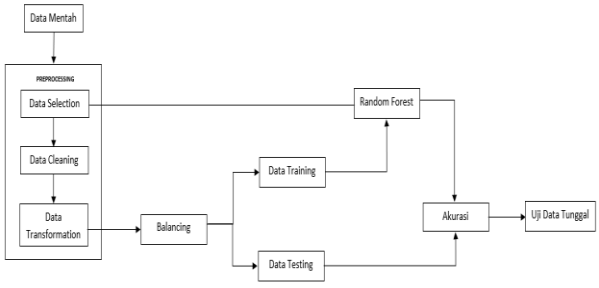
1. Lakukan penarikan contoh acak berukuran *n* dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan *bootstrap.*
2. Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih *m* peubah penjelas secara acak, dimana $m << p$. Pemilah terbaik dipilih dari *m* peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
3. Ulangi langkah 1 dan 2 sebanyak *k* kali, sehingga terbentuk sebuah hutan yang terdiri atas *k* pohon.
4. Selanjutnya *random forest* akan melakukan prediksi dengan mengkombinasikan hasil dari setiap pohon keputusan dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi.

Pembuatan pohon dalam random forest dilakukan dengan salah satu algoritma *decision tree* yang ada.

III. METODOLOGI PENELITIAN

A. *Skema Penelitian*

Skema penelitian digambarkan dalam Gambar 1 berikut ini. Data mentah yang dikumpulkan, dikenai *preprocessing* yang mencakup *data selection*, *data cleaning*, dan *data transformation*. Data hasil preprocessing selanjutnya dikenai proses balancing jika ditengarai ada ketidaksimbangan data. Selanjutnya data dipecah menjadi data training dan data testing untuk dikenai algoritma *random forest*. Hasilnya berupa akurasi. Pencarian model dengan akurasi terbaik dilakukan melalui eksperimen dengan mengubah-ubah jumlah pohon dan jumlah fold. Hasil akurasi terbaik dipilih sebagai model yang dapat dimanfaatkan untuk mengklasifikasikan data tunggal.



Gambar 1. Skema Penelitian

B. *Pengumpulan Data*

Tahapan pertama dalam penelitian ini adalah pengambilan data mentah yang diproses sesuai dengan kebutuhan dari penelitian. Data mentah yang dikumpulkan adalah data persalinan ibu hamil yang diambil dari Bangsal Kebidanan di Rumah Sakit Umum Argamakmur (RSUD) pada tanggal 18 November 2020 – 12 Februari 2021.

Data yang terkumpul berjumlah 302 baris data yang terdiri dari 25 atribut yaitu: nama, umur, usia kandungan, leukosit, limfosit, hemoglobin, trombosit, eritrosit, tekanan darah, glukosa, ureum, SGOT/SGPT, glukosa, protein, tunggal atau ganda, posisi bayi, panggul sempit, ketuban pecah, tali pusat, asma, hepatitis, HIS (kontraksi), riwayat partus, kondisi ketuban.

C. *Preprocessing*

Preprocessing merupakan tahapan yang terdiri dari proses *data selection* yang berfungsi untuk menghilangkan atribut yang tidak berpengaruh dalam analisis, *data cleaning* untuk pembersihan *noise data* serta *missing value*, selanjutnya pada *data transformation* untuk mengkonversi data terutama data yang belum numerik menjadi data numerik.

Tahap *data selection* dilakukan untuk menghilangkan atribut yang tidak berpengaruh dalam analisis. Proses seleksi data dilakukan berdasarkan peringkat atribut yang disusun berdasar nilai *information gain* dari setiap atribut seperti yang terdapat dalam Tabel 1.

Tabel 1. Peringkat Atribut

Peringkat	Atribut	*Information gain*
1	Riwayat Partus	0.064468
2	HIS (kontraksi)	0.061791
3	Kondisi Ketuban	0.04178
4	Umur	0.012442
5	Protein	0.009493
6	Posisi Bayi	0.006894
7	Panggul Sempit	0.004056
8	Hepatitis	0.003007
9	Ureum	0.003007
10	Usia Kandungan	0.002491
11	Tunggal/Ganda	0.002466
12	Asma	0.002462
13	Glukosa Urine	0.001496
14	Hemoglobin	0.001493
15	Tekanan Darah	0.001126
16	Limfosit (LYM)	0.001125
17	Ketuban Pecah Dini	0.001024
18	Trombosit	0.000620
19	Glukosa Sewaktu	0.000497
20	Eritrosit (RBC)	0.000250
21	SGPT	0.000000
22	SGOT	0.000000
23	Leukosit (WBC)	0.000000

Atribut SGPT, SGOT dan Leukosit memiliki *information gain* 0.00 sehingga keempat atribut tersebut dihapus karena tidak mempengaruhi penentuan jenis persalinan pada ibu hamil. Selanjutnya, dilakukan percobaan untuk menentukan jumlah atribut yang akan akan dilibatkan lebih lanjut dalam penelitian. Percobaan dilakukan dengan menerapkan algoritma *random forest* terhadap data asli yang berjumlah 302 records dan melakukan uji validasi memakai 3-fold cross validation untuk membagi data menjadi data training dan testing. Percobaan dilakukan sebanyak 20 kali, dimana dalam setiap percobaan dilakukan penambahan 1 atribut, urut berdasar peringkat dalam Tabel 1. Akurasi dari setiap percobaan tercantum dalam Tabel 2.

Tabel 2. Hasil Akurasi Percobaan Jumlah Atribut

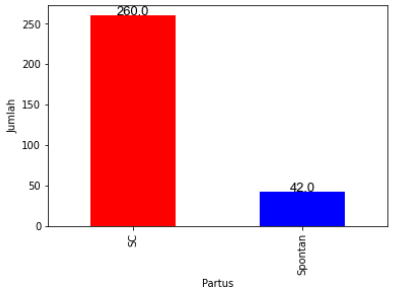
Jumlah atribut	Akurasi
1	81.09%
2	84.07%
3	84.07%
4	84.07%
5	84.07%
6	83.07%
7	83.07%
8	80.41%
9	80.07%
10	80.41%
11	80.41%
12	79.42%
13	80.07%
14	81.41%
15	82.07%
16	78.43%
17	82.74%
18	84.07%
19	83.06%
20	84.40%

Hasil Tabel 2 menunjukkan presentase tertinggi yaitu 84,40% yang diperoleh dalam percobaan ke-20 menggunakan 20 atribut yaitu Riwayat Partus, HIS (kontraksi), Kondisi Ketuban, Umur, Protein, Posisi Bayi, Panggul Sempit, Hepatitis, Ureum, Usia Kandungan, Tunggal/Ganda, Asma, Glukosa urine, Hemoglobin, Tekanan Darah, Limfosit (LYM), Ketuban Pecah Dini, Trombosit, Glukosa Sewaktu, Eritrosit (RBC). Dengan demikian selanjutnya dipergunakan 20 atribut tersebut dalam penelitian tahap berikutnya.

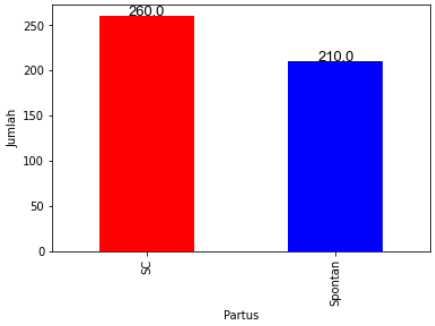
Dalam tahap ini juga ditetapkan atribut yang akan menjadi label kelas klasifikasi yaitu riwayat *partus* yang memiliki 2 kemungkinan nilai yaitu SC (*sectio caesaria*) dan Spontan. SC menandakan bahwa ibu hamil melahirkan dengan cara *caesar*, sedangkan spontan menandakan bahwa ibu hamil melahirkan secara alamiah.

D. Data Balancing

Terdapat ketidakseimbangan data yang dipergunakan dalam penelitian ini karena kelas persalinan spontan jauh lebih sedikit dibandingkan persalinan *caesar (SC).* Ketidakseimbangan kelas bisa menciptakan bias (kesalahan yang konsisten dalam memperkirakan sebuah nilai) sehingga model akan cenderung memprediksi kelas mayoritas. Oleh karena itu perlu dilakukan proses penyeimbangan data (*data balancing*) agar data minoritas seimbang dengan data mayoritas. Proses penyeimbangan data dilakukan dengan menerapkan algoritma *SMOTE. (Synthetic Minority Oversampling Technique)* [3]. Jumlah data sebelum dilakukan *balancing* berjumlah 302 *records*, setelah dilakukan proses *balancing*, jumlah *records* menjadi berjumlah 470 *records.* Gambar 2 dan gambar 3 menunjukkan distribusi data sebelum dan sesudah proses *balancing.*



Gambar 2. Distribusi Data Sebelum *Balancing*



Gambar 3. Distribusi Data Setelah *Balancing*

E. Pembagian Data Training dan Testing

Untuk membangun model dan mendapatkan akurasinya, diterapkan teknik *cross validation*. Dalam penelitian ini diujicobakan variasi jumlah fold 3, 5, dan 10.

F. Pemodelan Random Forest

Dalam membangun model *random forest* penulis menggunakan fungsi *RandomForestClassifier* dalam Python yang memiliki parameter *n_estimators, criterion* dan *random state*. Fungsi *n_estimators* digunakan untuk menunjukkan jumlah pohon dalam *random forest, criterion* digunakan mengukur kualitas *split,* dan *random_state* digunakan sebagai pembuat angka acak.

Dalam pengujian, penulis menggunakan variasi jumlah pohon 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096 untuk mendapatkan akurasi terbaik.

IV. PENGUJIAN

Tahap pegujian dilakukan menggunakan model 3-fold, 5-fold dan 10-fold. Penulis juga melakukan beberapa percobaan menggunakan variasi jumlah pohon 2, 4,8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. Pengujian dilakukan terhadap data sebelum dikenai proses *balancing* dan setelah dikenai proses *balancing* untuk mengetahui apakah proses *balancing* memberikan dampak pada peningkatan akurasi.

A. *Percobaan dengan 3-fold cross validation*

Tabel 3. Akurasi dengan 3-*fold Cross Validation*

No.	Jumlah Pohon	Akurasi Data Sebelum *Balancing*	Akurasi Data Setelah *Balancing*
1	2	80.79537%	84.88758%
2	4	85.75247%	88.72012%
3	8	87.73597%	90.21040%
4	16	80.79537%	91.27334%
5	32	88.39933%	92.12531%
6	64	89.06930%	92.55130%
7	128	88.07590%	91.48565%
8	256	87.41584%	91.69796%
9	512	88.40594%	91.69796%
10	1024	88.07590%	91.91300%
11	2048	88.07590%	91.69796%
12	4096	88.07590%	91.69796%

Berdasarkan hasil Tabel 3, pada percobaan dengan 3-fold cross validation, akurasi tertinggi diperoleh dalam percobaan ke-6 terhadap dataset yang telah dikenai proses balancing, dengan menggunakan 64 pohon, yaitu sebesar 92.55130%.

B. *Percobaan dengan 5-fold*

Tabel 4. Akurasi 5-*fold Cross Validation*

No.	Jumlah Pohon	Akurasi Data Sebelum *Balancing*	Akurasi Data Setelah *Balancing*
1	2	83.40437%	85.5319%
2	4	88.07103%	90%
3	8	88.07103%	91.27659%
4	16	89.73224%	92.12765%
5	32	89.40437%	91.70212%
6	64	89.07103%	91.91489%
7	128	88.74316%	92.12765%
8	256	89.40437%	91.91489%
9	512	89.06557%	92.12765%
10	1024	89.39890%	91.91489%
11	2048	88.07590%	91.91489%
12	4096	88.07590%	91.91489%

Berdasarkan hasil Tabel 4. akurasi tertinggi yaitu sebesar 92.12765%diperoleh pada percobaan ke-4 menggunakan 16 pohon terhadap dataset yang telah dikenai proses *balancing*.

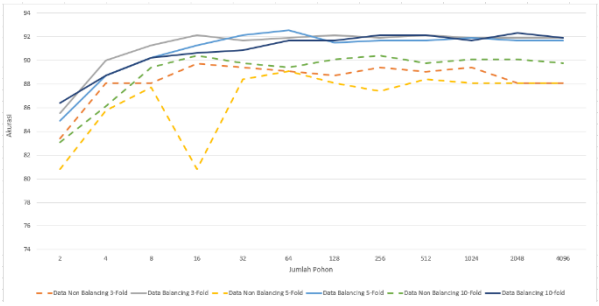
C. *Percobaan dengan 10-fold*

Tabel 5. Akurasi 10-*fold Cross Validation*

No.	Jumlah Pohon	Akurasi Data Sebelum *Balancing*	Akurasi Data Setelah *Balancing*
1	2	83.08602%	86.38297%
2	4	86.11827%	88.72340%
3	8	89.43010%	90.21276%
4	16	90.41935%	90.63829%
5	32	89.75268%	90.85106%
6	64	89.41935%	91.70212%
7	128	90.08602%	91.70212%
8	256	90.40860%	92.12765%
9	512	89.76344%	92.12765%
10	1024	90.08602%	91.70212%
11	2048	90.08602%	92.34042%
12	4096	89.75268%	91.91489%

Berdasarkan hasil Tabel 5, akurasi tertinggi pada 10-fold cross validation yaitu sebesar 92.34042% diperoleh dalam percobaan ke-11 menggunakan 2048 pohon terhadap dataset yang telah dikenai proses *balancing*.

Perbandingan hasil berbagai variasi eksperimen di atas dapat digambarkan dalam bentuk grafik seperti pada Gambar 3.



Gambar 4. Grafik Perbandingan Akurasi Percobaan

Berdasarkan hasil percobaan menggunakan 3 model *cross validation* terhadap dataset sebelum dan setelah dikenai proses *balancing*, diperoleh akurasi tertinggi sebesar 92.55130% dalam percobaan pada 3-*fold cross validation* dengan menggunakan 64 pohon. Dalam grafik terlihat bahwa akurasi dari klasifikasi terhadap data yang dikenai proses *balancing* selalu lebih tinggi dibandingkan akurasi dari data yang tidak dikenai proses *balancing*.

Selain itu, dari grafik terlihat bahwa peningkatan jumlah pohon dari 256 hingga 4096 tidak menghasilkan peningkatan akurasi yang signifikan. Pada eksperimen dengan 5-fold dan jumlah pohon 16, terdapat perbedaan akurasi yang sangat tajam. Penyebab terjadinya fenomena tersebut masih perlu diteliti lebih lanjut.

V. KESIMPULAN

Dari hasil penelitian ini dapat diambil kesimpulan bahwa:

1. Metode *random forest* dapat digunakan untuk melakukan klasifikasi terhadap jenis persalinan ibu hamil.
2. Berdasarkan hasil *preprocessing* dan pengujian data, disimpulkan bahwa terdapat 20 atribut yang paling berpengaruh dalam menentukan klasifikasi jenis persalinan pada ibu hamil dengan metode *random forest*.
3. Pada dataset yang diujicobakan, akurasi klasifikasi terhadap data yang telah dikenai proses *balancing* lebih baik daripada data yang tidak dikenai proses *balancing*.
4. Dari percobaan pengujian yang dilakukan dengan 3-*fold*, 5-*fold*, dan 10-*fold cross validation* didapatkan hasil akurasi terbaik sebesar 92.55130% dengan menggunakan model 3-*fold cross validation* dan jumlah pohon 64.

REFERENSI

[1] (2019) The WHO website [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality

[2] Pusdiknas, WHO, JHIPEGO, Buku III asuhan kebidanan pada ibu inpartu, Jakarta, 2001.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[4] F. Livingston, "Implementation of Breiman's Random Forest Machine Learning Algorithm," *ECE591Q Machine Learning Journal Paper*, Fall 2005.

[5] G. Louppe, "Understanding Random Forest," Ph.D dissertation: Faculty of Applied Sciences Department of Electrical Engineering & Computer Science, University of Liège, France, 2014.

[6] F. Gorunescu, Data Mining Concepts, Models and Techniques. Berlin: Springer, 2011.

[7] Płoński, P. How many trees in the Random Forest? https://mljar.com/blog/how-many-trees-in-random-forest/,2020