

# PERBANDINGAN BEBERAPA METODE KLASIFIKASI DALAM MEMPREDIKSI INTERAKSI FARMAKODINAMIK\*

Hasnita<sup>1</sup>, Farit Mochamad Afendi<sup>2‡</sup>, and Anwar Fitrianto<sup>3</sup>

<sup>1</sup>Department of Statistics, IPB University, Indonesia, hasnita\_handayani@apps.ipb.ac.id

<sup>1</sup>Department of Statistics, IPB University, Indonesia, fmafendi@apps.ipb.ac.id

<sup>1</sup>Department of Statistics, IPB University, Indonesia, anwarstat@gmail.com

‡corresponding author

**Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)**

**Vol 4 No 1 (2020), 11 - 21**

Copyright © 2020 Hasnita, Farit Mochamad Afendi, and Anwar Fitrianto. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## **Abstract**

One mechanism for Drug-Drug Interaction (DDI) is pharmacodynamic (PD) interactions. They are interactions by which the effects of a drug are changed by other drugs at the site of the receptor. The interactions can be predicted based on Side Effects Similarity (SES), Chemical Similarity (CS) and Target Protein Connectedness (TPC). This study aims to find the best classification technique by first applying the scaling process, variable interaction, discretization, and resampling technique. We used Random Forest, Support Vector Machines (SVM) and Binary Logistic Regression for the classification. Out of the three classification methods, we found the SVM classification method produces the highest Area Under Cover (AUC) value compared to the other, which is 67.91%.

**Keywords:** binary logistic regression, drug-drug interaction, random forest, support vector machines.

## **1. Pendahuluan**

Interaksi antar obat atau sering disebut *Drug-Drug Interaction* (DDI) merupakan penyebab yang signifikan dari reaksi obat yang merugikan dan juga dapat didefinisikan sebagai modifikasi efek suatu obat akibat obat lain yang diberikan secara bersamaan, sehingga keefektifan atau toksisitas satu obat atau lebih berubah (Fradgley, 2003).

---

\* Received Jun 2019; Accepted Oct 2019; Published online on Feb 2020

Interaksi obat perlu diperhatikan karena dapat mempengaruhi respon tubuh terhadap pengobatan (Cortes & Vapnik, 1995). Salah satu mekanisme interaksi obat adalah interaksi farmakodinamik. Interaksi farmakodinamik merupakan interaksi dimana efek suatu obat diubah oleh obat lain pada tempat aksi. Perubahan efek obat akibat interaksi obat dapat bersifat membahayakan pasien dengan berkurangnya khasiat obat (Fradgley, 2003). Dalam penelitian (Huang et al., 2013) interaksi farmakodinamik belum mempunyai kata kunci nama obat untuk menentukan kelas interaksi farmakodinamik. Oleh karena itu dibutuhkan suatu metode analisis yang baik untuk memprediksi interaksi farmakodinamik.

Beberapa penelitian sebelumnya telah membahas suatu nilai yang dapat digunakan dalam memprediksi interaksi obat diantaranya adalah penelitian yang dilakukan oleh Gottlieb et al. (2012) menggunakan pendekatan kesamaan basis dengan mengukur kesamaan informasi obat untuk memprediksi interaksi obat. Namun, pendekatan tersebut memiliki beberapa keterbatasan, seperti tidak dapat menangani obat-obatan baru. Penelitian serupa juga dilakukan oleh Huang et al. (2013) yang mengukur kekuatan koneksi jaringan antara target obat (S-score) dan menggunakan kemiripan efek samping untuk mengukur hubungan antara obat (P-score) dengan menerapkan pendekatan bayes untuk memprediksi interaksi obat. Winata et al. (2019) melakukan penelitian tentang peningkatan akurasi klasifikasi interaksi farmakodinamik obat berbasis seleksi pasangan obat takberinteraksi menggunakan algoritma DP-Clus.

Sejauh ini hanya terdapat beberapa penelitian yang mengkaji model prediksi yang hanya berdasarkan pada struktur obat tanpa memerlukan informasi latar belakang farmakologis atau biologis tambahan dari obat yang diprediksi.

Terdapat beberapa metode yang dapat digunakan untuk menghasilkan model yang akan digunakan dalam memprediksi interaksi farmakodinamik diantaranya adalah metode klasifikasi *random forest*, *Support Vector Machine* (SVM) dan regresi logistik biner. Metode klasifikasi *random forest* pertama kali dikenalkan oleh Breiman (2001). Menurut Breiman (2001), metode *random forest* merupakan metode klasifikasi yang membentuk kumpulan pohon regresi yang dapat menghasilkan *error* yang rendah dalam memprediksi suatu kejadian dan dapat mengatasi data latih dalam jumlah yang besar secara efisien. Dalam penelitian Lingga P et al. (2017) menunjukkan bahwa metode klasifikasi *random forest* memiliki tingkat akurasi yang lebih tinggi dalam memprediksi suatu kejadian dibandingkan dengan metode yang lain. Metode statistika lainnya yang dapat diterapkan untuk melakukan klasifikasi data adalah *Support Vector Machine* (SVM). SVM merupakan metode klasifikasi yang cara kerjanya menemukan hyperplane yang baik sebagai pemisah dua kelas yang berbeda (Cortes & Vapnik, 1995). Penelitian yang dilakukan oleh Rachman & Purnami (2012) yang menunjukkan bahwa metode klasifikasi SVM memiliki tingkat akurasi yang tinggi (98,11%) pada kasus klasifikasi tingkat keganasan kanker. Metode klasifikasi lainnya juga telah dibahas oleh Marino et al. (2014) yaitu metode klasifikasi regresi logistik biner yang merupakan metode klasifikasi yang dapat menangani kasus yang mempunyai peubah biner, dalam penelitian tersebut menunjukkan bahwa metode klasifikasi regresi biner memiliki ketepatan yang lebih akurat dalam mengklasifikasikan kejadian tertentu dibandingkan dengan metode analisis deskriminan linear. Penelitian serupa juga dilakukan oleh Fadhillah & Notobroto (2016) yang menunjukkan bahwa metode regresi

logistik biner baik digunakan dalam memprediksi penyakit struk (*Transient Ischemic Attack*).

Berdasarkan pertimbangan tersebut peneliti menggunakan beberapa algoritma untuk memprediksi farmakodinamik berdasarkan interaksi obat melalui interaksi antar protein pada data terbaru yaitu data obat tahun 2018. Peneliti menggunakan metode *semi-automatic text mining* untuk menentukan interaksi farmakodinamik (PD) dan menggunakan data *golden standar negative* (GSN) atau sering disebut sebagai pasangan obat yang tidak saling berinteraksi, serta mengukur jarak hubungan antar obat dengan melihat kemiripan efek samping obat (SES), mengukur kekuatan koneksi jaringan antar target obat (PTC) dan menghitung jarak kemiripan struktur obat menggunakan *fingerprint* klekota-roth (CS). Penelitian ini bertujuan untuk membandingkan beberapa metode pendekatan statistika untuk memprediksi interaksi farmakodinamik berdasarkan nilai PTC, SES dan CS. Metode pendekatan statistika yang digunakan adalah metode klasifikasi *random forest*, SVM dan regresi logistik biner.

## 2. Metodologi

### 2.1 Data

Data yang digunakan dalam penelitian ini adalah data obat yang diambil dari DrugBank (<https://www.drugbank.ca>) pada 9 juli 2018, namun hanya data obat yang tergolong dalam FDA-approved dan protein targetnya yang ada dalam tubuh manusia yang akan dianalisis lebih lanjut. Data bobot protein yang berinteraksi (*weight score*) dapat diperoleh dari HPRD (<http://www.hprd.org/>) yang digunakan dalam perhitungan nilai PTC. Data dari HPRD dapat juga digunakan untuk mengetahui pasangan obat mana saja yang saling berinteraksi dengan cara melakukan proses mapping dengan data dari DrugBank. Data efek samping obat dapat diperoleh dari MedDRA (<https://www.meddra.org/>) yang akan disambungkan dengan data dari DrugBank. Namun pada data dari MedDRA memerlukan ID-Chemical sebagai penghubung, maka untuk memperoleh ID-Chemical dari protein target data dari STITCH (<http://stitch.embl.de/>) akan disambungkan dengan data dari matador. Setelah ID-Chemical diketahui maka data MedDRA dapat dihubungkan pada data dari DrugBank. Pada perhitungan koefisien *tanimoto* data dibangkitkan melalui *fingerprint* Klekota Roth melalui aplikasi web ChemDes (<http://www.scbdd.com/chemdes/>). Aplikasi ChemDes digunakan untuk menerima rangkaian SMILES yang berkesesuaian dengan struktur kimia dari obat pada pembangkitan data *Fingerprint*. Sub-struktur kimia dapat diperoleh dari [http://www.scbdd.com/padel\\_desc/fpskleroht/](http://www.scbdd.com/padel_desc/fpskleroht/). Data yang akan dianalisis adalah hasil irisan semua data yaitu 50.000 pasangan obat. Teknik yang digunakan untuk mengumpulkan data ini menggunakan octoparse yang berfungsi untuk mengambil data dari berbagai website. Peubah yang digunakan dalam klasifikasi disajikan pada Tabel 1.

Tabel 1: Daftar peubah yang digunakan dalam klasifikasi.

| Peubah | Keterangan                                       | Tipe Peubah |
|--------|--|-------------|
| $y$    | Interaksi obat (PD/GSN)                          | Kategorik   |
| $x_1$  | Ukuran kemiripan efek samping obat (SES)         | Numerik     |
| $x_2$  | Keterhubungan protein target pasangan obat (PTC) | Numerik     |
| $x_3$  | Ukuran kemiripinan struktur obat (CS)            | Numerik     |

## 2.2 Metode Penelitian

Langkah-langkah Analisis yang dilakukan berdasarkan tujuan penelitian ini adalah sebagai berikut :

1. Menentukan peubah-peubah yang akan digunakan dalam prediksi interaksi obat farmakodinamik yaitu :

- a. Melakukan penyeleksian data yang akan digunakan pada penelitian ini.
- b. Membentuk peubah  $x_1$  dengan cara menghitung jarak terdekat dengan melihat kemiripan efek samping antar obat yang disimbolkan sebagai nilai SES. Data efek samping yang diperoleh dari DrugBank akan diubah kedalam bentuk biner (0 dan 1), kemudian menghitung nilai frekuensi efek samping masing-masing pasang obat dengan cara menjumlahkan efek samping yang sama. Jika suatu pasang obat tidak memiliki efek samping yang sama maka nilai SES akan bernilai nol. Nilai *weight* ( $g_i$ ) diperoleh dengan cara menghitung bobot pada efek samping yang berkorelasi. Perhitungan nilai SES dilakukan dengan memasukkan nilai frekuensi efek samping obat dan nilai *weight* kedalam persamaan (1). Perhitungan tersebut menggunakan bahasa pemrograman R.

$$SES = \sum_{i \in X \cap Y} (-\log f_i) \cdot g_i \quad (1)$$

dengan  $f_i$  merupakan frekuensi efek samping,  $X$  dan  $Y$  adalah dua obat yang berbeda dan  $g_i$  merupakan nilai bobot pada efek samping yang berkorelasi Campillos *et al.* (2008).

- c. Membentuk peubah  $x_2$  dengan cara menghitung jarak hubungan antara dua system target yang berpusat di jaringan yang disimbolkan sebagai PTC. Perhitungan nilai PTC dilakukan dengan cara memetakan obat kedalam jaringan PPI berdasarkan target protein dari obat, kemudian memboti protein yang tepat satu langkah pertama protein target. Jika dua system berpusat diberbagai gen maka ditambahkan satu diantara dua system tersebut. Perhitungan nilai PTC menggunakan persamaan (2) dengan bantuan bahasa pemrograman Pearl.

$$PTC = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (2)$$

dengan  $\bar{x}$  merupakan nilai rata-rata,  $s$  merupakan nilai standar deviasi,  $n$  merupakan jumlah data, dan  $\mu_0$  merupakan nilai rata-rata yang mewakili koefisien korelasi Pearson (PCC) pada seluruh data dalam jaringan (Huang *et al.*, 2013).

- d. Membentuk peubah  $x_3$  dengan cara menghitung jarak kemiripan obat menggunakan *fingerprint* klekota-roth yang disimbolkan sebagai  $CS$  menggunakan persamaan (3) dengan bantuan bahasa pemrograman SAS.

$$CS = \frac{a}{a+b+c} \quad (3)$$

dengan  $a$ ,  $b$  dan  $c$  merupakan frekuensi kemunculan dari kejadian ( $Obat_1 = 1$  dan  $Obat_2 = 1$ ), ( $Obat_1 = 0$  dan  $Obat_2 = 1$ ) dan ( $Obat_1 = 1$  dan  $Obat_2 = 0$ ) (Bajusz *et al.*, 2015).

- e. Menentukan peubah  $y$  (DDI) yang terdiri dari dua kelas data yaitu kelas interaksi farmakodinamik dan kelas GSN. Untuk menentukan interaksi farmakodinamik digunakan metode *semi-automatic text mining*, sedangkan untuk menentukan kelas GSN diambil secara acak kurang lebih 4 kali lebih banyak dari data PD pada semua data pasang obat.
2. Eksplorasi data pasangan obat pada tiga peubah prediktor (PTC, SES dan CS) dan data DDI sebagai peubah respon, untuk memberikan gambaran mengenai DDI.
  3. Pada proses klasifikasi interaksi obat, langkah pertama yang akan dilakukan adalah menginput data latih dan data validasi. Pada proses ini akan dilakukan pembagian data, yaitu data latih dan data validasi dengan menggunakan *5-fold cross validation*. Pada proses pemodelan klasifikasi menggunakan data latih pada beberapa pendekatan statistika yaitu *random forest*, SVM, dan regresi logistik biner. Pada data validasi untuk masing-masing pendekatan statistika akan dilakukan proses sebagai berikut:
    - a. Memprediksi interaksi obat dengan menggunakan model yang telah diperoleh.
    - b. Membentuk tabel ketetapan klasifikasi
    - c. Menghitung ketepatan prediksi.
  4. Melakukan *preprocessing* data diantaranya adalah melakukan proses *scaling*, membuat peubah baru (interaksi peubah) dengan cara mengalikan setiap kemungkinan kombinasi peubah, untuk mengatasi pencilan maka dilakukan metode diskritisasi pada data pasang obat dan dilakukan metode *undersampling* untuk menangani kelas data pasang obat yang tidak seimbang. Setelah melakukan *preprocessing* data maka proses 3 akan diulang kembali dengan menggunakan data yang telah di *preprocessing*.
  5. Membandingkan metode klasifikasi *random forest*, SVM, dan regresi logistik biner dari hasil klasifikasi interaksi obat tanpa *preprocessing* dan dengan dilakukan *preprocessing* berdasarkan nilai AUC.

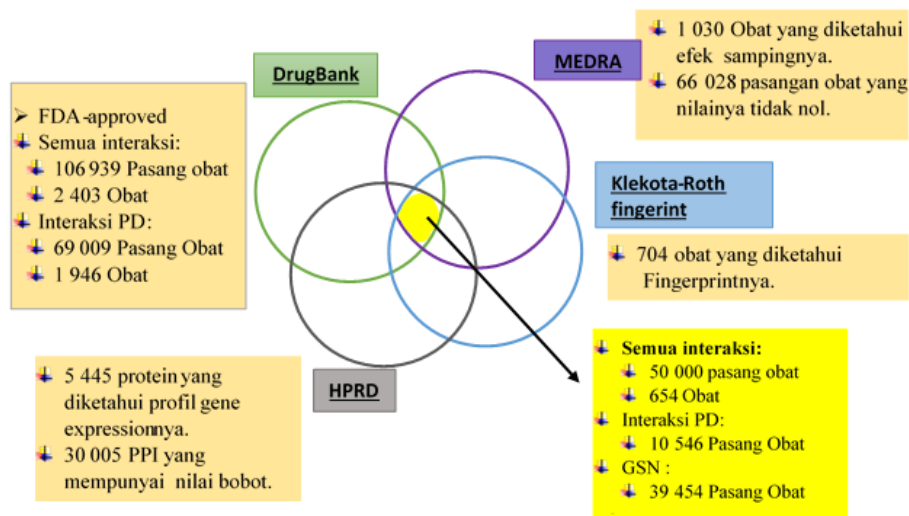
### 3. Hasil dan Pembahasan

#### 3.1 Penyeleksian Data

Data yang digunakan pada penelitian ini adalah data obat yang diambil dari berbagai sumber data. Berikut adalah ringkasan dari penyeleksian data.

Berdasarkan Gambar 1 terdapat beberapa sumber data yang menjadi acuan peneliti yang akan dikaitkan ke data drugbank. Data dari drugbank yang tergolong dalam FDA-approved terdiri dari 106 939 pasang obat dan terdiri dari 2 403 obat. Data drugbank yang saling berinteraksi farmakodinamik terdiri dari 69 009 pasang obat dan terdiri dari 1 946 obat. Pada perhitungan nilai kemiripan efek samping obat diperlukan data yang menyiapkan efek samping obat dan pada sumber data dari medra

menyiapkan data efek samping sebanyak 1 030 pasang obat dan terdapat 66 028 pasang obat yang mempunyai nilai PTC tidak nol yang juga terdapat pada drugbank. Data bobot protein yang berinteraksi (*weight score*) dapat diperoleh dari HPRD yang digunakan dalam perhitungan nilai kemiripan struktur jaringan protein target obat. Terdapat 30 005 nilai bobot protein yang diketahui pada data HPRD yang juga terdapat pada drugbank. Terdapat 704 obat dari Klekota-Roth yang juga terdapat pada drugbank dan data dari Klekota-Roth ini yang akan membantu dalam perhitungan nilai CS. Pengumpulan data terakhir dari berbagai sumber data yang semuanya terdapat juga pada drugbank adalah 50 000 pasang obat yang terdiri dari 10 546 pasang obat yang berinteraksi farmakodinamik dan 39 454 pasang obat yang termasuk GSN.



Gambar 1: Ringkasan pengumpulan data.

### 3.2 Pembentukan Peubah Penjelas

Penelitian ini menggunakan kemiripan efek samping obat pada pembentukan peubah penjelas ( $x_1$ ) dengan cara menghitung jarak efek samping antar pasang obat yang disimbolkan sebagai nilai SES. Hal pertama yang dilakukan dalam perhitungan ini adalah menentukan nilai weight dengan cara memboboti efek samping yang berkorelasi dengan bantuan bahasa pemrograman R diperoleh nilai minimum weight adalah 0.007, nilai mean adalah 1 dan nilai maksimum adalah 455.791. Selanjutnya menghitung nilai frekuensi efek samping obat dengan cara efek samping yang terdapat pada obat dibuat kedalam bentuk biner (0 dan 1). Kemudian, menjumlahkan efek samping yang sama pada suatu pasang obat dan membuatnya kedalam logaritma negative. Setelah nilai weight dan nilai logaritma negative dari efek samping obat didapatkan, maka nilai SES dapat dihitung dengan menggunakan persamaan (1). Sehingga dengan menggunakan bantuan bahasa pemrograman R didapatkan nilai minimum SES adalah nol, nilai rata-rata adalah 6.575 dan diperoleh nilai maksimumnya adalah 653.358. Semakin besar nilai SES menunjukkan bahwa satu pasang obat memiliki efek samping yang sangat mirip. Nilai SES pada suatu pasang

obat yang menghasilkan nilai nol menunjukkan bahwa data pasang obat tersebut memiliki efek samping yang berbeda.

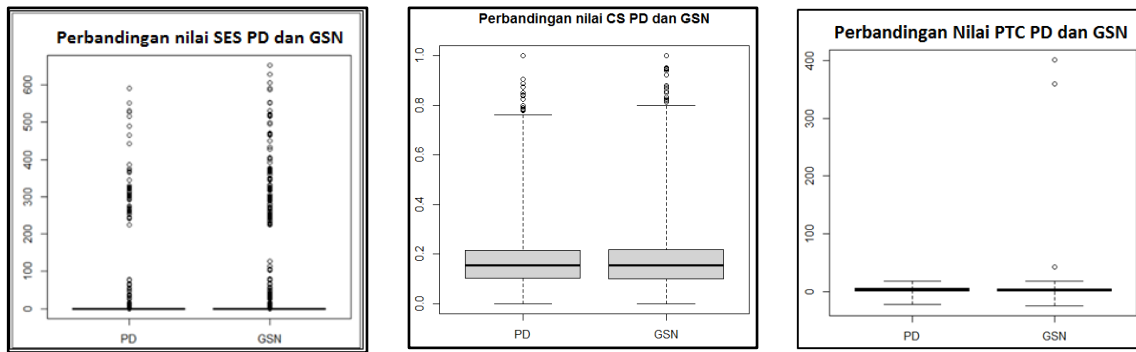
Pembentukan peubah penjelas ( $x_2$ ) menggunakan protein target obat dengan cara menghitung jarak hubungan antara dua system target yang berpusat di jaringan yang disimbolkan sebagai nilai PTC. Pertama kali yang harus dilakukan dalam perhitungan ini adalah menghitung nilai weight atau pembobotan pada setiap protein yang berpusat di jaringan PPI dimana protein yang akan diboboti adalah protein yang tepat satu langkah pertama dari protein target. Nilai bobot ini disediakan pada data dari HPRD sebanyak 30 005 protein yang mempunyai nilai bobot. Setelah mengetahui nilai bobot dari masing-masing protein maka nilai PTC dapat dihitung dengan cara nilai rata-rata dari bobot protein target akan dikurangkan dengan nilai rata-rata seluruh protein yang mempunyai nilai bobot dan hasil pengurangan tersebut akan dikalikan dengan akar dari banyaknya data protein yang mempunyai nilai bobot, selanjutnya akan dibagi dengan standar deviasi. Hal ini dilakukan untuk seluruh data pasang obat. Hasil dari perhitungan tersebut didapatkan nilai PTC minimum adalah -24.7030, nilai rata-rata adalah 2.3431, dan nilai maksimum adalah 400.9049.

Pembentukan peubah penjelas ( $x_3$ ) menggunakan *fingerprint* Klekota-roth dengan menghitung jarak kemiripan obat yang disimbolkan sebagai nilai CS. Perhitungan ini dilakukan dengan cara terlebih dahulu menghitung frekuensi kemunculan dari seluruh kemungkinan kombinasi suatu pasang obat kemudian perhitungan selanjutnya menggunakan koefisien Tanimoto yaitu frekuensi kemunculan dari suatu pasang obat yang keduanya terdapat fragmen pada data *fingerprint* akan dibagi dengan hasil penjumlahan frekuensi kemunculan dari suatu pasang obat yang keduanya mempunyai fragmen pada data *fingerprint* dan frekuensi kemunculan dari suatu pasang obat dimana salah satu obat tersebut tidak terdapat fragmen pada data *fingerprint*. Perhitungan tersebut akan dilakukan pada 50 000 data pasang obat. Sehingga dengan bantuan aplikasi R didapatkan nilai minimum CS adalah nol, nilai rata-rata adalah 0.1663 dan nilai maksimum adalah 1.

### 3.3 Deskripsi dan Korelasi Data Pasang Obat

Data yang digunakan adalah 50 000 data pasang obat yang terdiri dari kelas interaksi farmakodinamik (PD) dan kelas GSN. Perbandingan nilai peubah SES, CS dan PTC pada masing-masing kelas PD dan GSN dapat dilihat pada Gambar 2.

Berdasarkan Gambar 2 dapat disimpulkan bahwa perbandingan nilai SES pada interaksi PD memiliki nilai yang lebih kecil dibandingkan dengan nilai SES pada GSN namun tidak dapat dikatakan bahwa semakin kecil nilai SES, maka semakin besar peluang terjadi interaksi PD. Pada nilai CS memiliki nilai mean dan maksimal adalah sama baik untuk interaksi PD maupun GSN. Nilai PTC pada interaksi PD memiliki nilai mean berada disekitaran nol dan nilai PTC pada interaksi PD lebih kecil dibandingkan dengan nilai PTC pada GSN. Pada kelas GSN terdapat data yang sangat jauh dari nilai mean (ekstrim).



Gambar 2: Plot perbandingan nilai SES, CS dan PTC berdasarkan kelas PD dan GSN.

Uji rank Spearman digunakan untuk menguji hipotesis korelasi dengan skala pengukuran peubah minimal ordinal. Skala data untuk kedua peubah yang akan dikorelasikan dapat berasal dari skala yang berbeda. Korelasi antar peubah dapat dilihat pada Tabel 2.

Tabel 2: Uji korelasi pada masing-masing peubah.

|     | CS     | PTC    | SES    |
|-----|--------|--------|--------|
| CS  | 1.000  | -0.029 | 0.020  |
| PTC | -0.029 | 1.000  | -0.007 |
| SES | 0.020  | -0.007 | 1.000  |
| ddi | 0.011  | 0.109  | 0.033  |

Berdasarkan Tabel 2 dapat disimpulkan bahwa korelasi antara peubah SES dengan CS dan *ddi* adalah lemah.

### 3.4 Klasifikasi Tanpa *Preprocessing* Data

Metode klasifikasi yang digunakan pada penelitian ini adalah metode *random forest*, SVM, dan regresi logistik biner. Hasil dari metode ini akan dibandingkan dengan berbagai perlakuan pada data. Pada penelitian ini digunakan pembagian data menjadi data latih dan data uji dengan menggunakan metode *cross validation*. Metode *cross validation* digunakan untuk memperkecil bias yang terkait dengan pengambilan contoh secara acak dengan memilih *fold* sebanyak 5 sehingga data pasang obat dibagi menjadi 5 bagian. Data pasang obat dibagi menjadi 5 bagian dengan ukuran yang hampir sama sehingga data pasang obat akan pernah menjadi data latih dan data validasi.

Hal yang pertama kali dilakukan pada metode klasifikasi *random forest* adalah melakukan penarikan contoh acak pada tiga peubah bebas, lalu membangun pohon sebanyak 10 pohon dan melakukan proses looping untuk memperoleh pohon keputusan dengan bantuan aplikasi R sehingga diperoleh nilai akurasi, *sensitifity*, dan *specificity* pada data pasang obat yang dapat dilihat pada Tabel 3 Metode klasifikasi SVM dilakukan dengan cara terlebih dahulu menentukan nilai sigma dan nilai  $\sigma$  dengan melakukan proses tuning untuk menentukan parameter yang baik untuk data pasang obat pada metode klasifikasi SVM, dengan bantuan aplikasi R di dapatkan nilai sigma adalah 0.5 dan nilai  $\sigma$  adalah 0.5. Setelah nilai sigma dan  $\sigma$  diketahui maka



proses selanjutnya adalah melakukan klasifikasi SVM menggunakan kernel RBF, dengan bantuan bahasa pemrograman R didapatkan nilai akurasi, *sensitifity*, dan *specificity* dari model SVM yang dapat dilihat pada Tabel 3 Metode klasifikasi regresi logistik biner yang pertama kali dilakukan adalah mengestimasi parameter  $\beta$  dengan menggunakan metode *maximum likelihood* dan selanjutnya dilakukan klasifikasi menggunakan parameter  $\beta$  yang telah di estimasi dengan bantuan aplikasi R sehingga didapatkan nilai akurasi, *sensitifity*, dan *specificity* dari model Regresi logistik biner yang dapat dilihat pada Tabel 3.

Tabel 3: Perbandingan metode klasifikasi tanpa *preprocessing* data

| No | Metode                 | Akurasi(%) | <i>Sensitifity</i> (%) | <i>Specificity</i> (%) | AUC(%) |
|----|------------------------|------------|------------------------|------------------------|--------|
| 1  | <i>Random forest</i>   | 75.76      | 22.04                  | 90.12                  | 56.08  |
| 2  | SVM                    | 78.91      | 0.11                   | 99.70                  | 50.00  |
| 3  | Regresi logistik biner | 78.86      | 0.00                   | 100.00                 | 50.00  |

Berdasarkan Tabel 3 dapat disimpulkan bahwa tingkat akurasi yang dihasilkan dari metode klasifikasi *random forest* tanpa *preprocessing* data adalah 75.76% dengan nilai AUC sebesar 56.08%. Metode klasifikasi *random forest* dapat memprediksi interaksi farmakodinamik sebesar 22.04% dan dapat memprediksi GSN sebesar 90.12%. Pada metode klasifikasi SVM menghasilkan nilai akurasi sebesar 78.91% dan nilai AUC sebesar 50%, namun hanya dapat memprediksi interaksi farmakodinamik sebesar 0.10%. Dan untuk metode klasifikasi regresi logistik biner menghasilkan nilai akurasi 78.86%, namun tidak dapat memprediksi kelas interaksi farmakodinamik. Metode klasifikasi regresi logistik biner tanpa *preprocessing* data hanya dapat memprediksi kelas GSN. Berdasarkan nilai *sensitifity* dari ketiga metode tersebut dapat disimpulkan bahwa ketiga metode tidak dapat memprediksi interaksi farmakodinamik dengan baik oleh karena itu peneliti akan melakukan berbagai *preprocessing* data untuk meningkatkan nilai *sensitifity*.

### 3.5 Klasifikasi dengan *Preprocessing* Data

Data yang terdiri dari 50000 pasang obat dilakukan *preprocessing* data untuk meningkatkan nilai *sensitifity* pada setiap metode klasifikasi. Berdasarkan pembentukan peubah SES, CS dan PTC terlihat bahwa *range* pada setiap peubah sangat berbeda oleh karena itu peneliti melakukan proses standarisasi (*scaling*) agar *range* pada setiap peubah tidak jauh berbeda. Berdasarkan data yang digunakan yang terdiri dari kelas interaksi PD sebanyak 10546 pasang obat dan GSN sebanyak 39454 pasang obat. Hal ini menunjukkan bahwa data pasang obat memiliki kelas yang tidak seimbang oleh karena itu peneliti menggunakan teknik *resampling* dengan menggunakan metode *undersampling* untuk mengatasi kelas yang tidak seimbang. Kemudian peneliti melakukan interaksi antar peubah bebas dengan cara mengalikan disetiap kombinasi peubah yang terjadi sehingga peubah bebas dari data pasang obat yang akan dilakukan dalam klasifikasi adalah 10 peubah bebas yaitu nilai SES, CS, PTC, interaksi antara SES-CS, SES-PTC, CS-PTC, SES-SES, CS-CS, PTC-PTC dan SES-CS-PTC. Berdasarkan eksplorasi yang telah dilakukan terlihat bahwa terdapat

data pencilan oleh karena itu peneliti melakukan proses diskritisasi untuk menangani data pencilan. Proses tersebut dilakukan dengan menggunakan metode diskritisasi yang berbasis *entropy*. Metode diskritisasi akan diterapkan pada data pasang obat dengan 10 peubah bebas.

Perbandingan metode klasifikasi berdasarkan nilai akurasi, *sensitifity*, *specificity* dan nilai AUC dari ketiga model klasifikasi dengan melakukan proses *scaling*, teknik *resampling*, interaksi peubah dan melakukan proses diskritisasi adalah disajikan pada Tabel 4.

Tabel 4: Perbandingan metode klasifikasi dengan *preprocessing* data.

| No | Metode                 | Akurasi(%) | <i>Sensitifity</i> (%) | <i>Specificity</i> (%) | AUC(%) |
|----|------------------------|------------|------------------------|------------------------|--------|
| 1  | <i>Random forest</i>   | 62.78      | 50.60                  | 73.01                  | 64.31  |
| 2  | SVM                    | 62.24      | 57.75                  | 78.07                  | 67.91  |
| 3  | Regresi logistik biner | 63.15      | 55.04                  | 72.34                  | 63.69  |

Berdasarkan Tabel 4 dapat disimpulkan bahwa tingkat akurasi yang dihasilkan dari metode klasifikasi *random forest* dengan melakukan *preprocessing* data adalah 62.78% dengan nilai AUC sebesar 64.31%. Metode klasifikasi *random forest* dapat memprediksi interaksi farmakodinamik sebesar 55.6% dan dapat memprediksi GSN sebesar 73.01%. Pada metode klasifikasi SVM menghasilkan nilai akurasi sebesar 62.24%, nilai AUC sebesar 67.91%, dan dapat memprediksi interaksi farmakodinamik sebesar 57.75%. Metode klasifikasi regresi logistik biner menghasilkan nilai akurasi 63.15%, dan dapat memprediksi kelas interaksi farmakodinamik dan GSN berturut-turut sebesar 55.04% dan 72.34%.

Berdasarkan Tabel 3 dan Tabel 4 dapat disimpulkan bahwa ketika tidak dilakukan *preprocessing* maka metode klasifikasi yang baik digunakan untuk data pasang obat adalah metode klasifikasi *random forest*. Metode klasifikasi *random forest* tanpa *preprocessing* data memiliki nilai AUC yang paling tinggi (56.08%) dibandingkan dengan metode klasifikasi SVM dan regresi logistik biner. Namun ketika melakukan *preprocessing* yaitu melakukan proses *scaling*, teknik *resampling*, interaksi antar peubah, dan melakukan diskritisasi peubah maka dapat meningkatkan nilai *sensitifity* yang artinya bahwa dengan dilakukan *preprocessing* maka dapat meningkatkan kemampuan dalam memprediksi interaksi farmakodinamik. Berdasarkan nilai AUC ketiga metode tersebut tidak jauh berbeda namun metode klasifikasi SVM yang memiliki nilai AUC tertinggi (67.91%) dibandingkan dengan metode klasifikasi *random forest* dan regresi logistik biner.

#### 4. Simpulan

Pada penelitian ini dapat disimpulkan bahwa metode klasifikasi yang baik digunakan dalam memprediksi interaksi farmakodinamik adalah metode SVM dengan syarat harus dilakukan *preprocessing* data karena dengan dilakukan *preprocessing* data yaitu melakukan proses *scaling*, teknik *resampling*, interaksi antar peubah, dan dilakukan diskritisasi peubah pada data pasang obat maka dapat meningkatkan kemampuan suatu metode klasifikasi dalam memprediksi interaksi farmakodinamik.

## Daftar Pustaka

- Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1): 20.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5–32.
- Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321(5886): 263–266.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3): 273–297.
- Fadhilah, A. R., & Notobroto, H. B. (2016). Analisis regresi logistik biner pada kejadian transient ischemic attack (TIA) di RSUD Dr. Soetomo Surabaya. *Jurnal Biometrika Dan Kependudukan*, 5(2): 157–165.
- Fradgley, S. (2003). *Interaksi Obat dalam Farmasi Klinis (Clinical Pharmacy) Menuju Pengobatan Rasional dan Penghargaan Pilihan Pasien (Aslam M, Tan CK, Prayitno A, Ed)*. Jakarta (ID): PT Elex Media Komputindo Kelompok Gramedia.
- Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E., & Sharan, R. (2012). INDI: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology*, 8(1).
- Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H., & Han, J.-D. J. (2013). Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. *PLoS Computational Biology*, 9(3): e1002998. <https://doi.org/10.1371/journal.pcbi.1002998>
- Lingga P, R. D., Fatichah, C., & Purwitasari, D. (2017). Deteksi gempa berdasarkan data twitter menggunakan decision tree, random forest, dan SVM. *Jurnal Teknik ITS*, 6(1): A159–A162.
- Marino, M., Tirta, I. M., & Dewi, Y. S. (2014). Perbandingan analisis diskriminan linier, diskriminan linier robust dan regresi logistik biner. Studi kasus pada penjurusan bidang IPA / IPS siswa tingkat SMA Negeri 1 Bangorejo Banyuwangi. *Prosiding Seminar Nasional Matematika*, 192–200. Jember (ID): Universitas Jember.
- Rachman, F., & Purnami, S. W. (2012). Perbandingan klasifikasi tingkat keganasan breast cancer dengan menggunakan regresi logistik ordinal dan support vector machine (SVM). *Jurnal Sains Dan Seni ITS*, 1(1): D130–D135.
- Winata, H. M., Afendi, F. M., & Fitrianto, A. (2019). Peningkatan akurasi klasifikasi interaksi farmakodinamik obat berbasis seleksi pasangan obat takberinteraksi. *Indonesian Journal of Statistics and Its Applications*, 3(3): 247–259.