KOLONI: Jurnal Multidisiplin Ilmu, 2 (1), Tahun 2023
e-ISSN: 2828-6863

PERBANDINGAN METODE KLASIFIKASI KEGAGALAN SIMULASI MODEL IKLIM

Caecilia Bintang Girik Allo[1], Nicea Roona Paranoan[2], Winda Ade Fitriya B[3], Sitti Rosnafi'an Sumardi[4]
[1]Universitas Cenderawasih, Jayapura, Indonesia
[2,3,4]Universitas Cenderawasih, Jayapura, Indonesia
caecilia.bintang@fmipa.uncen.ac.id

Abstract: *Simulation of climate model is used to produce climate models used to estimate climate in the future using some software. Simulation of climate model has two probability, they are success or failure. The problem is when the simulation is fail. There are 18 variables that used to predict the simulation. Feature selection is used to reduce the dimension of variables using RFECV method. There are 11 variables that important to simulation of climate. There are 46 from 540 simulations that fail. Furthermore, SMOTE is used to handle imbalance cases. The classification method used in this paper are Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and Random Forest. The AUC value were not significantly different for the four methods when using SMOTE. However, the highest AUC was obtained by SVM method, so the simulation of climate model can be predicted by SVM method.*

Keywords: *AUC, SMOTE, RFECV, Logistic Regression, SVM, Random Forest, Naïve Bayes*

Abstrak: Simulasi model iklim digunakan untuk menghasilkan model iklim yang dapat berguna untuk memproyeksikan iklim di masa depan dengan menggunakan suatu software tertentu. Simulasi model iklim yang dilakukan mempunyai dua kemungkinan yaitu berhasil atau gagal. Permasalahan yang terjadi adalah adanya kegagalan simulasi model iklim. Simulasi model iklim dilakukan berdasarkan nilai-nilai parameter. Terdapat 18 parameter yang digunakan untuk membuat simulasi. Metode klasifikasi dapat digunakan untuk mengetahui suatu simulasi berhasil atau gagal. Dilakukan seleksi variabel dengan metode RFECV untuk mengetahui variabel yang penting terhadap simulasi model iklim. Hasil seleksi variabel diperoleh 11 variabel. Terdapat 46 dari 540 simulasi model iklim yang gagal. Selanjutnya dilakukan proses SMOTE untuk menyeimbangkan data simulasi model iklim yang tidak seimbang. Metode klasifikasi yang digunakan dalam paper ini adalah Regresi Logistik, Naïve Bayes, SVM, dan Random Forest. Hasil yang diperoleh adalah nilai AUC tidak begitu berbeda secara signifikan untuk keempat metode tersebut. Namun, nilai AUC tertinggi diperoleh dengan metode SVM, sehingga model kegagalan simulasi model iklim diprediksi dengan metode SVM.

Kata kunci: AUC, SMOTE, RFECV, Regresi Logistik, SVM, Random Forest, Naïve Bayes

Pendahuluan

Model iklim menggunakan metode kuantitatif untuk mensimulasikan interaksi pergerakan penting iklim, termasuk atmosfer, lautan, permukaan tanah, dan es. Contoh kegunaan model iklim adalah untuk memprediksi curah hujan di suatu daerah dan mengukur panas suhu permukaan laut. Dibutuhkan simulasi model iklim agar dapat menghasilkan model iklim yang dapat berguna untuk memproyeksikan iklim di masa depan. Terdapat beberapa parameter yang dibutuhkan untuk membuat model iklim. Oleh karena itu, parameter-parameter tersebut yang akan digunakan untuk simulasi model iklim. Simulasi model iklim menggunakan software tertentu. Penting untuk mengetahui apakah simulasi model iklim yang dibangun berdasarkan parameter-parameter model iklim dapat berhasil atau gagal. Kegagalan dalam simulasi adalah ketidakstabilan numerik atau fenomena iklim tertentu yang dapat memberikan pengaruh positif yang sangat besar terhadap model (Lucas *et al*, 2013).

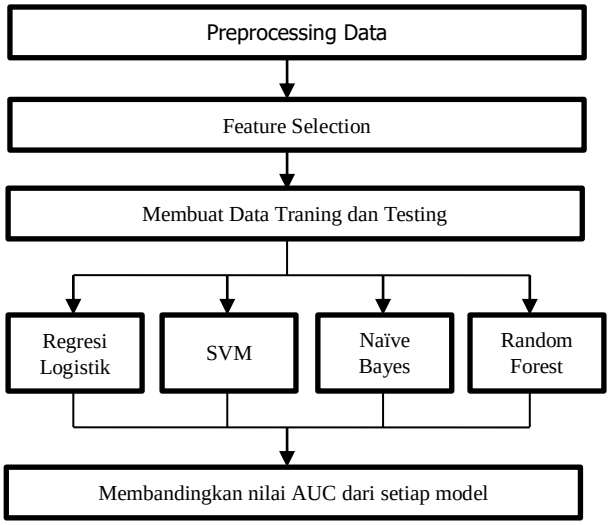
Bryne *et al* (2011) menggunakan Supprot Vector Machine untuk mendeteksi kehadiran atau keaadaan tumor pada kanker payudara. Khandezamin, Naderan, & Rasthi (2020) menggunakan Regresi Logistik untuk melakukan *feature selection*. Jacob *et al* (2012) mengaplikasikan metode klasifikasi menggunakan Naïve Bayes dan K-Nearest Neighbor pada datanya. Kegagalan simulasi model iklim dapat diketahui menggunakan metodde klasifiaksi. Lucas *et al* (2013) melakukan penelitian mengenai klasifikasi kegagalan simulasi model iklim. Namun pada penelitian tersebut tidak mempertimbangkan kasus tidak seimbangnya data dan tidak dilakukannya seleksi variabel. Metode klasifikasi yang digunakan hanyalah *Support Vector Machine* (SVM). Berdasarkan penelitian tersebut, peneliti coba mengembangkan penelitian tersebut dengan mempertingkan aspek ketidakseimbangan data dan metode klasifikasi lainnya.

Terdapat berbagai macam metode klasifikasi. Salah satunya adalah regresi logistik. Penggunaan metode klasifikasi berupa Regresi Logistik membutuhkan asumsi, yaitu tidak terjadi multikolinearitas antar variabel prediktor. Sedangkan metode klasifikasi lainnya seperti Support Vector Machine (SVM), Naïve Bayes, dan Random Forest tidak membutuhkan asumsi yang harus dipenuhi. Tujuan penelitian ini adalah memperoleh metode klasifikasi yang terbaik dari empat metode klasifikasi yang digunakan untuk kasus simulasi model iklim. Terdapat 540 simulasi model iklim. Setiap simulasi mempunyai 18 parameter. Penentuan model terbaik dilihat dari nilai AUC setiap model. Hasil dari penelitian ini dapat membantu pemerintah, agar simulasi model iklim yang digunakan tidak gagal. Apabila simulasi model iklim berhasil maka model iklim tersebut dapat digunakan untuk memproyeksikan perubahan iklim di masa depan.

Metode

Data yang digunakan diperoleh dari UCI Machine Learning Classification Dataset. Data dibuat di bawah naungan Departemen Energi Amerika Serikat oleh Lawrence Livermore National Laboratory (LLNL). Data didapat dari simulasi model iklim selama 10 tahun yang selanjutnya dicatat 18 nilai parameter yang membuat simulasi model iklim tersebut gagal atau berhasil (Lucas *et al*, 2013). Terdapat 540 simulasi, 46 diantaranya menghasilkan simulasi model iklim yang gagal dibeberapa waktu selama masa studi. Simulasi yang gagal diberi kode "0" dan Simulasi yang berhasi diberi kode "1". Nilai parameter model yang digunakan merupakan nilai parameter yang sudah distandarisasi menggunakan nilai minimum dan maksimum sehinggal nilai parameternya berada diantara [0,1].

Langkah-langkah pada penilitian ini dapat dilihat pada Gambar 1. Proses mengolah data dimulai dengan preprocessing data, *feature selection*, *modelling*, dan membandingkan model. Proses membagi data menjadi testing dan test menggunakan K-Fold Cross Validation.



Gambar 1. Langkah Analisis

A. Naïve Bayes

Klasifikasi naïve bayes adalah klasifikasi yang berdasarkan pada teorema bayes dan digunakan untuk menghitung probabilitas masing-masing kelas dengan asumsi bahwa tidak ada relasi antara kelas independen. Misalkan $\{x_1, x_2, ..., x_n\}$ merupakan atribut yang digunakan untuk mendefinisikan kelas. Perhitungan probabilitas prosterior untuk setiap kelas pasa satu atribut adalah (Gorunescu, 2011)

$$P(x_i \mid y_I) = \frac{1}{\sigma_{iI}\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_{iI})^2}{2\sigma_{iI}^2}\right) \quad (1)$$

Kelas yang terpilih adalah kelas yang memaksimumkan nilai dari $P(x_1, x_2, ..., x_p \mid y_I)$. Setiap atribut diasumsikan saling bebas dari kelas Y. Di mana $P(x_i \mid y_I)$, $i = 1,2,...,p$ dapat dihitung untuk setiap x_i. Data akan diklasifikasi ke dalam kelas y_I jika probabilitas yang dihasilkan bernilai paling tinggi di antara yang lain.

B. Regresi Logistik

Regresi salah satu metode yang digunakan untuk menggambarkan hubungan antara variabel respon dengan variabel prediktor. Apabila tipe dari variabel respon adalah kategori maka regresi yang digunakan adalah regresi logistik. Regresi logistik dapat digunakan sebgai salah satu metode klasifikasi.

Terdapat tiga jenis regresi logistik (Agresti, 2007). Pertama, regresi logistik biner. Regresi logistik biner digunakan ketika variabel respon memiliki dua kategori. Kedua, regresi logistik multinomial yang digunakan ketika variabel respon memiliki lebih dari dua kategori. Ketiga, regresi logistik ordinal. Regresi logistik ordinal digunakan ketika tipe data dari variabel respon adalah ordinal. Model regresi logistik sebagai berikut:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)} \quad (2)$$

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (3)$$

C. Support Vector Machine (SVM)

Selain regresi logistik, SVM juga merupakan salah satu metode klasifikasi. SVM digunakan untuk data linear dan nonlinear. Pemetaan nonlinear digunakan untuk mentransform data ke dimensi yang lebih tinggi. SVM menemukan pemisah yang linear atau *hyper plane* pada dimensi yang lebih tinggi (Han *et al*, 2012). SVM mencari maksimal jarak *hyper plane* (MMH) yang memenuhi persamaan (4) dengan syarat (5).

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{l} \xi \quad (4)$$

$$y_i\left(w\phi\left(x_i\right) + b\right) + \xi_i \geq 1$$
$$\xi_i > 0 \;; i = 1,\ldots,l \quad (5)$$

D. Random Forest

Metode *random forest* adalah pengembangan dari metode CART (Classification and Regression Tree), yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection* (Breiman, 2001). Dalam *random forest*, banyak pohon ditumbuhkan sehingga terbentuk hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas *n* amatan dan *p* peubah penjelas, *random forest* dilakukan dengan cara (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.

E. Synthetic Minority Oversampling Technique (SMOTE)

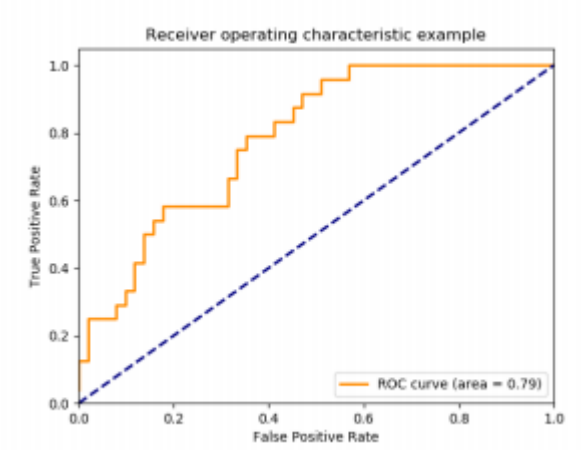
SMOTE merupakan salah satu metode untuk mengatasi kasus *imbalance,* yaitu jika terjadi proporsi jumlah tiap kelas sangat berbeda jauh. SMOTE yang diusulkan oleh Chawla *et al* (2002) melakukan penambahan jumlah data kelas minor agar setara dengan jumlah data pada kelas mayor dengan membangkitkan data buatan (sintesis) yang dibentuk berdasarkan *k-nearest*

neighbors. Prosedur pembangkitan data buatan untuk data numerik:

1. Hitung perbedaan antar vektor utama dengan ke tetangga terdekatnya.
2. Kalikan perbedaan dengan angka yang diacak di antara 0 dan 1.
3. Tambahkan perbedaan tersebut ke dalam nilai utama pada vektor utama asal sehingga diperoleh vektor utama baru.

F. Area Under Receiver Operating Characteristics Curve (AUC)

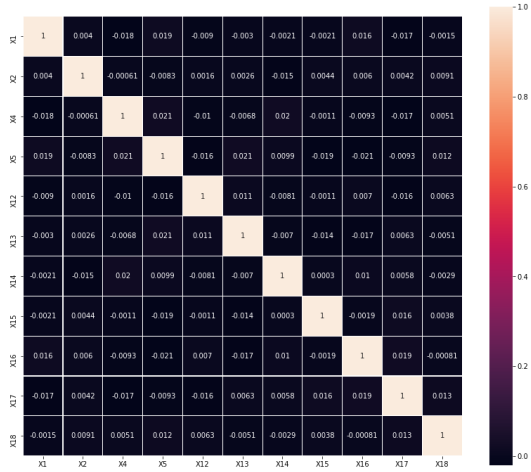
Area Under Receiver Operating Characteristic Curve (AUC) didefinisikan sebagai luas area trapezoid pada kurva ROC. Kurva ROC sendiri merupakan kurva yang dibentuk dari pasangan titik-titik True Positive Rate (TPR) terhadap False Positive Rate (FPR) dengan nilai threshold yang berbeda-beda.



Gambar 2. Contoh Kurva ROC dengan Nilai AUC

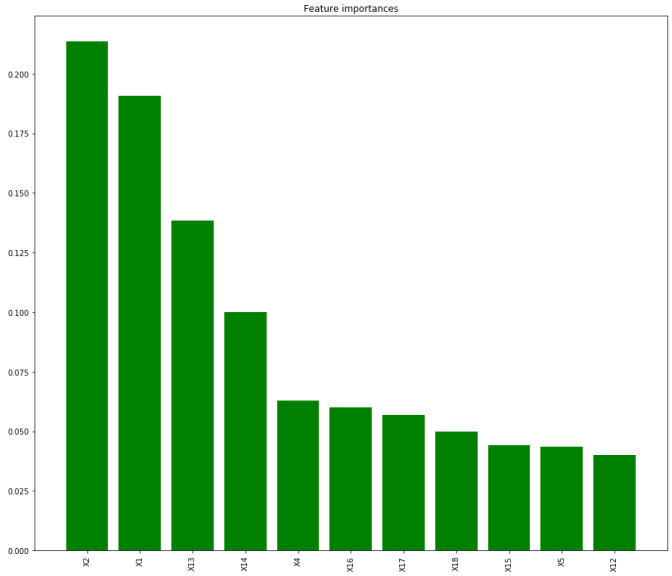
Hasil dan Pembahasan

Sebelum membuat model klasifikasi, tahap pertama yang paling penting adalah *preprocessing* data. Pada tahap ini pertama-tama dilakukan pengecekan adanya *missingvalue*. Dalam data yang digunakan dalam paper tidak terdapat *missingvalue.* Kemudian dilakukan pengecekan outlier terhadap data. Pengecekan outlier menggunakan boxplot. Dari hasil boxplot tiap variabel dapat dilihat bahwa tidak terdapat outlier pada data. Selanjutnya tidak dilakukan standarisasi pada data karena setiap variabel prediktor berada pada interval [0,1]. Pengecekan multikolinearitas dapat dilihat pada Gambar 3. Berdasarkan Gambar 3 dapat disimpulkan bahwa tidak terjadi multikolinearitas antar variabel.



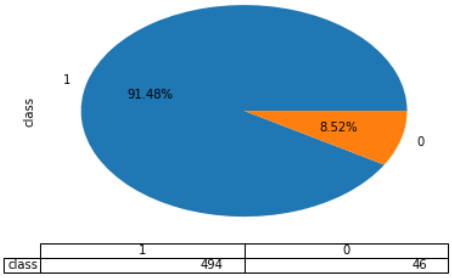
Gambar 3. Heatmap Korelasi

Setelah melakukan *preprocessing,* dilakukan seleksi variabel. Berdasarkan hasil dari seleksi variabel menggunakan metode RFECV, diperoleh 11 variabel prediktor yang berpengaruh terhadap variabel respon. Berdasarkan Gambar 4 dapat dikatakan bahwa terdapat empat variabel prediktor yang paling mempengaruhi model, yaitu X_2 , X_1 , X_{13} , dan X_{14} . Variabel-varaibel tersebut adalah viskositas parameter kedua, korelasi viskositas parameter pertama dan keenam, korelasi antara *tracer* dan momentum, dan difusivitas dasar yang vertikal mempengaruhi kegagalan simulasi model iklim.



Gambar 4. Hasil Ranking Fitur Berdasarkan Metode RFECV

Berdasarkan **Error! Reference source not found.**, diketahui bahwa data tidak seimbang (*imbalance*) antara kelas "0" dan kelas "1". Berdasarkan hal itu, maka dilakukan proses SMOTE untuk mengatasi data yang tidak seimbang, sehingga data menjadi seimbang.



Gambar 5. Perbandingan Jumlah Masing-Masing Kelas

Setelah data seimbang, dilakukan pemodelan. Pada penelitian ini juga dilakukan perbandingan untuk data yang seimbang dan tidak seimbang. Nilai AUC dari empat metode, yaitu Regresi Logistik, SVM, Random Forest, dan Naïve Bayes pada data Imbalanced yang diatasi dengan SMOTE telah dirangkum pada Tabel 2. Sedangkan nilai AUC dari empat metode, yaitu Regresi Logistik, SVM, Random Forest, dan Naïve Bayes pada data Imbalanced dapat dilihat pada Tabel 1.

Tabel 1. Hasil AUC Empat Metode Klasifikasi pada Data Imbalanced

Fold	AUC Random Forest	AUC SVM	AUC Regresi Logistik	AUC Naïve Bayes
1	0,9282828	0,9873737	0,75	0,70
2	0,9320988	0,9882155	0,7727273	0,6666667
3	0,8843537	0,9705215	0,8282313	0,7171202
4	0,9051627	0,8647587	0,7777778	0,7222222
5	0,9702581	0,9769921	0,9292929	0,8333333
Rata-rata	0,9240312	0,9575723	0,8116059	0,7278685

Tabel 2. Hasil AUC Empat Metode Klasifikasi Menggunakan SMOTE

Fold	AUC Random Forest	AUC SVM	AUC Regresi Logistik	AUC Naïve Bayes
1	0,963131	0,991919	0,991919	0,982828
2	0,93266	0,974186	0,960718	0,971942
3	0,939955	0,986532	0,989899	0,973064
4	0,86532	0,861953	0,843996	0,845679
5	0,984694	0,993197	0,988662	0,988662
Rata-rata	0,937152	0,961557	0,955039	0,952435

Apabila Tabel 1 dan Tabel 2 dibandingkan maka diperoleh bahwa metode terbaik untuk data yang mengalami imbalanced dan data yang sudah dilakukan SMOTE adalah metode SVM dengan kernel Linear. Nilai rata-rata AUC yang dihasilkan pada data yang mengalamai imbalaced relatif tinggi. Berdasarkan Tabel 2 dapat dilihat bahwa hasil AUC tidak berbeda secara signifikan untuk setiap metode. Nilai AUC setiap model berada di atas 0.90. Metode yang menghasilkan rata-rata nilai AUC terendah adalah metode Random Forest. Sedangkan metode yang menghasilkan rata-rata nilai AUC tertinggi adalah metode SVM. Sehingga model yang akan digunakan untuk memprediksi kelas model simulasi iklim adalah model SVM.

Kesimpulan

Data simulasi model iklim yang digunakan tidak seimbang antara kelas "0" dan kelas "1", sehingga digunakan SMOTE untuk menyeimbangkan kelas. Data Imbalanced pelru diatasi karena dapat meningkatkan nilai AUC. Variabel prediktor yang didapatkan dari seleksi variabel sebanyak 11 variabel dengan empat variabel tertingginya adalah viskositas parameter kedua, korelasi viskositas parameter pertama dan keenam, korelasi antara *tracer* dan momentum, dan difusivitas dasar yang vertikal. Kemudian, model klasifikasi terbaik adalah dengan menggunakan model SVM. Pemerintah dapat memperhatikan 11 variabel dalam membuat simualsi model iklim. Selanjutnya model klasifikasi SVM yang telah dibuat dapat digunakan untuk mengetahui apakah simulasi tersebut berhasil atau gagal.

Ucapan Terima Kasih

Ucapan terima kasih kepada seluruh pihak yang telah membantu segala proses mulai dari proses perancangan penelitian, pengolahan data, penulisan, dan terbit.

Referensi

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New Jersey: John Wiley & Sons, Inc.

Breiman, L. (2001). Random Forest. *Machine Learning, 45, 5-32.*

Breiman, L., & Cutler, A. (2003). Manual on Setting Up, Using, and Understanding Random Forest V4.0, *33*. Available: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.

Byrne, et al (2011). Support Vector Machine-Based Ultrawideband Breast Cancer Detection System. *Journal of Electromagnetics Waves and Applications, 25(13), 1807-1816*.

Chawla, et al. (2002). SMOTE: Synthetic Minority Over Sampling Technique. *In Journal of Artificial Intelligence Research, 16, 321-357*.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques*. Australia: Springer-Verlag Berlin Heidelberg.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. United States of America: Elsevier Inc.

Jacob *et al*. (2012). Efficient Classifier for Classification of Prognostic Breast Cancer Data Through Data Mining Techniques. *Proceedings of the World Congress on Engineering and Computer Science*. San Fransisco.

Khandezamin, Ziba., Naderan, Marjan., & Rasthi, M. J. (2020). Detection and Classification of Breast Cancer Using Logistic Regression Feature Selection and GMDH Calssifier. *Journal of Biomedical Informatics*.

Lucas, D.D. et al. (2013). Failure Analysis of Parameter-Induced Simulation Crashes in Climate Models. *Geoscientific Model Development, 6, 1157-1171*.