Analisis Efektivitas Teknik Resampling dalam Meningkatkan Kinerja Model Prediksi Kanker Payudara pada Dataset Wisconsin Breast Cancer

Nadinda Aurora
121450004

Meinisa
121450076

Saiful Haris Muhammad
121450115

Dede Masita
121450007

Helma Lia Putri
121450100

***Abstract*—Metode resampling merupakan teknik untuk mengukur model dengan memprediksi model. Breast Cancer dataset mengenai penyakit kanker payudara. *Cross Validation* adalah metode partisi data yang digunakan untuk mengetahui kestabilan penduga parameter sesuai model. Metode Holdout bergantung pada pembagian training set dan test set yang dilakukan secara acak. *Receiver Operating Characteristic* analisis dengan memanfaatkan *Confusion Matrix-based Measures. Oversampling melakukan standarisasi menggunakan z-score dan SMOTE (Syntehentic Minority Over-sampling Technique)* metode oversampling untuk menangani masalah ketidakseimbangan kelas dalam dataset.**

Keywords—Resampling, Cross Validation, Holdout, Receiver Operating Characteristic, SMOTE.

I. Pendahuluan

Kanker payudara *(Breast Cancer)* merupakan penyakit yang termasuk ke dalam penyakit yang cukup berbahaya, terlebih pada kaum wanita, walaupun tidak menutup kemungkinan terjadi pada kaum pria. Penyakit ini merupakan penyakit kedua terbesar dalam penyumbang angka kematian setelah kanker paru - paru. Bukan tanpa sebab, penyakit ini dapat mudah menghinggapi manusia dikarenakan penyakit ini menyerang tubuh manusia yang banyak komponen kemungkinan penyebabnya. Deteksi dini dalam memprediksi kanker payudara sangat diperlukan dengan tujuan meningkatkan pemahaman masyarakat dan mampu mengurangi resiko yang ditimbulkan serta mengobati pada tahap penanganan pertama. Contohnya seperti diagnosis yang ada, tingkat keparahan atau kedalaman dari bagian yang cekung permukaan kanker pada payudara, hingga perbedaan tingkat warna keabuan pada gambar kanker payudara yang terdeteksi dan faktor lainnya. Meskipun kanker payudara termasuk ke dalam kategori penyakit yang ganas terdapat pula tingkatan golongan dari penyakit ini. Dalam upaya pengelompokkan dari identifikasi kanker payudara untuk dapat diklasifikasikan dan dicocokkan ke golongan yang 1 = *Malignant* (Ganas) dan 0 = *Benign* (Jinak). Oleh karena itu, pemanfaatan teknik resampling menjadi sangat penting karena dapat meningkatkan keakuratan model, mengatasi ketidakseimbangan kelas, mengurangi *overfitting*, mempercepat proses pelatihan model, serta meningkatkan kemampuan untuk menafsirkan hasil model.

Dengan semakin berkembangnya teknologi informasi, metode *machine learning* digunakan untuk mempelajari pola-pola dan hubungan yang ada dalam data kanker payudara, sehingga dapat membantu dalam penentuan golongan penyakit dan pengambilan keputusan terkait pengobatan yang tepat. namun, dalam pengolahan data kanker payudara terdapat masalah klasifikasi yang cukup kompleks karena data yang tidak seimbang antara kelas ganas dan jinak. Teknik resampling merupakan teknik yang berfungsi untuk menyelesaikan masalah ketidakseimbangan kelas melalui tahap eliminasi data dari kelas mayoritas atau duplikasi data pada kelas minoritas (Oversampling). Oleh karena itu, teknik resampling menjadi penting untuk dilakukan agar model yang dibangun dapat memiliki akurasi yang tinggi dan dapat diandalkan dalam pengambilan keputusan. Metode ini juga dapat membantu para ahli medis untuk mengurangi kesalahan diagnosis dalam waktu yang lebih singkat namun lebih rinci. Metode ini akan mendeteksi apakah kanker tersebut masuk ke dalam kelompok kanker ganas atau kanker jinak

Adapun tujuan Analisis ini adalah memperoleh hasil untuk meningkatkan akurasi model dengan penerapan teknik Resampling (*Holdout*, *Cross-Validation*), Standarisasi *(Z-Score)*, Oversampling (SMOTE), *ROC (Receiver Operating Characteristic)* pada data kanker payudara.

II. Metode

2.1. Data Penelitian

Teknik Diagnosis FNA (Aspirat Jarum Halus) oleh interpretasi komputasi melalui metode machine learning dengan tujuan menghasilkan akurasi yang tinggi pada classifier dan tingkat negatif palsu yang rendah. Fitur dihitung dari gambar digital dari aspirasi jarum halus (FNA) dari massa payudara yang ada. Hal ini menggambarkan karakteristik inti sel yang ada pada gambar ruang 3 dimensi yang dijelaskan dalam: [K. P. Bennett dan O. L. Mangasarian: "Diskriminasi Pemrograman Linear Kuat dari Dua Set Tak Terpisahkan Linear", Metode Optimasi dan Perangkat Lunak 1, 1992, 23-34]. Dengan informasi atribut berupa nomor identitas serta diagnosis (M = ganas, B = jinak)3-32), Data yang digunakan adalah data komponen atau faktor penyebab terjadinya kanker payudara yang terjadi terhadap orang pada umumnya seperti radius (*mean of distances from center to points on the perimeter*), texture (*standard deviation of gray-scale values*), perimeter, area, smoothness (*local variation in radius lengths*), compactness (perimeter^2 / area - 1.0), concavity (*severity of concave portions of the contour*), concave points (*number of concave portions of the contour*), symmetry, fractal dimension *("coastline approximation"* - 1). Terdapat pula ketetapan yaitu nilai atribut yang hilang: tidak ada dan distribusi kelas tergolong sebanyak 357 jinak, 212 ganas.

2.2. Variabel Tugas Besar

Variabel yang digunakan dalam tugas besar ini terdiri dari variabel dependen (Y) dan beberapa variabel independen (X) dengan rincian sebagai berikut :

1. Variabel dependen

   Y = Diagnosis , 1 = Ganas, 0 = Jinak

2. Variabel independen

   X_1= radius_mean, X_2= texture_mean, X_3= perimeter_mean, X_4= area_mean dan X_5= smoothness_mean.

2.3. Analisis Data

Software yang digunakan pada pengolahan data ini berupa software RStudio. Metode Resampling dengan penerapan teknik resampling (*Holdout*, *Cross-Validation*), Standarisasi *(Z-Score)*, Oversampling (SMOTE), *ROC (Receiver Operating Characteristic)* . Berikut adalah prosedur analisis data yang diterapkan dalam penelitian ini:

1. Mengimport dataset menggunakan library readr dan membaca file "Breast_Cancer.csv" ke dalam variabel cancer.

2. Mengubah kolom "diagnosis" menjadi variabel respon biner dengan nilai 0 dan 1.

3. Mendefinisikan objek task_cancerdiagnose sebagai objek klasifikasi menggunakan fungsi TaskClassif$new().

4. Menggunakan learner lrn("classif.log_reg") untuk menerapkan regresi logistik pada data menggunakan MLR3. Learner ini akan digunakan untuk memodelkan dan melakukan prediksi pada data.

5. Melihat hasil metrik evaluasi yang tersedia dalam MLR3 menggunakan fungsi mlr_measures dan mlr_measures$msr_tbl(). Fungsi ini memberikan informasi tentang metrik yang dapat digunakan untuk mengukur performa model.

6. Melakukan resampling menggunakan metode holdout menggunakan fungsi rsmp("holdout"). Resampling ini membagi data menjadi dua bagian: data pelatihan dan data pengujian.

7. Melakukan pelatihan dan pengujian model menggunakan fungsi resample() dan benchmark(). Fungsi ini menghitung performa model dengan menggunakan strategi resampling yang telah ditentukan.

8. Melihat hasil performa model dengan menggunakan metrik akurasi (classif.acc).

9. Membangun resampling dengan menggunakan fungsi rsmp(). Resampling ini mempartisi data menjadi subset pelatihan dan subset pengujian.

10. Melakukan evaluasi model dengan menggunakan fungsi resample() dan rsmp(). Hasil performa model diukur menggunakan metrik akurasi.

11. Melakukan inspeksi terhadap objek hasil resample menggunakan fungsi as.data.table(). Hal ini membantu dalam melihat prediksi dan confusion matrix yang dihasilkan oleh model.

12. Melakukan plot visualisasi hasil resampling menggunakan fungsi autoplot(). Plot ini dapat berupa histogram atau boxplot.

13. Melakukan evaluasi ROC (Receiver Operating Characteristic) menggunakan fungsi confusion_matrix() dan roc() dari package precrec. ROC digunakan untuk mengevaluasi kemampuan model dalam membedakan dua kelas.

14. Menampilkan kurva ROC menggunakan fungsi autoplot() dari package precrec.

III. HASIL DAN PEMBAHASAN

Resampling ini menggunakan *Cross Validation* dan *Holdout* untuk melakukan pengambilan sampel dari dataset *Breast Cancer* serta membagi data menjadi dua subset yaitu *training set* dan *test set* dalam membangun model. Analisis model dengan ROC dalam memanfaatkan *Confusion matrix-based measure* sepeti *True Positive Rate* (TPR), *False Positive Rate* (FPR), dan *Area Under the Curve* (AUC) dalam mengukur performa dari model untuk dapat membedakan *class* positif dan negatif. Analisis ROC pada nilai TPR dan FPR setiap *threshold* memiliki nilai probabilitas untuk memprediksi hasil dari ROC *Curve* yang menggambarkan hubungan antara TPR dan FPR.

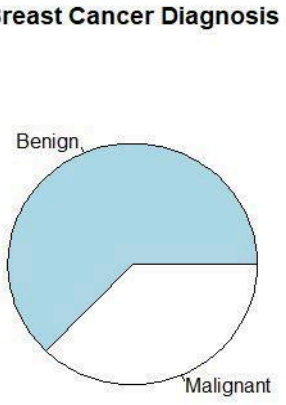
Resampling untuk memperoleh hasil distribusi nilai TPR, FPR, dan AUC. Hal tersebut dapat lebih stabil dan akurat, sehingga evaluasi dari performa model dengan ROC lebih efektif.

A. Proses Resampling

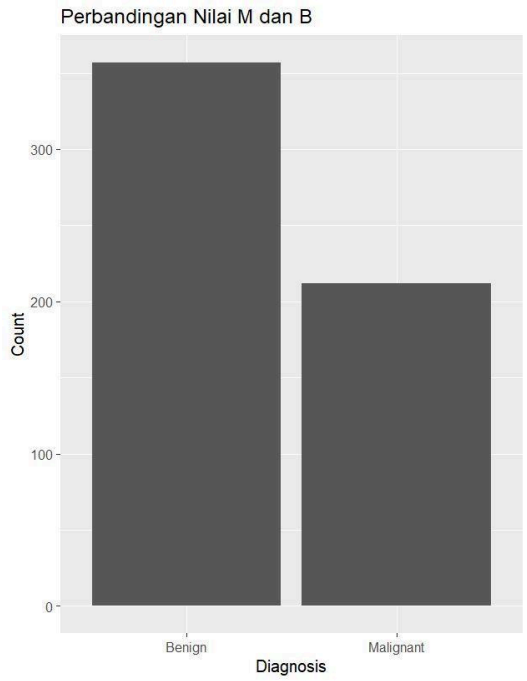
Proses resampling, dilakukan pemeriksaan proporsi dari hasil diagnosis pasien kanker yaitu 0 diartikan sebagai kanker jinak dan 1 berarti kanker ganas. Dapat dilihat bahwa proporsi kanker yang bersifat jinak lebih banyak dibandingkan kanker ganas pada dataset yang ada.

```
diagnosis
        1         0
0.3725835 0.6274165
```

Berikut visualisasi diagram pie chart.



Gambar. Pie chart diagnosis breast cancer nilai Malignant dan Benign.



Gambar. histogram hasil perbandingan nilai M dan B.

Interpretasi model dilakukan untuk melihat bagaimana pengaruh peubah prediktor terhadap respon menurut model. Model yang digunakan ini yaitu regresi logistik, maka dilakukan interpretasi model untuk mengukur besarnya nilai koefisien dan p-value yang dapat menggambarkan bagaimana pengaruh peubah prediktor terhadap respon.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95590  -0.14839  -0.03943   0.00429   2.91690

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.35952   12.85259  -0.573   0.5669
area_mean               0.03980    0.01674   2.377   0.0174 *
compactness_mean       -1.46242   20.34249  -0.072   0.9427
concave.points_mean    66.82176   28.52910   2.342   0.0192 *
concavity_mean          8.46870    8.12003   1.043   0.2970
fractal_dimension_mean -68.33703   85.55666  -0.799   0.4244
perimeter_mean         -0.07151    0.50516  -0.142   0.8874
radius_mean            -2.04930    3.71588  -0.551   0.5813
smoothness_mean        76.43227   31.95492   2.392   0.0168 *
symmetry_mean          16.27824   10.63059   1.531   0.1257
texture_mean            0.38473    0.06454   5.961  2.5e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

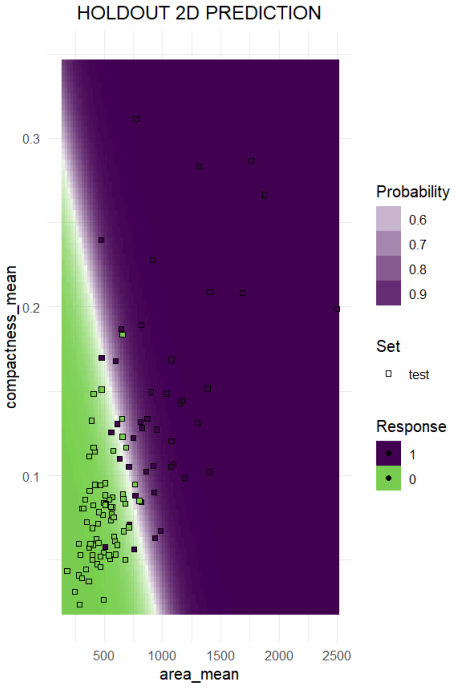
Sebelum masuk ke dalam metode holdout, terapkan metode holdout secara manual untuk mempartisi data yang terdapat dalam objek Task ke dalam satu set pelatihan (untuk melatih model) dan satu set pengujian (untuk memperkirakan kinerja generalisasi). Sebagai awal cepat untuk melakukan resampling dan benchmarking dengan package mlr3, kita menunjukkan contoh singkat bagaimana melakukannya dengan fungsi convenience resample() dan benchmark() .

Secara khusus, kita menunjukkan cara memperkirakan kinerja generalisasi learner pada tugas yang diberikan dengan metode holdout menggunakan resample() dan cara menggunakan benchmark() untuk membandingkan dua learner pada task.

B. Holdout dan Cross Validation (CV) Prediction

- Resampling menggunakan metode holdout. Secara default, holdout metode akan menggunakan 2/3 data sebagai data training dan 1/3 sebagai data test. kita dapat mengatur lebih spesifik lagi parameter rasio untuk holdout dengan meng-update rasio. sebagai contoh kita membangun objek resampling untuk holdout dengan split 80:20.

- Holdout hanya mengestimasi performa dengan menggunakan single set atau satu set. Untuk mendapatkan estimasi kinerja yang lebih andal dengan memanfaatkan semua data yang tersedia, digunakan metode resampling lainnya yaitu resampling CV dengan menyiapkan 10-fold cross-validation.



Gambar. Scatter Plot Holdout Prediction

- Kemudian, metode resampling CV terlebih dahulu dilakukan pengujian model menggunakan fungsi resample() dan benchmark() yang digunakan untuk menghitung performa model dengan teknik resampling holdout dan didapatkan hasil sebesar 0.9 yang menunjukkan bahwa model yang digunakan dalam resampling holdout memiliki tingkat akurasi yang tinggi yaitu 90%.

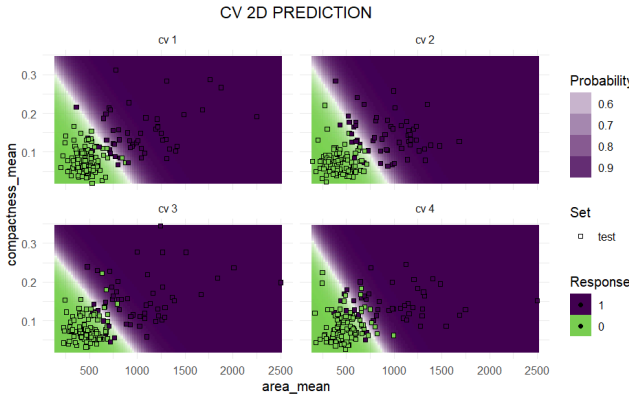
```
classif.acc
        0.9
```

- Lakukan teknik resampling dengan metode CV dengan folds = 4. Metode ini akan membagi data menjadi 4 subset yang berbeda untuk dilakukan evaluasi. Selanjutnya, menghitung skor akurasi setiap folds yang ada, dan didapatkan hasil yang cukup akurat yang mana nilai akurasi yang berbeda dalam setiap folds menggambarkan sejauh mana

model dapat menggeneralisasi dengan baik pada data yang tidak terlihat selama proses pelatihan.

iteration <int>	classif.acc <dbl>
1	0.9160839
2	0.9366197
3	0.9788732
4	0.9014085

- Secara keseluruhan nilai akurasi dari teknik resampling menggunakan metode CV menunjukkan nilai sebesar 0.9332463.



Gambar. Scatter Plot Cross Validation Prediction

Proses instansiasi strategi resampling dilakukan untuk menghasilkan pemisahan uji train untuk task tertentu, hal ini dilakukan dengan memanggil metode $instantiate() dari objek Resampling yang dibangun sebelumnya pada Task. Memanifestasikan partisi tetap dan menyimpan indeks baris untuk set pelatihan dan pengujian langsung di objek resampling Instansiasi sangat relevan adalah ketika tujuannya adalah untuk membandingkan beberapa learner secara adil.

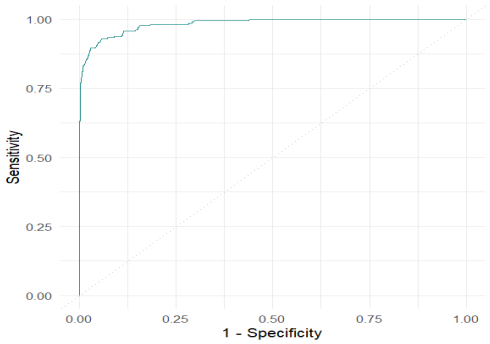
Menggunakan pemisahan tes-latihan yang sama untuk mendapatkan hasil yang sebanding. Artinya, kita perlu memastikan bahwa semua learner yang akan dibandingkan menggunakan data pelatihan yang sama untuk membuat model dan bahwa learner menggunakan data pengujian yang sama untuk mengevaluasi kinerja model.

Kemudian, membuat plot visualisasi hasil resampling. Kali ini dilakukan pembuatan plot visualisasi prediksi 2 dimensi dari hasil resampling dalam task menggunakan teknik resampling cv dengan dua fitur sebagai prediktor yaitu compactness_mean dan area_mean pada metode holdout dan cross validation prediction.

C. ROC (Receiver Operating Characteristic)

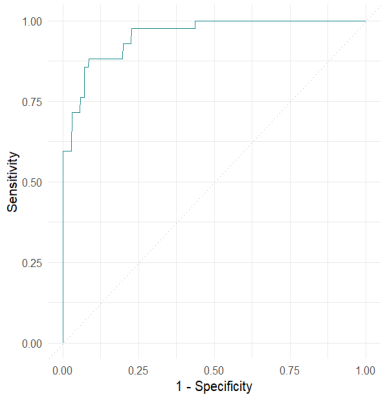
Analisis ROC yang banyak digunakan untuk mengevaluasi pengklasifikasian biner pada confusion matrix dan mengukur kemampuan pengklasifikasian untuk memisahkan dua kelas (yaitu, kinerja diskriminasi). Pertama dilakukan pembuatan confusion matrix yang menunjukkan jumlah kelas true positif, true negatif, false positif, dan false negatif dari data latih dan data uji yang ditentukan.

Berdasarkan hasil yang didapatkan, nilai dikaitkan dengan proses diagnosis kanker maka nilai yang memprediksi bahwa pasien menderita kanker ganas dan hasilnya benar bahwa pasien menderita kanker ganas sebanyak 34 kasus. Tetapi terdapat 2 nilai yang memprediksi bahwa pasien menderita kanker ganas tetapi pada hasil yang sebenarnya pasien menderita kanker jinak. Hal ini tentu menunjukkan keakuratan yang tidak tepat 100% namun persentase nilai akurasinya sebesar 91.15% dan error klasifikasi sebar 9.73%. Kurva ROC dari proses resampling holdout dan resampling cv sebagai berikut.



Gambar. *kurva ROC cross validation*

melakukan standarisasi data dan oversampling menggunakan fungsi po() dari package mlr3 pipelines. Dalam contoh ini, dilakukan standarisasi menggunakan z-score dan oversampling menggunakan SMOTE (Synthetic Minority Over-sampling Technique).



Gambar. *kurva ROC holdout*

Hasil dari kurva ROC adalah grafik yang berbentuk kurva yang menggambarkan hubungan antara TPR (Sensitivity) dan FPR. Pada kurva ROC, nilai TPR diplot pada sumbu y (ordinat) sedangkan nilai FPR diplot pada sumbu x (absis). Setiap titik pada kurva ROC mewakili pasangan nilai TPR dan FPR yang didapatkan pada suatu ambang batas tertentu.

D. SMOTE (Synthetic Minority Over-sampling Technique)

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla sebagai salah satu solusi dalam menangani data tidak seimbang dengan prinsip yang berbeda dengan metode oversampling yang telah diusulkan sebelumnya[8]. Bila Metode oversampling berprinsip memperbanyak pengamatan secara acak, Metode SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan. Data buatan atau sintesis tersebut dibuat berdasarkan k-tetangga terdekat (k nearest neighbor). Jumlah k-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya[9].

Dalam hal ini, kita ingin meningkatkan jumlah amatan pada kelas minoritas (diagnosis = 0) agar sebanding dengan jumlah amatan pada kelas mayoritas (diagnosis = 1). Dalam kode tersebut, parameter "dup_size=1" menunjukkan bahwa kita ingin menambahkan jumlah amatan kelas minoritas sebanyak 1 kali dari jumlah amatan awal.

metode ini bekerja dengan mengelompokkan data terdekat yang dipilih berdasarkan jarak Euclidean antara kedua data. penentuan jumlah replikasi harus sesuai dengan jumlah k pada *nearest neighbour*, jika jumlah replikasi sebanyak n maka jumlah k harus sebanyak n-1.

Misalkan terdapat dua struktur data dengan p dimensi yaitu $x^t = [x_1, x_2, \dots, x_p]$ dan $y^t = [y_1, y_2, \dots, y_p]$, maka jarak Euclidean d$(x, y)$ yang dihasilkan antara kedua data ditunjukan pada persamaan sebagai berikut :

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2}$$

Tujuan dari proses tersebut adalah untuk mempersiapkan data agar siap digunakan dalam analisis lebih lanjut, seperti pembuatan model klasifikasi untuk mendiagnosis kanker payudara berdasarkan variabel-variabel yang ada.

```
Call:
stats::glm(formula = task$formula(), family = "binomial", data = data,
    model = FALSE)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.82661  -0.29015   0.00181   0.15640   2.79850

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.3438     0.1526   2.252   0.0243 *
area_mean         4.3132     0.3913  11.022   <2e-16 ***
compactness_mean  1.7486     0.1938   9.021   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1076.94  on 780  degrees of freedom
Residual deviance:  340.87  on 778  degrees of freedom
AIC: 346.87

Number of Fisher Scoring iterations: 7
```

Gambar. Ringkasan model setelah dilakukan oversampling dengan SMOTE

IV. KESIMPULAN

A. Resampling

Penggunaan teknik resampling dalam metode cross validation sangat efektif dalam meningkatkan kinerja model dengan memprediksi penyakit kanker payudara pada dataset *Wisconsin Breast Cancer*. Metode ini memberikan estimasi kinerja yang lebih baik dalam memaksimalkan penggunaan data dan memberikan wawasan yang lebih baik mengenai kemampuan model dalam melakukan prediksi kanker jinak atau ganas.

B. Efektivitas Teknik Resampling

Berdasarkan analisis proses resampling menggunakan metode holdout, model regresi logistik berhasil mencapai tingkat akurasi sebesar 90% dengan menggunakan metode cross validation untuk melakukan evaluasi dengan membagi data menjadi 4 subset yang berbeda. Kinerja model meningkat menjadi 93.32% yang menunjukkan bahwa pengguna metode cross validation memberikan estimasi kinerja yang lebih andal dan dapat mengoptimalkan pemanfaatan data yang tersedia.

Analisis ROC dan AUC memberikan informasi penting tentang kinerja model dalam membedakan kelas kanker jinak dan ganas dengan tingkat akurasi sebesar 91.15% dan nilai error dari klasifikasi sebesar 9.73%. Pada kurva ROC ditunjukkan bahwa model memiliki kemampuan diskriminasi yang cukup baik dengan AUC yang dapat memberikan gambaran kinerja model secara agregat.

C. Plotting Resampling

Visualisasi prediksi dari 2 dimensi menggunakan scatter plot untuk memberikan gambaran sejauh mana model tersebut menggeneralisasikan data yang tidak terlihat selama proses pelatihan. Scatter plot yang dihasilkan dari visualisasi 2 dimensi prediksi resampling dengan metode holdout dan CV menggambarkan hubungan antara dua fitur (compactness_mean dan area_mean) dalam memprediksi kelas kanker payudara (jinak atau ganas).

Dalam scatter plot, setiap titik merepresentasikan satu data pasien. Sumbu x (horizontal) mewakili nilai compactness_mean, sedangkan sumbu y (vertikal) mewakili nilai area_mean. Pada scatter plot, titik-titik tersebut diberi warna berdasarkan kelas prediksi dari model.

Dengan menggunakan metode holdout, scatter plot menunjukkan sebaran titik-titik data dan pemisahan kelas kanker jinak dan ganas berdasarkan fitur compactness_mean dan area_mean. Karena warna plot terlihat pekat, hal ini menandakan data test yang digunakan menghasilkan probabilitas keakuratan data sebesar 90% dalam mengklasifikasikan kanker bersifat ganas atau jinak.

Dalam metode CV, scatter plot menggambarkan hasil prediksi dari setiap fold dalam cross-validation. Scatter plot ini dapat memberikan gambaran tentang sejauh mana model dapat menggeneralisasi data yang tidak terlihat selama proses pelatihan. Karena warna plot dari keempat folds CV yang ada terlihat pekat, hal ini menandakan data test yang digunakan menghasilkan probabilitas keakuratan data sebesar 90%, maka dapat dikatakan bahwa model memiliki kemampuan yang baik dalam membedakan kanker bersifat ganas atau jinak berdasarkan fitur yang diberikan.

Daftar Pustaka

Fitriani, H., Yasin, & Tarno. "Penanganan klasifikasi kelas data tidak seimbanh dengan random oversampling pada naive bayes", (2021)

Wahyu Nugraha, dkk. "Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4.5, Random Forest, dan SVM", (2021)

Robert Dessaix, "Russia: The End of an Affair," Australian Humanities Review 6 (1997), diakses pada 28 Februari 2010, https://www.australianhumanitiesreview.org/archive/Issue-June-1997/dessaix.html

Chawla, Nitesh V., dkk. 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 61, 2018, 863-905.

Barro, Rossi Azmatul, dkk. (2013). Penerapan Synthetic Minority Oversampling Technique (SMOTE) terhadap Data Tidak Seimbang pada Pembuatan Model Komposisi Jamu. Jurnal Xplore Vol.1, 2013, 1-6.

Azis, dkk. "Pendekatan machine learning yang Efisien untuk Kanker Payudara" , (2009).

Borges, Lucas Rodrigues. "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection Borges", (2015)