Klasifikasi Golongan Penyakit Kanker Payudara Menggunakan Model Regresi Logistik

1st Dimas Rizky Ramadhani
121450025

2nd M. Akbar Resdika
121450066

3nd Veni Zahara Kartika
121450075

4th Mujadid Choirus Surya
121450015

5th Muhammad Fahrul Aditya
121450156

***Abstract*— Laporan ini melakukan analisis pengklasifikasian golongan penyakit kanker payudara. Penyakit kanker payudara memiliki dua golongan yaitu biasa dan ganas. Model yang dilakukan pada analisis ini menggunakan model regresi logistik untuk memperoleh model berdasarkan variabel-variabel prediksi yang berhubungan dengan variabel respon. Dilakukan model regresi sebelum dilakukan PCA dan sesudah dilakukan PCA. Mendapatkan hasil prediksi model regresi logistik setelah dilakukan PCA memiliki performa yang lebih baik dalam melakukan prediksi kategori pasien yang terkena kanker payudara jinak dan kanker payudara ganas.**

Kata Kunci : Pengklasifikasian, Golongan, Regresi, Logistik, PCA, Performa

I. PENDAHULUAN

Kanker payudara merupakan penyakit tidak menular dimana penyakit ini sangat rentan terhadap wanita dan merupakan jenis kanker paling mendominasi yang terjangkit penyakit ini, bahkan melebihi kanker serviks (Hayati & Purwaningsih, 2017, 122). kanker payudara menempati posisi pertama dalam kategori penyakit kanker terbanyak di indonesia, data dari Global Burden of Cancer (GLOBOCAN) 2020 menunjukkan kasus kanker payudara mencapai 68.856 kasus (16,6%) dari total 396.914 kasus kanker di indonesia (*Kementerian Kesehatan Republik Indonesia*, 2022). Faktor-faktor yang berpengaruh pada penyakit kanker payudara ini menurut Moningkey dan Kodim(2008) adalah *menarche* pada umur muda, *menopause* pada umur lebih tua, kehamilan pertama pada umur tua, penggunaan hormon-hormon eksogen, penyakit fibrokistik, obesitas, konsumsi lemak, riwayat keluarga dengan kanker payudara (Nani, 2009, 61)

Kanker Payudara dibagi menjadi 2 golongan yaitu kanker payudara biasa atau sering disebut jinak dan kanker payudara malignant atau biasa disebut ganas, kanker payudara jinak umumnya ditandai dengan benjolan kecil bulat dan lembut, keadaan perkembangan serta pertumbuhan kanker payudara jinak biasanya tidak bersifat kanker, sehingga kanker ini akan terdeteksi tetapi tidak merusak jaringan yang ada di dekatnya. Pada kanker payudara tingkat ganas ditandai dengan bentuk yang tidak simetris, kasar, terasa nyeri dan yang lainnya, pada kanker payudara tingkat ganas akan menjalar serta merusak jaringan dan organ lain yang ada di dekatnya. (Atthalla et al., 2018, 148)

Dalam hal ini kami ingin melakukan pengklasifikasian golongan penyakit kanker payudara menggunakan model regresi logistik. Menurut Hosmer dan Lemeshow tujuan melakukan analisis data menggunakan regresi logistik adalah menjelaskan hubungan antara variabel respon dengan variable-variabel prediktornya sehingga akan menemukan model yang terbaik (Wella Putri, 2021, 1).

II. METODE

Regresi logistik adalah pendekatan untuk membangun model prediksi yang digunakan ketika variabel terikat memiliki skala dikotomis, seperti Ya/Tidak, Baik/Buruk, atau Tinggi/Rendah. Berbeda dengan regresi linear (Ordinary Least Squares/OLS), regresi logistik tidak mengasumsikan distribusi normal untuk residual (kesalahan prediksi), melainkan mengikuti distribusi logistik.Dalam regresi logistik, tujuan utamanya adalah mengestimasi probabilitas

kejadian kategori tertentu berdasarkan nilai-nilai prediktor yang diberikan. (*Regresi Logistik*, n.d.)

Dalam penelitian kami dilaksanakan dengan cara pengambilan data set secara sekunder dari sumber kaggle dengan judul dataset " Breast Cancer Wisconsin (Diagnostic) Data Set ". Data yang memuat tentang Fine Needle Aspirasi (FNA) dari massa payudara, yang menggambarkan sebuah karakteristik inti sel dengan data yang ditampilkan dari dataset tersebut sebesar 569 observasi dari 33 variabel.

Dengan beberapa hal singkat yang telah dijabarkan tadi pada penelitian kali ini, kami akan menganalisis golongan penyakit kanker payudara menggunakan model regresi logistik. Fokus yang akan kami analisis adalah pengklasifikasian golongan kanker payudara jinak(B) dan golongan kanker payudara ganas(M) yang merupakan variabel dependen penggolongan tersebut dikategorikan dengan nilai 0 (golongan kanker payudara Jinak) dan 1 (golongan kanker payudara ganas).

Dalam menganalisis data yang telah kami dapatkan, kami menggunakan software R studio sebagai alat pengolahan data dan menggunakan metode *Principal Component Analysis* (PCA) untuk mengurangi jumlah variabel yang digunakan dan mengubah variabel yang berkorelasi menjadi tidak berkorelasi. (*Principal Component Analysis*, 2019)

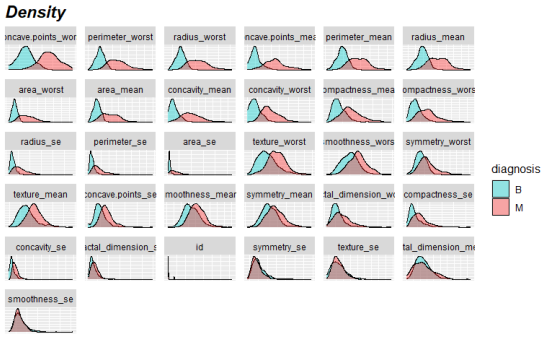
III. HASIL DAN PEMBAHASAN

A. Karakteristik dataset

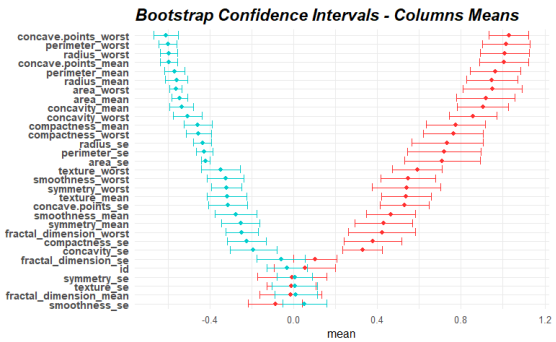
Penyakit kanker payudara merupakan salah satu kanker yang mendominasi di Indonesia terutama di kalangan perempuan. Pada analisis ini akan dibahas mengenai model dari faktor-faktor yang berpengaruh terhadap peluang terjadinya penyakit kanker payudara dengan dua kategori yakni kanker payudara jinak dan ganas. Data penelitian yang digunakan diambil melalui dataset pada website UCI Machine Learning Repository dengan judul dataset "Breast Cancer Wisconsin (Diagnostic)" (*Breast Cancer Wisconsin (Diagnostic) Data Set*, 1995). Dataset ini memiliki 32 atribut yakni *id*, *diagnosis* (*M* = ganas dan *B* = jinak), *radius_mean*, *texture_mean*, *perimeter_mean*, *area_mean*, *smoothness_mean*, *compactness_mean*, *concavity_mean*, *concave points_mean*, *symmetry_mean*, *fractal_dimension_mean*, *radius_se*, *texture_se*, *perimeter_se*, *area_se*, *smoothness_se*, *compactness_se*, *concavity_se*, *concave points_se*, *symmetry_se*, *fractal_dimension_se*, *radius_worst*, *texture_worst*, *perimeter_worst*, *area_worst*, *smoothness_worst*, *compactness_worst*, *concavity_worst*, *concave points_worst*, *symmetry_worst*, dan *fractal_dimension_worst*. Pada dataset ini variabel y atau variabel dependen merupakan diagnosis dengan dua kemungkinan yakni jinak yang didefinisikan sebagai 0 dan ganas yang didefinisikan sebagai 1.

B. Eksplorasi Data dan Metode PCA

Eksplorasi data merupakan suatu pendekatan untuk mengetahui informasi terkait kondisi dan sebaran yang dimiliki oleh sebuah data. Hal ini dilakukan sebagai langkah awal sebelum melakukan pengelolaan data lebih lanjut. Pada dataset "Breast Cancer Wiconsin (Diagnostic)" sebaran data antara variabel-variabel prediksi dengan variabel respon yang memiliki 2 kategori yaitu jinak dan ganas akan terlihat seperti pada gambar 1



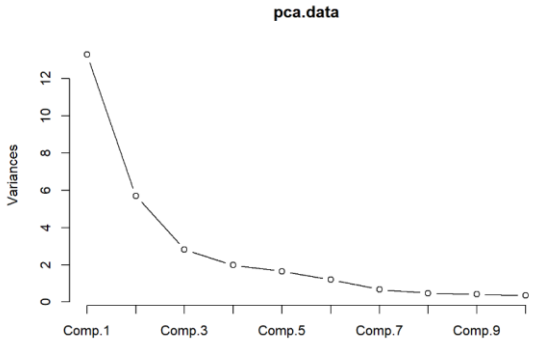
Gambar 1. Sebaran Data "Breast Cancer Wisconsin (Diagnostic)"



Gambar 2. *Bootstrap* seberan data "Breast Cancer Wisconsin(Diagnostic)"

Terlihat bahwa dataset "Breast Cancer Wisconsin (Diagnostic)" memiliki banyak variabel serta berdimensi tinggi sehingga rentan mengalami overfitting dan interpretasi yang rumit (Ahn et al., n.d.). Saat mengatasi data yang berdimensi tinggi maka diperlukan pendekatan untuk mengendalikan

kompleksitas model melalui metode reduksi dimensi salah satunya yakni metode PCA yang merupakan singkatan dari *Principal Component Analysis*. PCA adalah teknik mereduksi suatu set variabel yang memiliki dimensi cukup tinggi menjadi lebih rendah dan tetap mengandung sebagian besar data awal sehingga hasil analisis yang diberikan terhadap model nantinya akan sama dengan data aslinya. PCA memiliki cara kerja dengan memilih variabel yang dapat menjelaskan sebagian variabilitas pada data dan mengurangi dimensi data dengan membentuk variabel baru yang disebut sebagai *principal component*. *Principal component* yang dipilih harus memenuhi syarat yakni mempunyai *variance* terbesar. Pada dataset ini ketika dilakukan PCA dengan variabel terbaik yang digunakan maka terbentuk variabel baru sebanyak 6 *principal component*. Dapat dilihat pada gambar 3 yang menjelaskan hubungan antara *principal component* yang digunakan dengan *variance* yang dimiliki.



Gambar 3. Hubungan *Principal Component* dengan *Variance*

Berdasarkan gambar 3 terlihat bahwa comp.1 sampai comp.6 memiliki tingkat variance yang lebih tinggi dibandingkan dengan comp.7 hingga comp.10. Sehingga comp.1 sampai comp.6 yang akan digunakan sebagai variabel baru dalam membuat model regresi logistik.

C. Model Regresi Logistik

Pada dataset "Breast Cancer Wisconsin (Diagnostic)" analisis regresi logistik digunakan sebagai salah satu metode untuk memperoleh model berdasarkan variabel-variabel prediksi yang berhubungan dengan variabel respon yakni diagnosa kanker payudara termasuk kedalam golongan ganas atau tidak. Saat melakukan metode regresi logistik tahapan awal yang dilakukan adalah membagi data menjadi data train dan data set dengan partisi 70:30 sehingga data train pada dataset yang digunakan untuk memperoleh model memiliki 263 data yang tergolong kedalam kanker payudara jinak dan 142 data tergolong kedalam kanker payudara ganas. Model regresi logistik yang didapat sebelum dilakukan PCA adalah sebagai berikut

$$\pi(x) = \frac{e^{-9.864e+03 - 4.073e+02\ radius\_mean + 8.950e+01\ texture\_mean + 3.715e+04\ smoothness\_mean + 2.822e+04\ compactnes\_mean + 1.294e+04 concavity\_mean - 7.351e+04\ smoothness\_se + 2.219e+04 compactness\_se + 1.269e+05 concave.points\_se - 3.665e+05\ fractal\_dimension\_se + 8.071e+02 radius\_worst - 4.435e+01 perimeter\_worst - 9.986e+02 concavity\_worst + 4.418e+04\ fractal\_dimension\_worst}}{1+e^{-9.864e+03 - 4.073e+02\ radius\_mean + 8.950e+01\ texture\_mean + 3.715e+04\ smoothness\_mean + 2.822e+04\ compactnes\_mean + 1.294e+04 concavity\_mean - 7.351e+04\ smoothness\_se + 2.219e+04 compactness\_se + 1.269e+05 concave.points\_se - 3.665e+05\ fractal\_dimension\_se + 8.071e+02 radius\_worst - 4.435e+01 perimeter\_worst - 9.986e+02 concavity\_worst + 4.418e+04\ fractal\_dimension\_worst}} \quad (1)$$

Sedangkan ketika variabel yang digunakan merupakan variabel baru yang berasal dari metode PCA maka model regresi logistik yang diperoleh adalah sebagai berikut.

$$\pi(x) = \frac{e^{-0.4214 + 2.6929\ comp.1 + 1.4443\ comp.2 + 0.4237\ comp.3 - 0.8664\ comp.4 - 1.8044\ comp.5 - 0.6545\ comp.6}}{1+e^{-0.4214 + 2.6929\ comp.1 + 1.4443\ comp.2 + 0.4237\ comp.3 - 0.8664\ comp.4 - 1.8044\ comp.5 - 0.6545\ comp.6}} \quad (2)$$

D. Evaluasi Model Regresi Logistik

Pada regresi logistik untuk mengetahui performa dari sebuah model dapat dilakukan evaluasi menggunakan *confussion matrix*. *confussion matrix* merupakan tabel yang berisikan 4 buah kombinasi berdasarkan nilai aktual dan nilai prediksi dari data. Empat buah komponen ini terdiri dari *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN)

Pada confusion matrix model regresi logistik sebelum dilakukan PCA terlihat bahwa:

		Actual Condition	
	Total Samples	Jinak (0)	Ganas (1)
Output of Classifier	Jinak (0)	86	6
	Ganas (1)	8	64

Gambar 4. Confusion matrix sebelum dilakukan PCA

- TP yang didapat adalah 86 yang berarti model dapat memprediksi dengan benar atau mengklasifikasi 86 pasien didiagnosa terkena kanker payudara jinak.
- FP yang didapat adalah 6 yang berarti model memprediksi 6 pasien didiagnosa terkena kanker payudara jinak namun

ternyata pasien terkena kanker payudara ganas.

- FN yang didapat adalah 8 yang berarti model memprediksi 8 pasien didiagnosa terkena kanker payudara ganas namun ternyata pasien hanya terkena kanker payudara jinak.
- TN yang didapat adalah 64 yang berarti model dapat memprediksi dengan benar atau mengklasifikasi 64 pasien didiagnosa terkena kanker payudara ganas dengan benar.

```
               Accuracy : 0.9146
                 95% CI : (0.8609, 0.9525)
    No Information Rate : 0.5732
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8262

 Mcnemar's Test P-Value : 0.7893

            Sensitivity : 0.9149
            Specificity : 0.9143
         Pos Pred Value : 0.9348
         Neg Pred Value : 0.8889
             Prevalence : 0.5732
         Detection Rate : 0.5244
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.9146

       'Positive' Class : 0
```

Gambar 5. Nilai statistik model sebelum dilakukan PCA

Prediksi model ini memiliki akurasi sebesar 91.46%, *sensitivity* sebesar 91.49%, dan *specificity* sebesar 91.43% sehingga dapat dikatakan bahwa model memiliki performa yang cukup baik dalam memprediksi kategori pasien yang terkena kanker payudara jinak dan kanker payudara ganas.

Jika dibandingkan dengan model regresi logistik setelah dilakukan PCA terlihat bahwa:

		Actual Condition	
	Total Samples	Jinak (0)	Ganas (1)
Output of Classifier	Jinak (0)	83	0
	Ganas (1)	11	70

Gambar 6. Confussion matrix setelah dilakukan PCA

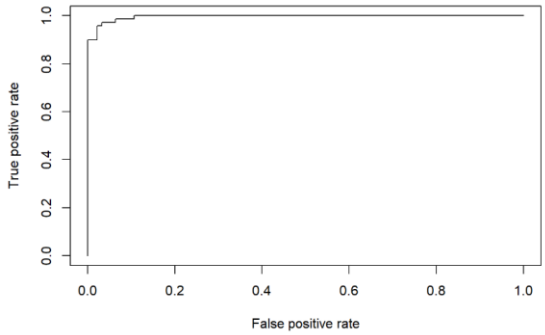
- TP yang didapat adalah 83 yang berarti model dapat memprediksi dengan benar atau mengklasifikasi 83 pasien didiagnosa terkena kanker payudara jinak.
- FP yang didapat adalah 0 yang berarti model memprediksi 0 pasien didiagnosa terkena kanker payudara jinak namun ternyata pasien terkena kanker payudara ganas.
- FN yang didapat adalah 11 yang berarti model memprediksi 11 pasien didiagnosa terkena kanker payudara ganas namun ternyata pasien hanya terkena kanker payudara jinak.
- TN yang didapat adalah 70 yang berarti model dapat memprediksi dengan benar atau mengklasifikasi 70 pasien didiagnosa terkena kanker payudara ganas .

```
               Accuracy : 0.9329
                 95% CI : (0.8832, 0.966)
    No Information Rate : 0.5732
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8656

 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 0.8830
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8642
             Prevalence : 0.5732
         Detection Rate : 0.5061
   Detection Prevalence : 0.5061
      Balanced Accuracy : 0.9415

       'Positive' Class : 0
```

Gambar 7. Nilai statistik model setelah dilakukan PCA

Prediksi model ini memiliki akurasi sebesar 93.29%, *sensitivity* sebesar 88.30%, dan *specificity* sebesar 100% sehingga dapat dikatakan bahwa model memiliki performa yang lebih baik dalam memprediksi kategori pasien yang terkena kanker payudara jinak dan kanker payudara ganas dibandingkan dengan menggunakan model regresi logistik tanpa dilakukan PCA.

Selain dengan *confussion matrix*, evaluasi model dapat diketahui melalui kurva ROC seperti pada gambar 8.



Gambar 8. Kurva ROC

Berdasarkan kurva ROC pada gambar 8 terlihat bahwa akurasi yang dimiliki oleh model berdasarkan perfoma dalam memprediksi sangatlah baik karena hampir mendekati nilai 100% yakni

99.5%. Model ini berarti hampir tidak melakukan kesalahan dalam mengklasifikasi data terhadap diagnosa kanker payudara jinak dan kanker payudara ganas.

IV. KESIMPULAN

Dalam analisis regresi logistik pada dataset "Breast Cancer Wisconsin (Diagnostic)", dilakukan pemodelan untuk memprediksi kategori diagnosa kanker payudara sebagai jinak atau ganas. Terdapat dua model regresi logistik yang dievaluasi: sebelum dilakukan PCA dan setelah dilakukan PCA.

Didapatkan hasil dari model regresi logistik dengan performa lebih baik setelah dilakukan PCA dengan akurasi sebesar 93.29%, sensitivity sebesar 88.30%, dan specificity sebesar 100%. Hal ini dibuktikan dengan hasil evaluasi kurva ROC menunjukkan bahwa model regresi logistik setelah dilakukan PCA memiliki akurasi yang sangat baik, mendekati 99.5%. Sehingga model ini hampir tidak melakukan kesalahan dalam mengklasifikasikan data terkait diagnosa kanker payudara jinak dan ganas.

Dengan hasil dari analisis model yang kami buat, dapat diambil kesimpulan bahwa model regresi logistik setelah dilakukan PCA memberikan performa yang lebih baik dalam memprediksi kategori pasien yang terkena kanker payudara jinak dan ganas. Meskipun demikian, masih perlu diperhatikan dan diperbaiki beberapa kesalahan prediksi yang terjadi untuk meningkatkan performa model secara keseluruhan.

DAFTAR PUSTAKA

[1] D. L. Wella Putri, Penyunt.“Peningkatan Ketepatan Klasifikasi Model Regresi Logistik Biner dengan Metode Bagging (Bootstrap Aggregating),” *Indonesian Journal of Mathematics and Natural Sciences,* p. 72, 2021.

[2] I. N. Atthalla, A. Jovandy dan H. Habibie, Penyunt.“Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor,” *Computer Science and ICT,* vol. Vol.4 No.1, p. 151, 2018.

[3] “Kementerian Kesehatan Republik Indonesia,” 4 February 2022. [Online]. Available: https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html. [Diakses 15 May 2023].

[4] S. Hayati dan D. Purwaningsih, “Hubungan Dukungan Keluarga Dengan Kualitas Hidup Penderita Kanker Payudara,” *Jurnal Keperawatan BSI,* vol. Vol. V No 2, p. 129, 2 9 2017.

[5] D. Nani, Penyunt.“ANALISIS FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN KEJADIAN KANKER PAYUDARA DI RUMAH SAKIT PERTAMINA CILACAP,” *Jurnal Keperawatan Soedirman,* vol. Volume r. no 2, p. 66, 07 2009.

[6] “Regresi Logistik,” [Online]. Available: https://www.statistikian.com/2015/02/regresi-logistik.html. [Diakses 17 May 2023].

[7] [Online]. Available: https://eksplorasidata.mipa.ugm.ac.id/. [Diakses 17 May 2023].

[8] “Principal Component Analysis,” 22 September 2019. [Online]. Available: https://arofiqimaulana.com/principal-component-analysis/. [Diakses 17 May 2023].

[9] “Breast Cancer Wisconsin (Diagnostic) Data Set,” 1 November 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. [Diakses 17 May 2023].

[10] H. Ahn, H. Moon, M. j. Fazzari, N. Lim, J. J. Chen dan R. L. Kodell, “Classification by ensembles from random partitions of high-dimensional data,” [Online].