LAPORAN TUGAS BESAR
STATISTIKA DAN PROBABILITAS

Penulis

Restu Firmansyah

PROGRAM STUDI S1 TEKNIK INFORMATIKA

FAKULTAS INFORMATIKA

INSTITUT TEKNOLOGI TELKOM PURWOKERTO

2024

a. Tentang Data Set

Dataset Cryotherapy digunakan untuk memprediksi hasil perawatan cryotherapy berdasarkan beberapa fitur atau atribut pasien. Cryotherapy adalah metode perawatan medis yang menggunakan suhu dingin untuk menghilangkan kutil dan lesi kulit lainnya. Dalam dataset ini, setiap baris mewakili data dari satu pasien yang menjalani cryotherapy.

Dataset ini mengandung variabel-variabel seperti usia (age), waktu (time), area yang diobati (area), jumlah kutil (number of warts), jenis kelamin (sex), dan hasil dari perawatan (result of treatment). Data ini penting untuk menganalisis hubungan antara faktor-faktor ini dengan efektivitas cryotherapy dalam mengobati kutil, memberikan wawasan tentang seberapa baik metode ini berfungsi pada berbagai kelompok usia, jenis kelamin, dan ukuran area yang diobati.

b. Frekuensi Tiap Kelas

Untuk menentukan jumlah kelas yang akan dipakai pada data, kita bisa menggunakan rumus 1+(3,3*LOG(n)). Dimana n adalah jumlah data dari dataset, yakni 90. Sehingga didapatkan jumlah kelasnya adalah 7,44 dan dibulatkan ke atas menjadi 8.

	MIN	MAX	RANGE	MODE	Lebar Kelas	
Age	15	67	52	15	6,5	7
Time	0,25	12	11,75	12	1,46875	2
Number_of_Warts	1	12	11	2	1,375	2
Area	4	750	746	100	93,25	94

Pada tabel di atas, dapat diketahui nilai MIN, MAX, RANGE, dan MODE dari tiap-tiap variabel. Dimana RANGE didapatkan dari selisih antara nilai MIN dan MAX dari variabel. Kemudian untuk menentukan lebar kelas, kita bisa membagi nilai RANGE tiap variabel dengan jumlah kelas yang telah ditentukan, yakni 8. Adapun untuk frekuensi tiap kelas sebagai berikut:

1. Variabel Time

Interval	Frekuensi
0,25 - 1,75	4
1,75 - 3,25	5
3,25 - 4,75	14
4,75 - 6,25	11
6,25 - 7,75	3
7,75 - 9,25	13
9,25 - 10,75	17
10,75 - 12,25	23

2. Variabel Age

Interval	Frekuensi
15 - 21	36
22 - 28	15
29 - 35	21
36 - 42	9
43 - 49	0
50 - 56	3
57 - 63	2
64 - 70	4

3. Variabel Number_of_Warts

Interval	Frekuensi
1 - 2,374	24
2,375 - 3,749	13
3,750 - 5,124	14
5,125 - 6,499	6
6,500 - 7,874	6
7,875 - 9,249	10
9,250 - 10,624	4
10,625 - 12,000	13

4. Variabel Area

Interval	Frekuensi
4 - 97	66
98 - 191	21
192 - 285	0
286 - 379	0
380 - 473	0
474 - 567	0
568 - 661	0
662 - 755	3

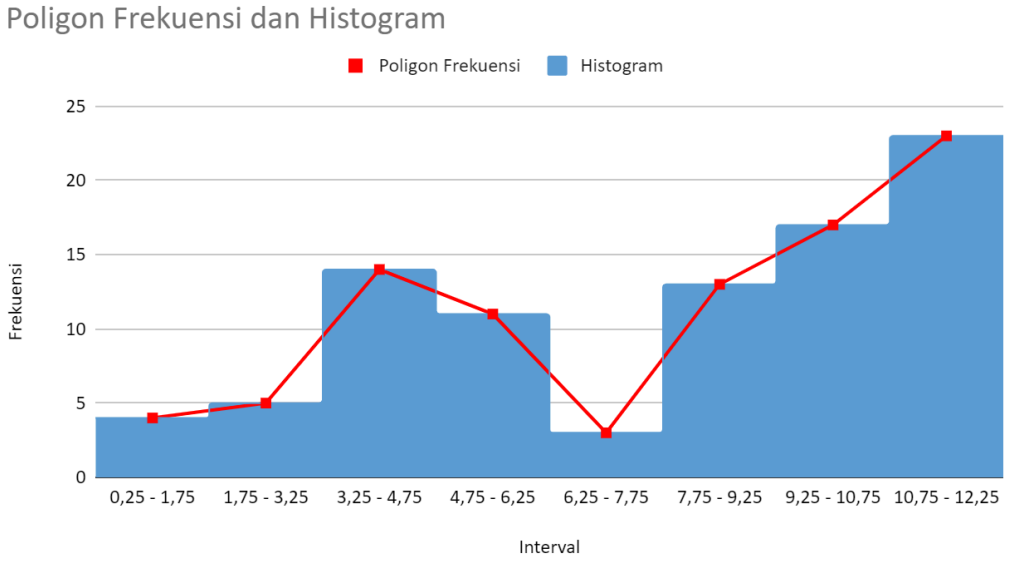
c.

d. Analisis Statistik Deskriptif

1. Variabel Time

Ukuran Pemusatan					
Mean	Median	Modus	Kuartil 1	Kuartil 2	Kuartil 3
7,67	8,173076923	10,56034483	4,196428571	8,173076923	10,2826087

Interval	Frekuensi	Batas Bawah	Batas Atas	Frekuensi Kumulatif	Histogram	Poligon Frekuensi
0,25 - 1,75	4	-0,25	2,25	4	4	4
1,75 - 3,25	5	1,25	3,75	9	5	5
3,25 - 4,75	14	2,75	5,25	23	14	14
4,75 - 6,25	11	4,25	6,75	34	11	11
6,25 - 7,75	3	5,75	8,25	37	3	3
7,75 - 9,25	13	7,25	9,75	50	13	13
9,25 - 10,75	17	8,75	11,25	67	17	17
10,75 - 12,25	23	10,25	12,75	90	23	23



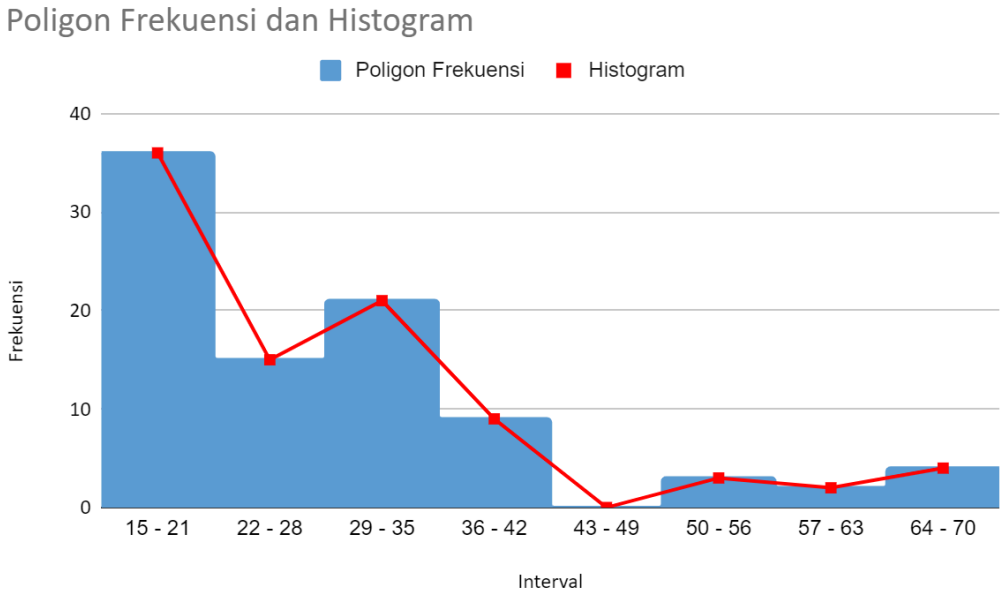
Interval	Frekuensi	Nilai Tengah	|x-xmi|	fi(|x-xmi|)	|x-xmi|^2	fi*|x-xmi|^2
0,25 - 1,75	3	1	6,666666667	20	44,44444444	133,3333333

1,75 - 3,25	3	2,5	5,166666667	15,5	26,69444444	80,08333333
3,25 - 4,75	10	4	3,666666667	36,66666667	13,44444444	134,4444444
4,75 - 6,25	6	5,5	2,166666667	13	4,694444444	28,16666667
6,25 - 7,75	2	7	0,666666667	1,333333333	0,444444444	0,888888889
7,75 - 9,25	8	8,5	0,833333333	6,66666667	0,694444444	5,555555556
9,25 - 10,75	13	10	2,333333333	30,33333333	5,444444444	70,77777778
10,75 - 12,25	14	11,5	3,833333333	53,66666667	14,69444444	205,7222222
SImpangan Mutlak Rata-Rata			MDx =	1,968518519		658,9722222
					Varian =	7,404182272
					(Deviasi Standar) sx =	2,72106271

2. Variabel Age

Ukuran Pemusatan					
Mean	Median	Modus	Kuartil 1	Kuartil 2	Kuartil 3
28,6	25,7	18,92105263	18,875	22,73529412	22,60416667

Interval	Frekuensi	Batas Bawah	Batas Atas	Frekuensi Kumulatif	Histogram	Poligon Frekuensi
15 - 21	36	14,5	21,5	36	36	36
22 - 28	15	21,5	29,5	51	15	15
29 - 35	21	29,5	36,5	72	21	21
36 - 42	9	36,5	42,5	81	9	9
43 - 49	0	42,5	49,5	81	0	0
50 - 56	3	49,5	56,5	84	3	3
57 - 63	2	56,5	63,5	86	2	2
64 - 70	4	63,5	69,5	90	4	4



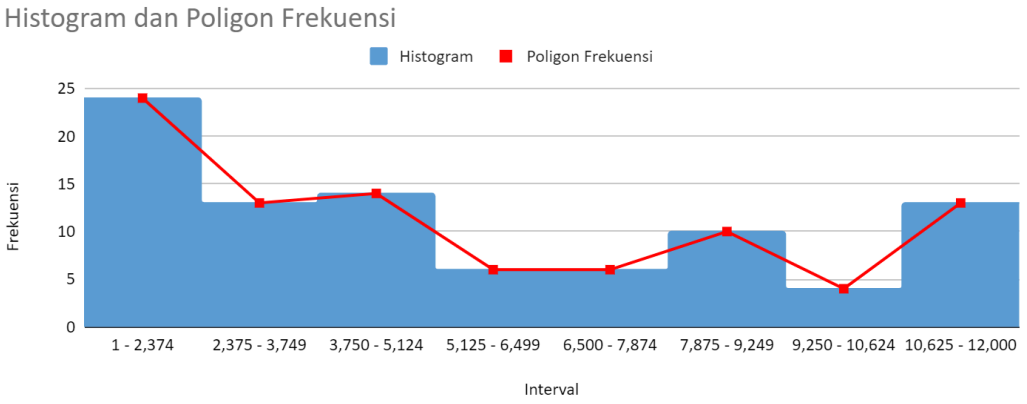
Interval	Frekuensi	Titik Tengah	|x-xmi|	fi(|x-xmi|)	|x-xmi|^2	fi*|x-xmi|^2
15 - 21	25	18	18	450	324	8100
22 - 28	11	25	25	275	625	6875
29 - 35	14	32	32	448	1024	14336
36 - 42	6	39	39	234	1521	9126
43 - 49	0	46	46	0	2116	0
50 - 56	2	53	53	106	2809	5618
57 - 63	1	60	60	60	3600	3600
64 - 70	3	67	67	201	4489	13467
SImpangan Mutlak Rata-Rata			MDx =	19,71111111		61122
					Varian =	686,7640449
					(Deviasi Standar) sx =	26,20618333

3. Variabel Number_0f_Warts

Ukuran Pemusatan

Mean	Median	Modus	Kuartil 1	Kuartil 2	Kuartil 3
5,511111111	4,036	1,443	1,789	4,036	7,994

Interval	Frekuensi	Batas bawah	Batas atas	Frekuensi Kumulatif	Histogram	Poligon Frekuensi	Nilai Tengah
1 - 2,374	24	0,500	1,875	24	24	24	1,687
2,375 - 3,749	13	1,875	3,250	37	13	13	3,062
3,750 - 5,124	14	3,250	4,625	51	14	14	4,437
5,125 - 6,499	6	4,625	6,000	57	6	6	5,812
6,500 - 7,874	6	6,000	7,375	63	6	6	7,187
7,875 - 9,249	10	7,375	8,750	73	10	10	8,562
9,250 - 10,624	4	8,750	10,125	77	4	4	9,937
10,625 - 12,000	13	10,125	12,500	90	13	13	11,3125



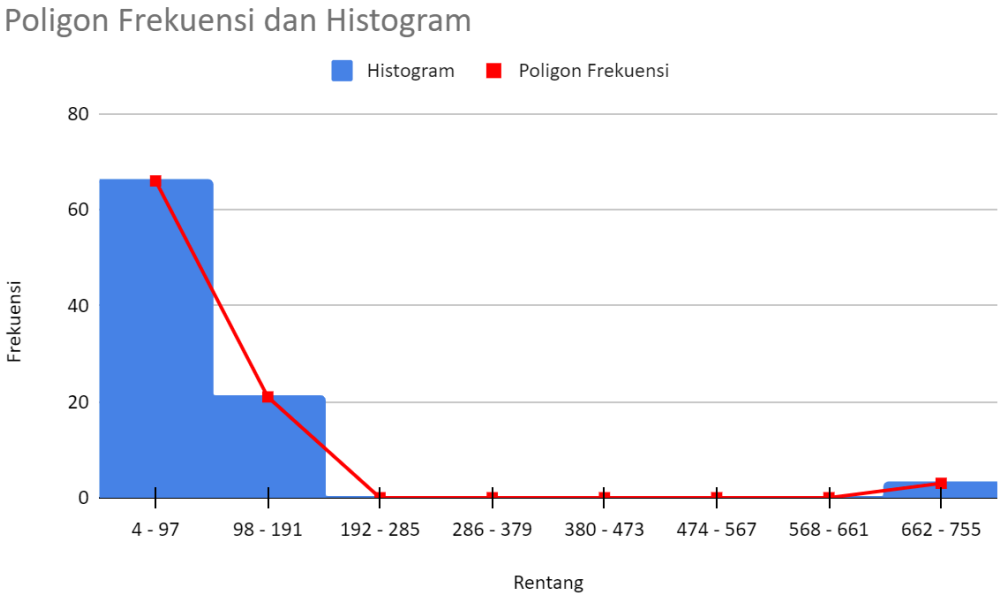
Interval	Frekuensi	Nilai Tengah	|x-xmi|	fi(|x-xmi|)	|x-xmi|^2	fi*|x-xmi|^2
1 - 2,374	10	1,687	3,824111111	38,24111111	14,62382579	146,2382579
2,375 - 3,749	7	3,062	2,449111111	17,14377778	5,998145235	41,98701664
3,750 - 5,124	11	4,437	1,074111111	11,81522222	1,153714679	12,69086147
5,125 - 6,499	6	5,812	0,3008888889	1,805333333	0,090534123 46	0,543204740 7
6,500 - 7,874	5	7,187	1,675888889	8,379444444	2,808603568	14,04301784
7,875 - 9,249	10	8,562	3,050888	30,50888	9,3079230	93,079230

			889	889	12	12
9,250 - 10,624	4	9,937	4,425888889	17,70355556	19,58849246	78,35396983
10,625 - 12,000	11	11,3125	5,801388889	63,81527778	33,65611304	370,2172434
SImpangan Mutlak Rata-Rata			MDx =	2,104584568		757,152802
					Varian =	8,507334854
					(Deviasi Standar) sx =	2,916733593

4. Variabel Area

Ukuran Pemusatan					
Mean	Median	Modus	Kuartil 1	Kuartil 2	Kuartil 3
85,83333333	67,59090909	59,39189189	35,54545455	67,59090909	99,12068966

Interval	Frekuensi	Batas Bawah	Batas Atas	Frekuensi Kumulatif	Histogram	Poligon Frekuensi
4 - 97	66	3,5	97.5	66	66	66
98 - 191	21	97,5	191.5	87	21	21
192 - 285	0	191,5	285.5	87	0	0
286 - 379	0	285,5	379.5	87	0	0
380 - 473	0	379,5	473.5	87	0	0
474 - 567	0	473,5	567.5	87	0	0
568 - 661	0	567,5	661.5	87	0	0
662 - 755	3	661,5	755.5	90	3	3

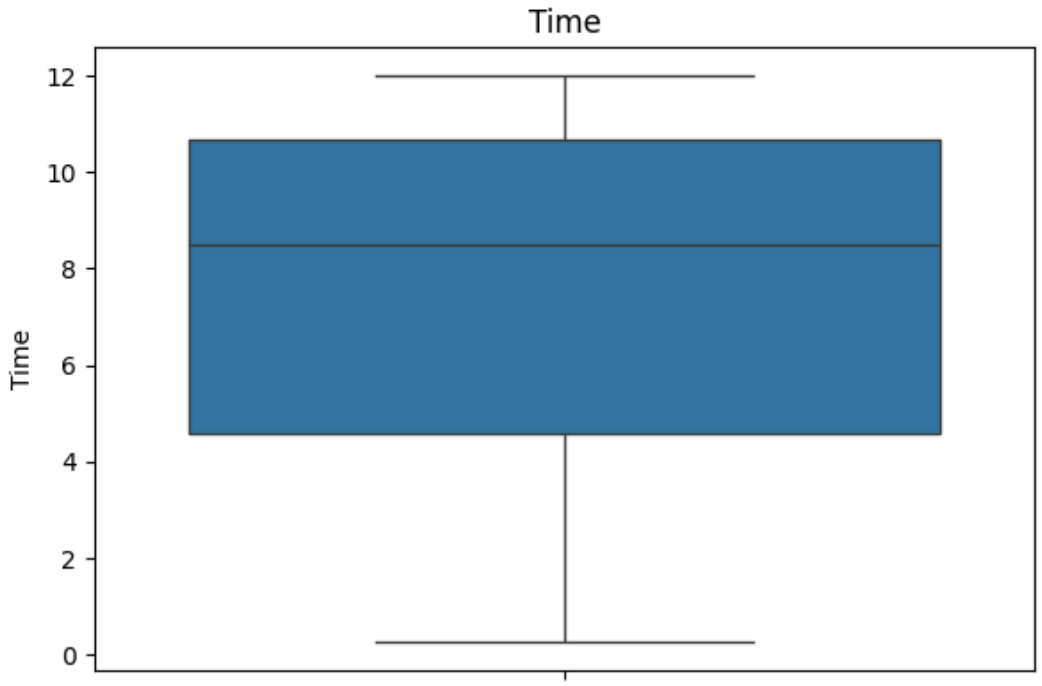


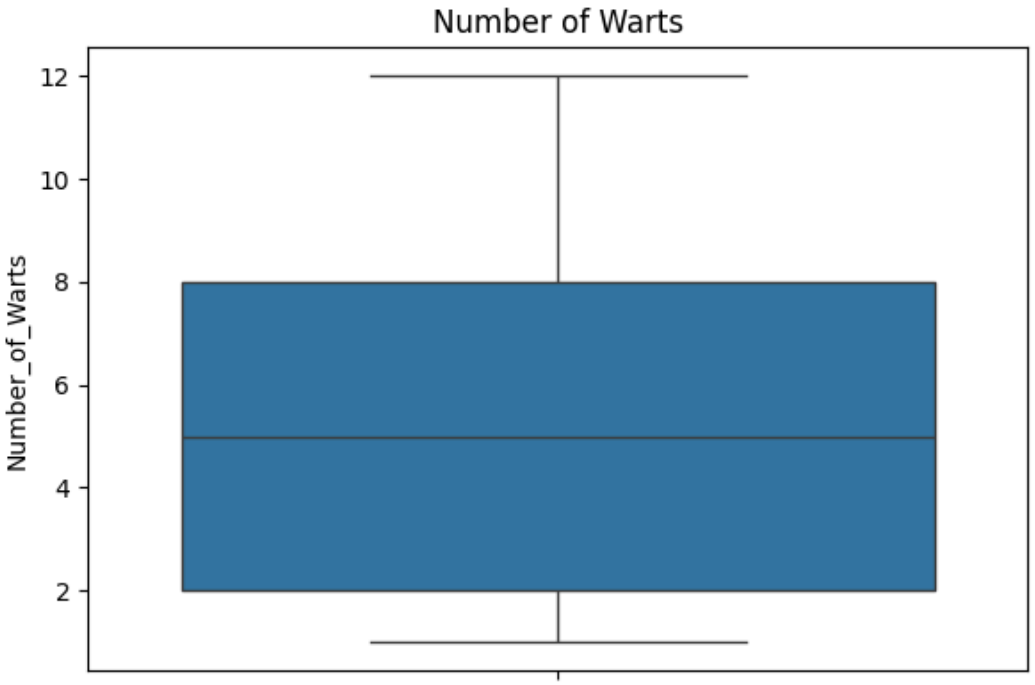
Interval	Frekuensi	Nilai Tengah	|x-xmi|	fi(|x-xmi|)	|x-xmi|^2	fi*|x-xmi|^2
4 - 97	44	50,5	35,33333333	1554,666667	1248,444444	54931,55556
98 - 191	15	144,5	58,66666667	880	3441,777778	51626,66667
192 - 285	0	238,5	152,6666667	0	23307,11111	0
286 - 379	0	332,5	246,6666667	0	60844,44444	0
380 - 473	0	426,5	340,6666667	0	116053,7778	0
474 - 567	0	520,5	434,6666667	0	188935,1111	0
568 - 661	0	614,5	528,6666667	0	279488,4444	0
662 - 755	2	708,5	622,6666667	1245,333333	387713,7778	775427,5556
SImpangan Mutlak Rata-Rata			MDx =	40,88888889		881985,7778
					Varian =	9909,952559
					(Deviasi Standar) sx =	99,54874464

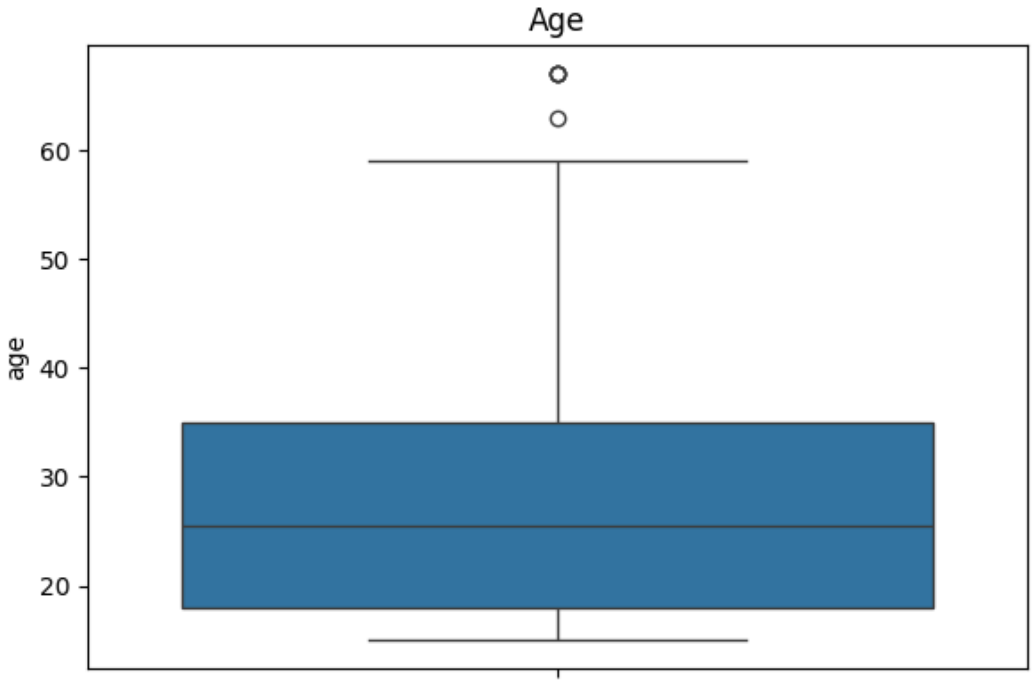
e.

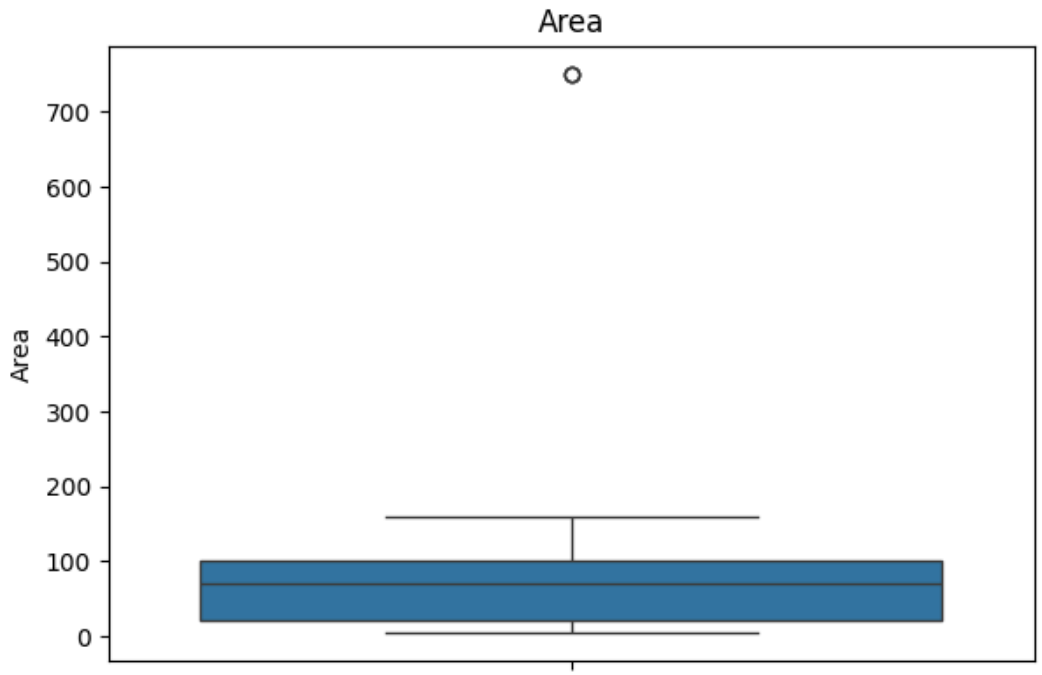
f. Analisis Outlier

Outlier disini dapat membantu kita untuk menganalisa data-data penderita Cryotherapy. Outlier bisa menunjukkan data yang tidak biasa, misalnya, anak-anak atau orang tua yang mungkin memiliki respons yang berbeda terhadap perawatan ini dibandingkan dengan kelompok usia lainnya. Dengan mengidentifikasi outlier ini, kita dapat mengeksplorasi faktor-faktor yang mempengaruhi respons individu terhadap Cryotherapy.




Number of Warts
Number_of_Warts
12
10
8
6
4
2


Age
age
60
50
40
30
20

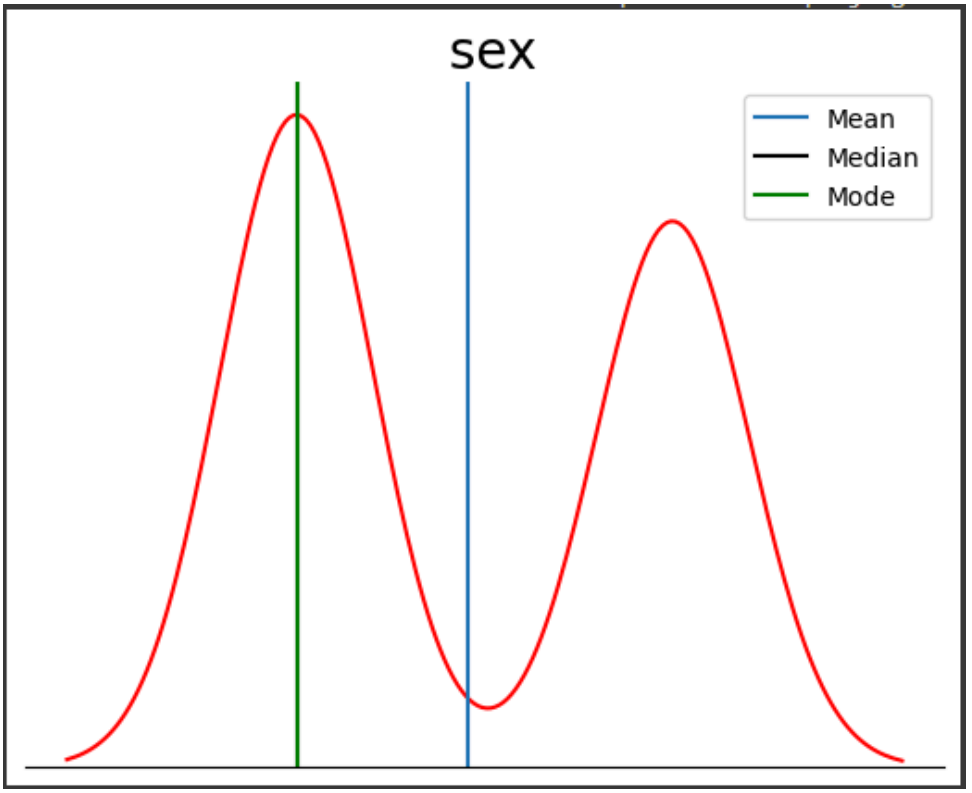


Data yang dibutuhkan untuk membuat outlier adalah Kuartil 1 (Q1), Kuartil 3 (Q3), Nilai Maximum, dan Nilai Minimum. Quartil 1 dan 3 digunakan untuk mencari nilai dimana sebagian besar data berada, dan nilai max, min digunakan untuk menjadikan batas atas dan bawah.Dengan informasi ini, kita dapat menetapkan batas untuk menentukan apakah suatu nilai dianggap sebagai outlier atau tidak.

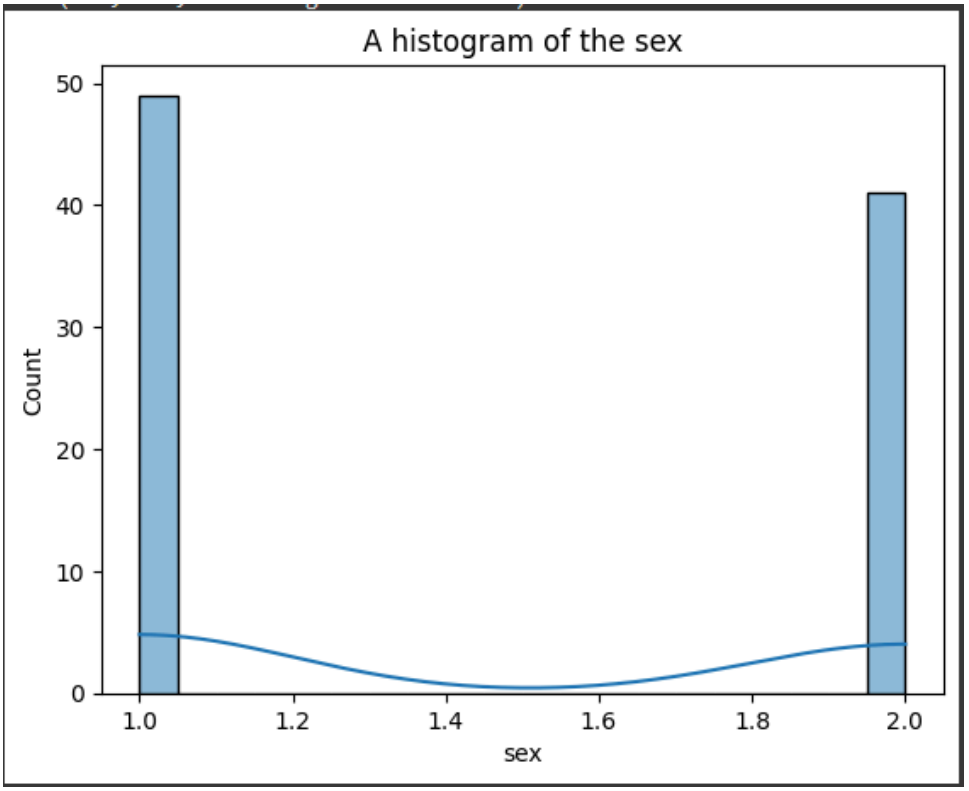
g. Skewness dan Kurtosis

Skewness adalah sebuah konsep dalam statistika yang mengukur seberapa tidak simetris distribusi data. Distribusi dikatakan memiliki skewness positif jika ekornya lebih panjang di sebelah kanan mean, negatif jika lebih panjang di sebelah kiri, dan nol jika simetris. Skewness membantu dalam memahami pola distribusi data serta memberikan informasi tentang kecenderungan data dalam menyebar di sekitar nilai rata-rata.

- Sex

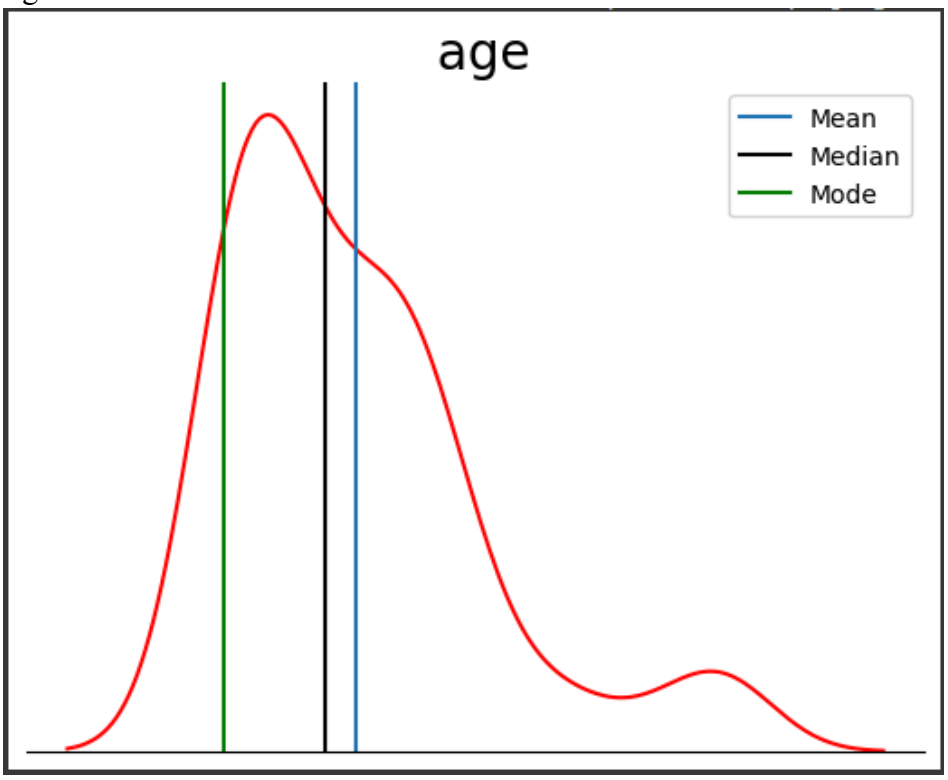


Skewness : -2.35763

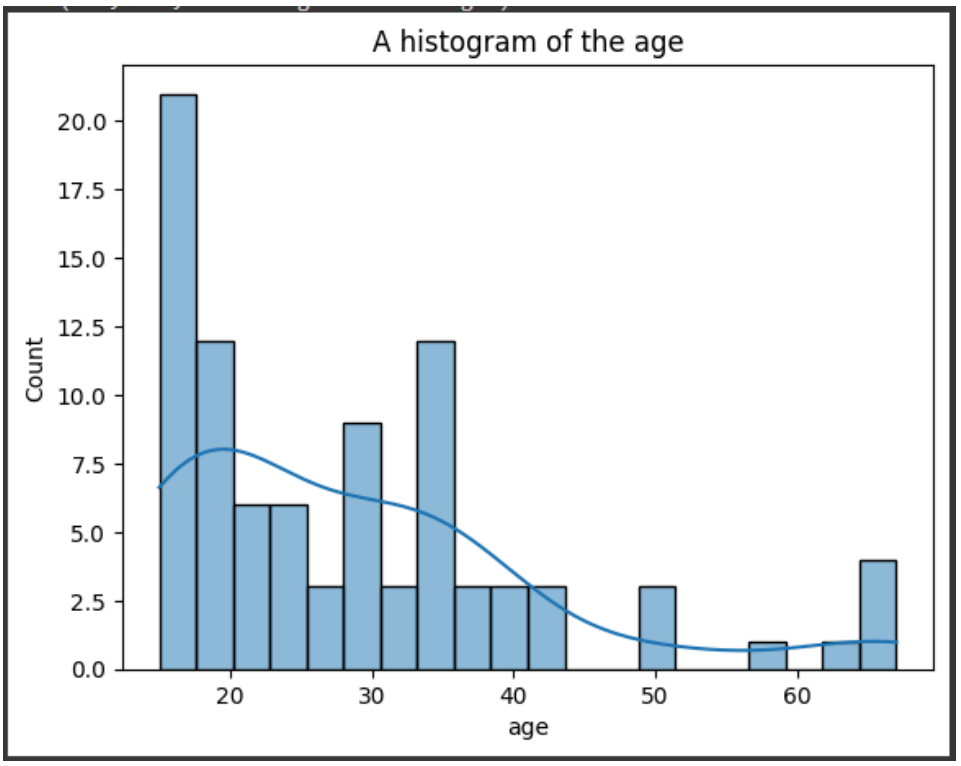


Kurtosis : -2.0122770417135425

- Age

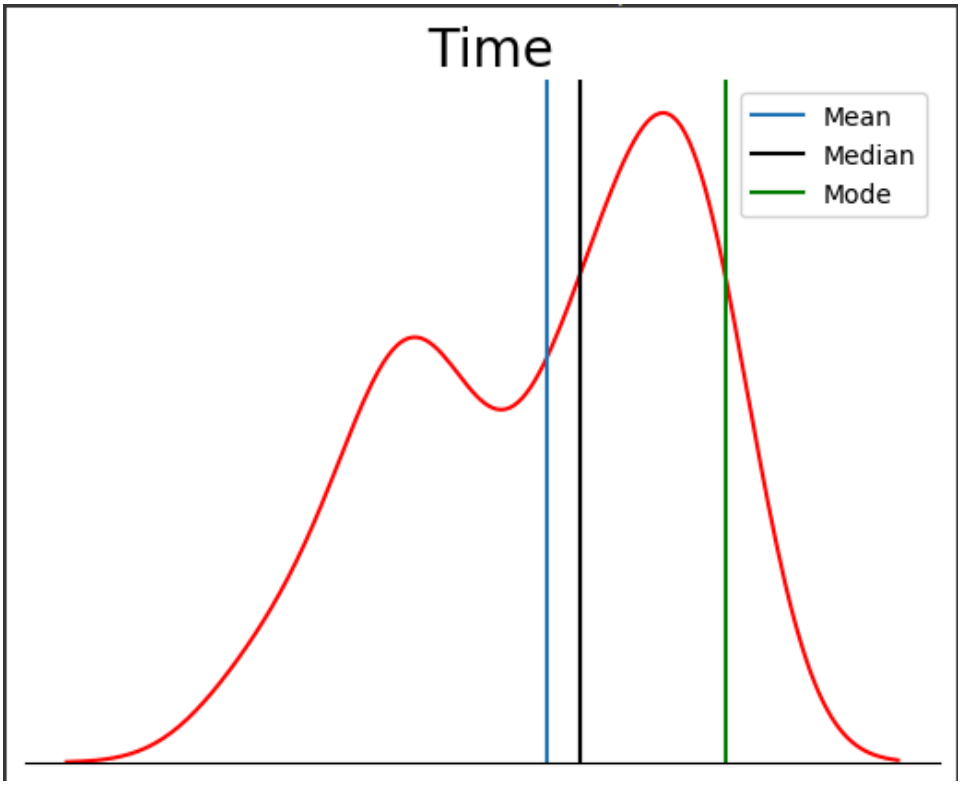


Skewness : -2.35763

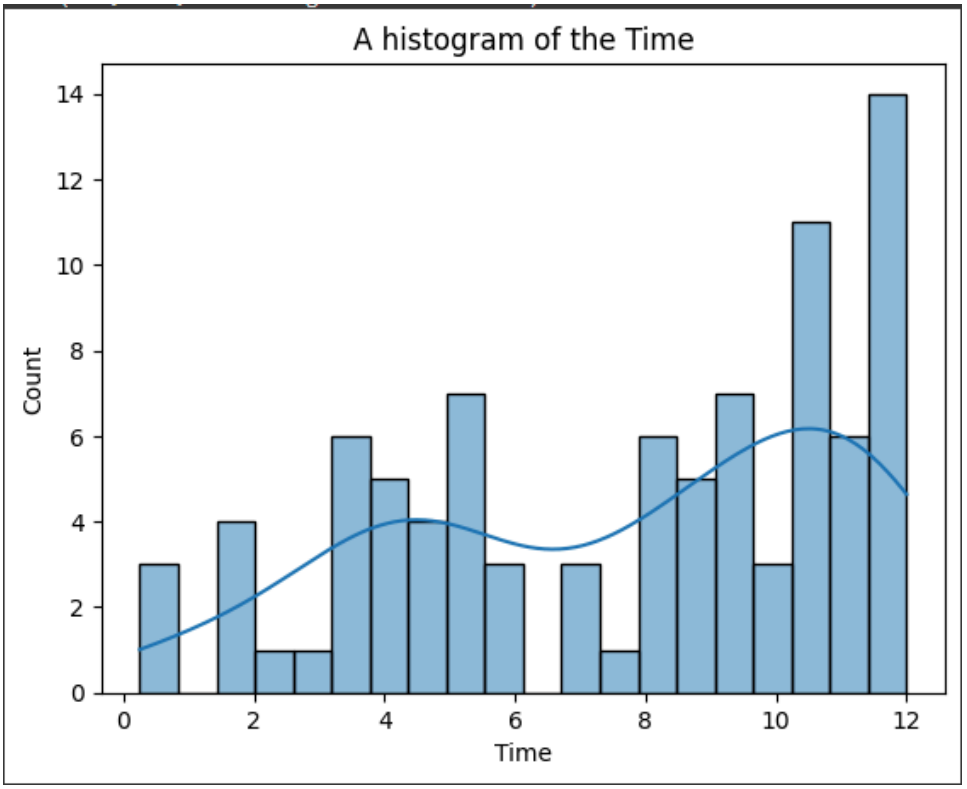


Kurtosis : 1.5876126152298173

- Time

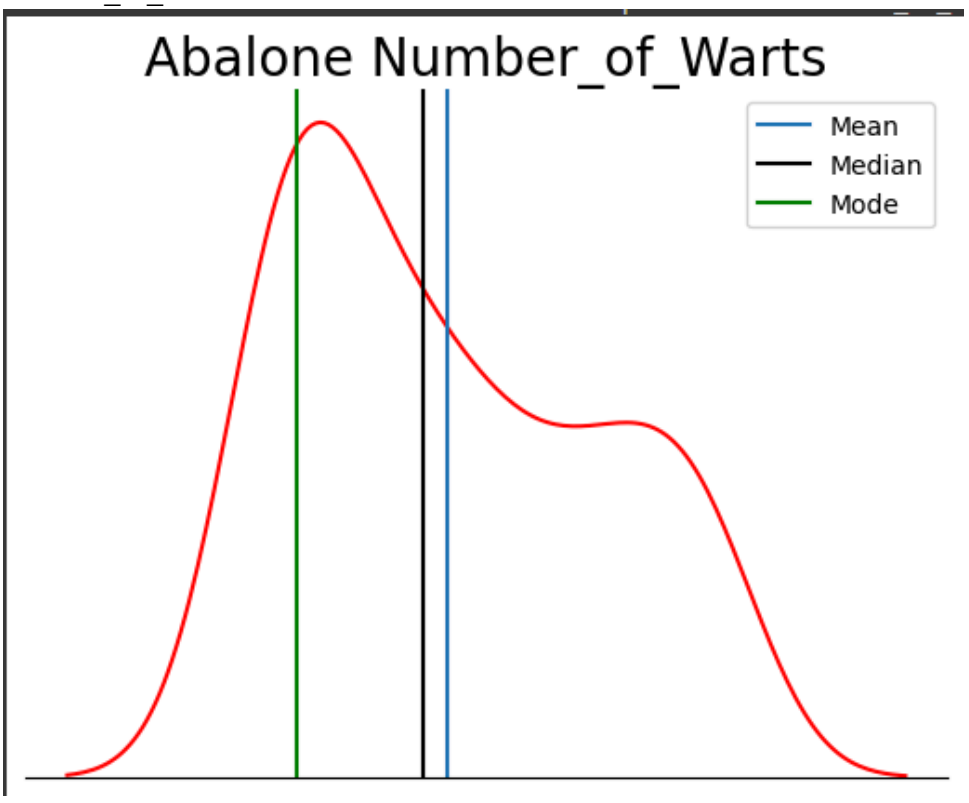


Skewness : -0.70360

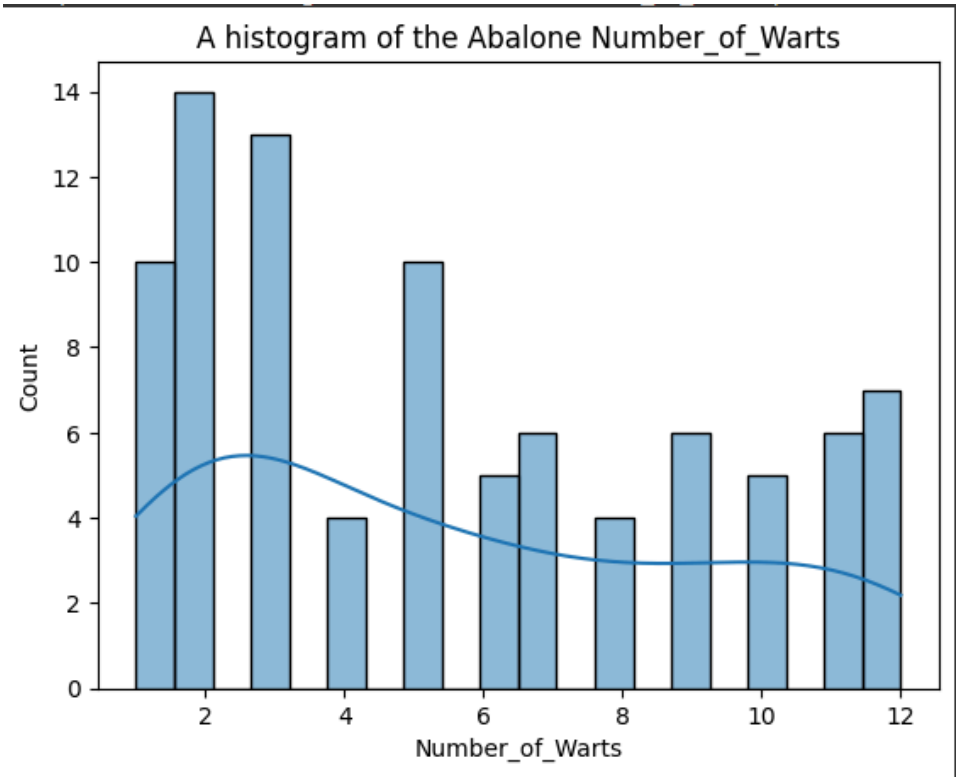


Kurtosis : -1.0410479583634418

- Number_of_Warts

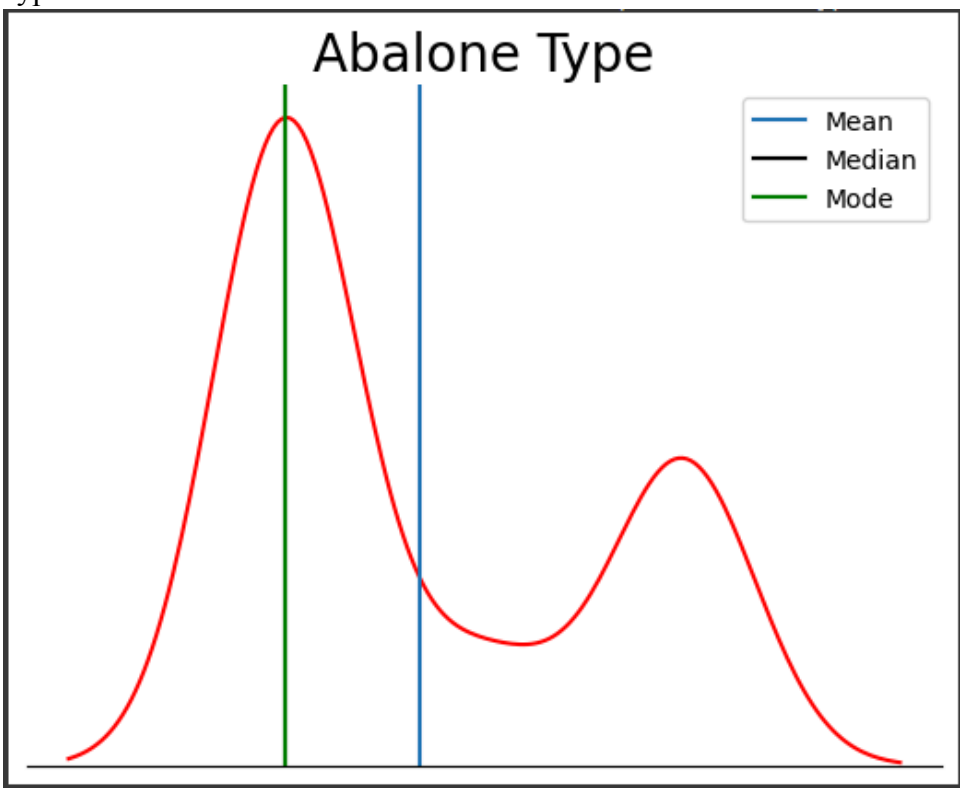


Skewness : 0.47127

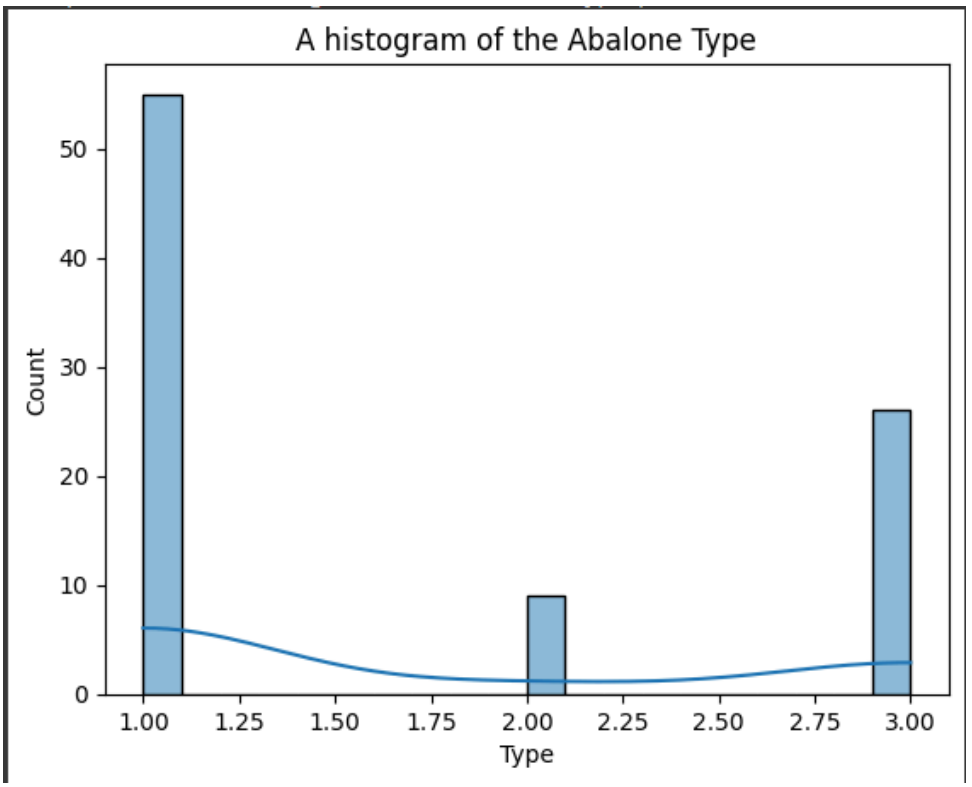


Kurtosis : -1.1620206713899788

- Type

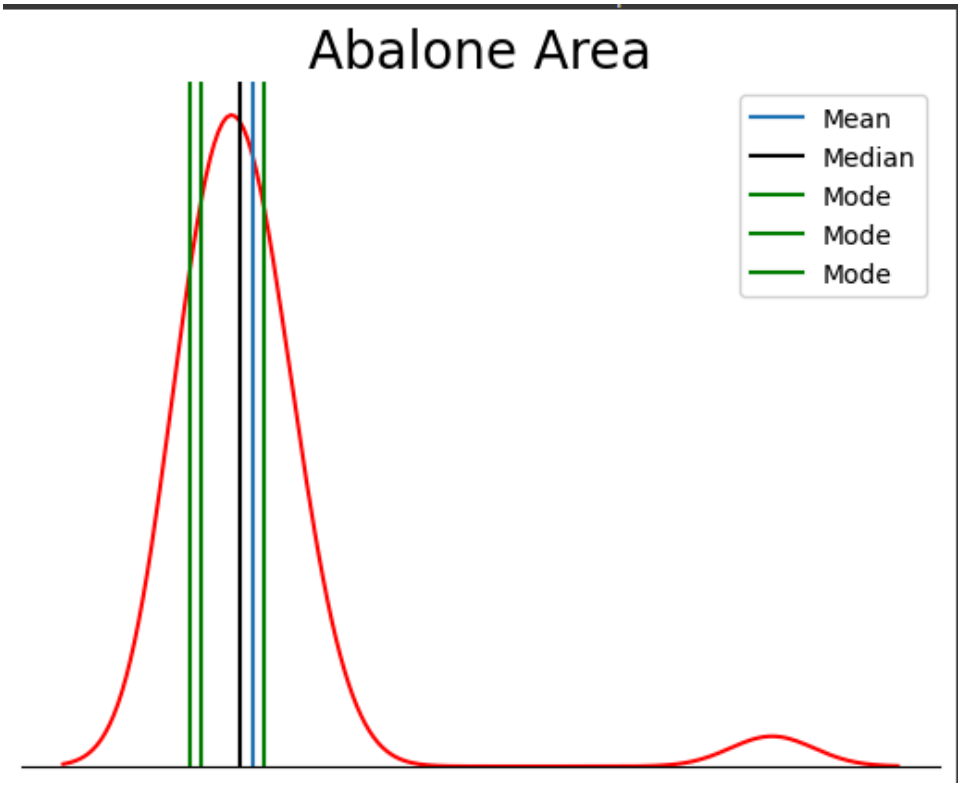


Skewness : 2.26610

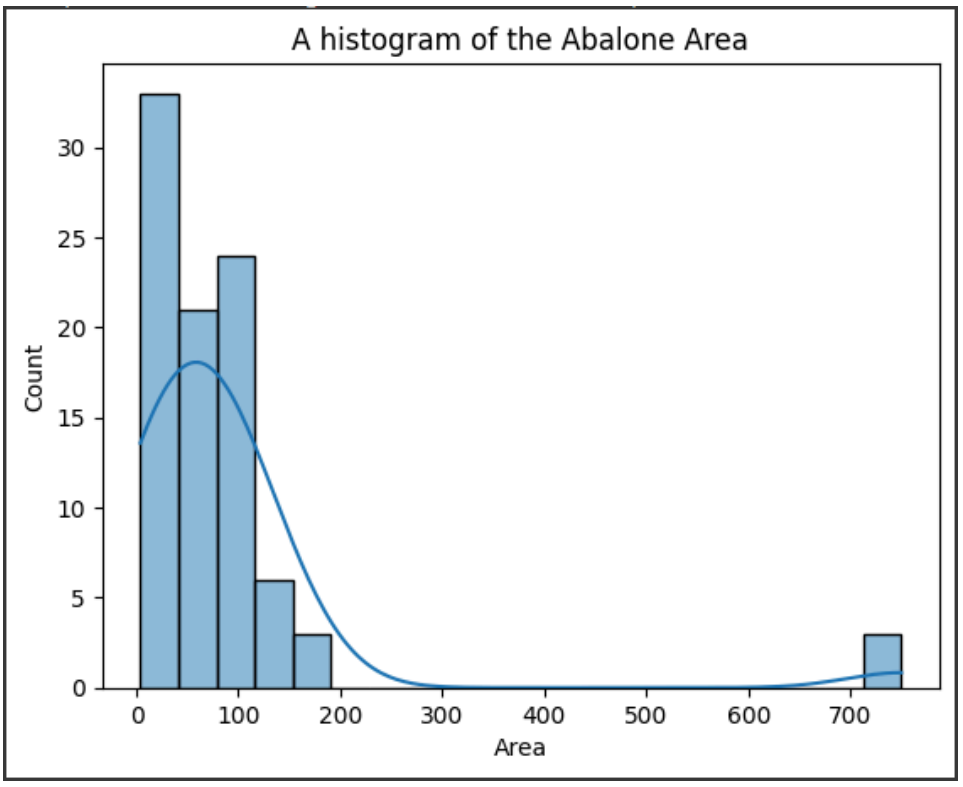


Kurtosis : -1.4132752985061114

- Area

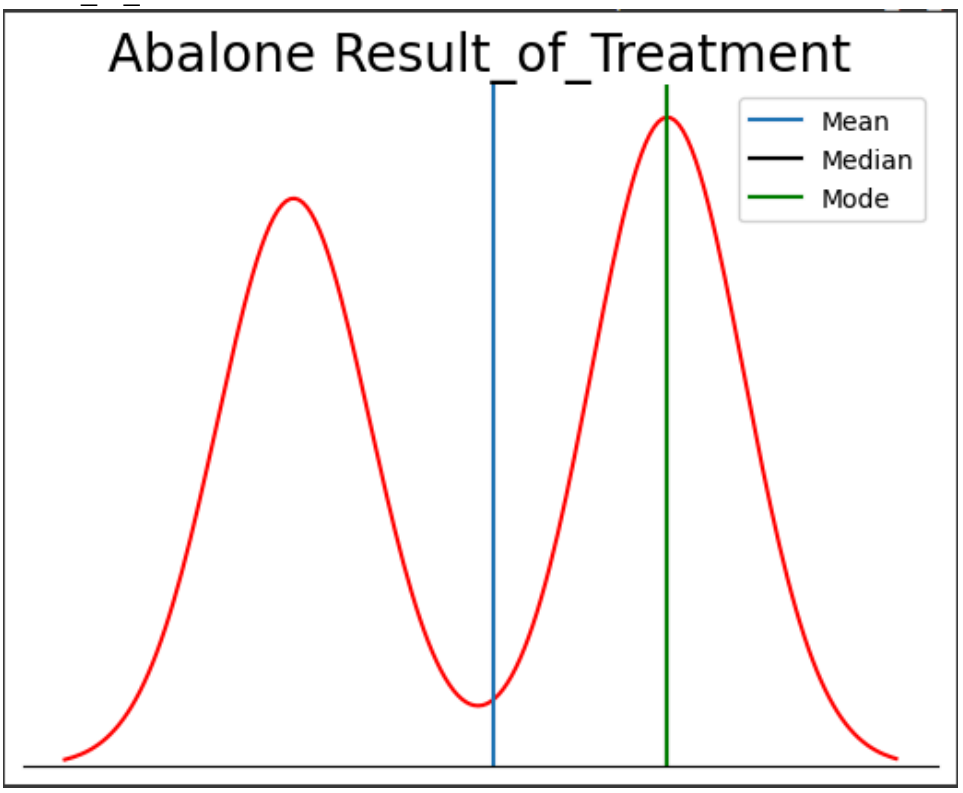


Skewness : 0.36058

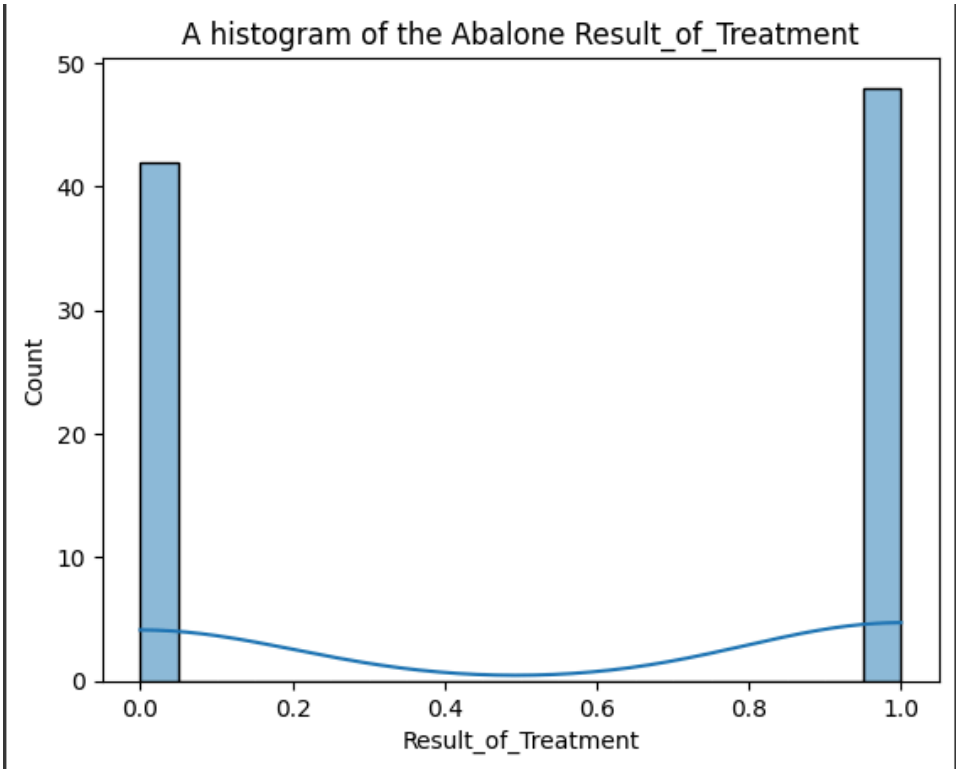


Kurtosis : 20.239244004409066/

- Result_of_Treatment



Skewness : -2.79061



Kurtosis : -2.0270865987460818

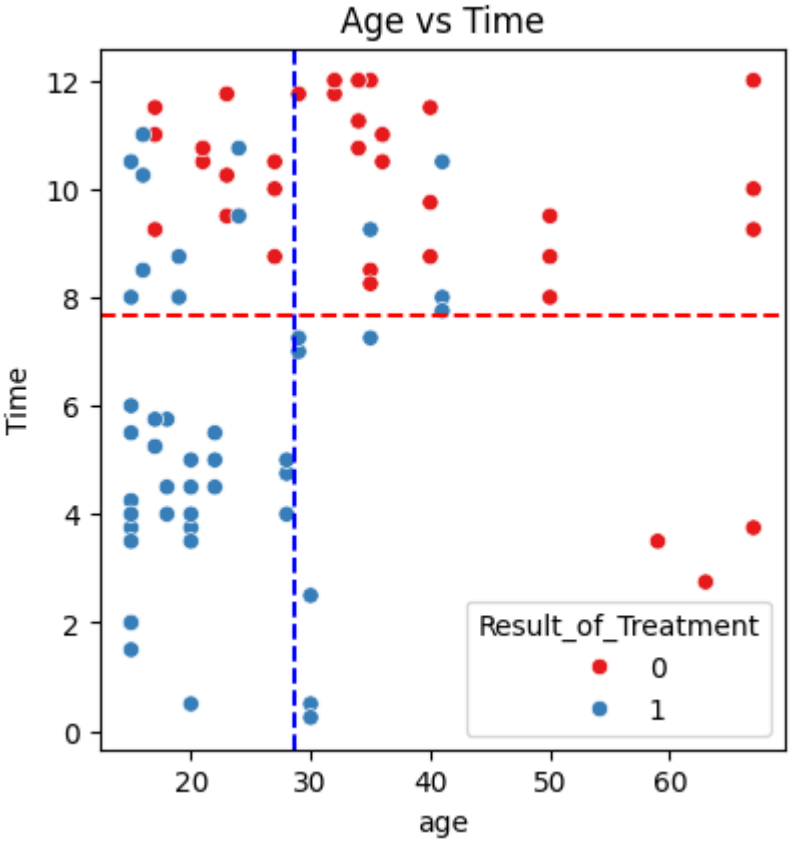
berikut untuk link google Colab :
https://colab.research.google.com/drive/1BLX22Hk7WxJmMs5mWWDJicwNX3sYr8PC?usp=sharing

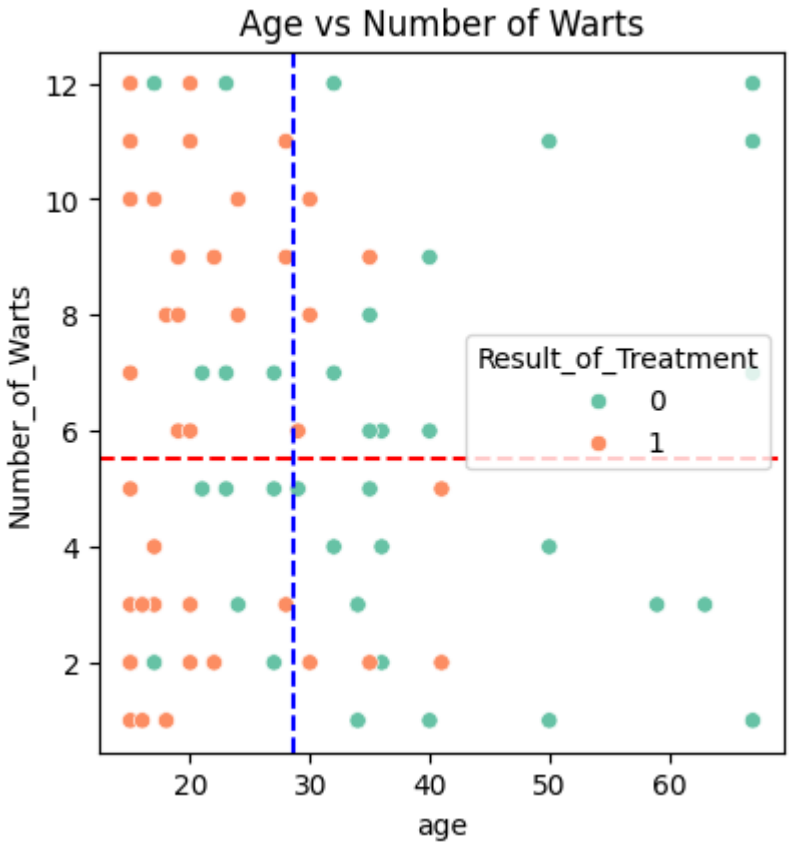
```
[ ]  abalone_sex = df['sex']
     sns.kdeplot(abalone_sex, color="red")

     sns.despine(top=True, right=True, left=True)
     plt.xticks([])
     plt.yticks([])
     plt.ylabel("")
     plt.xlabel("")
     plt.title("Abalone sex", fontdict=dict(fontsize=20))

     # Find the mean, median, mode
     mean_rings = df["sex"].mean()
     median_sex = df["sex"].median()
     mode_sex = df["sex"].mode().iloc[0]
     std = abalone_sex.std()

     skewness = (3 * (mean_sex - median_sex)) / std

     print(
        f"Skor dari Pierson's second skewness terhadap distribusi panjang abalone adalah: {skewness:.5f}")

     # Add vertical lines at the position of mean, median, mode
     plt.axvline(mean_sex, label="Mean")
     plt.axvline(median_sex, color="black", label="Median")
     plt.axvline(mode_sex, color="green", label="Mode")

     plt.legend();
```

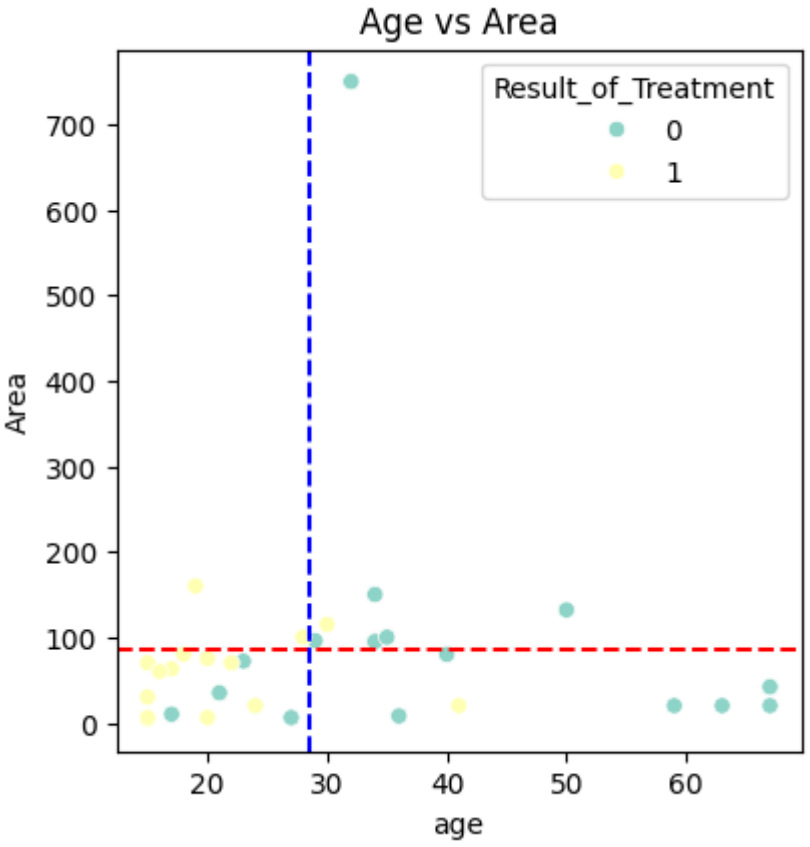
Visualisasi dimulai dengan kernel density estimation (KDE) plot yang menunjukkan distribusi frekuensi dengan warna merah. Grafik tersebut kemudian didekorasi dengan menghilangkan batas-batas (spines) dan label sumbu, serta mengatur judul. Selanjutnya, ditambahkan nilai statistik seperti mean, median, mode, dan skewness. Skewness dihitung menggunakan rumus Pierson's second skewness untuk mengevaluasi ketidaksimetrisan distribusi. Tiga garis vertikal ditambahkan pada posisi mean, median, dan mode pada plot untuk memberikan representasi visual dari nilai-nilai tersebut. Dengan demikian, kode ini secara

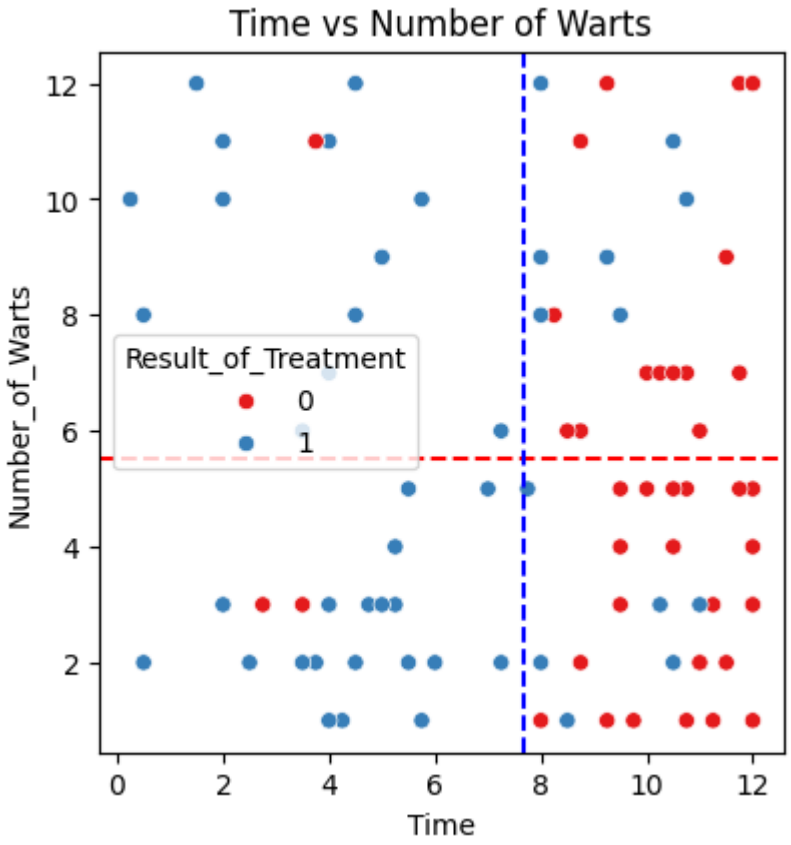
komprehensif menggambarkan distribusi dan statistik penting terkait dalam dataset.

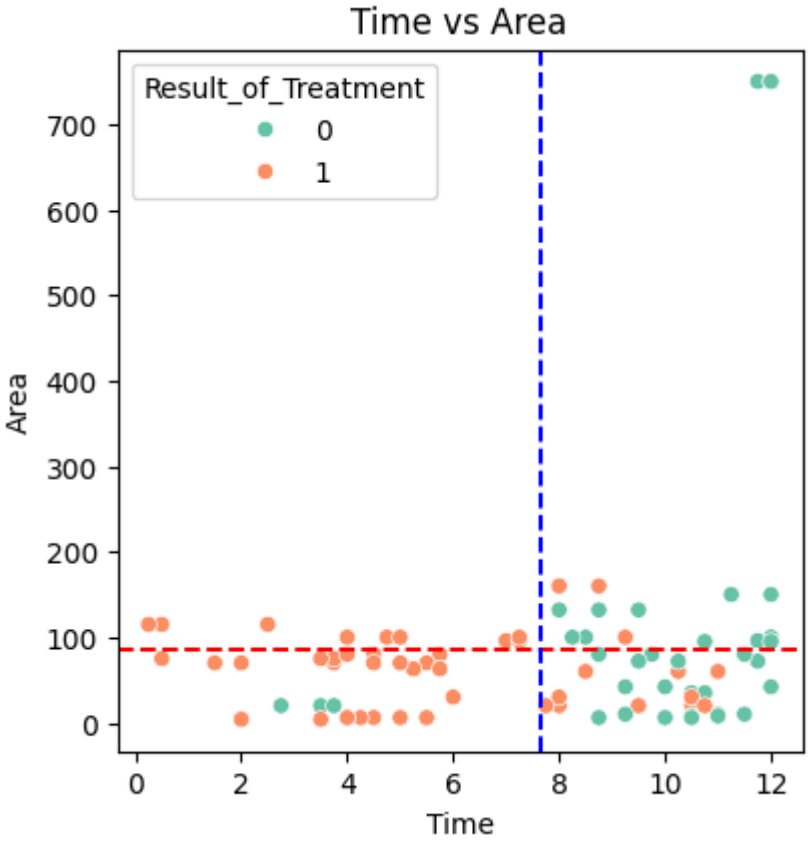
h. Visualisasi Scatterplot

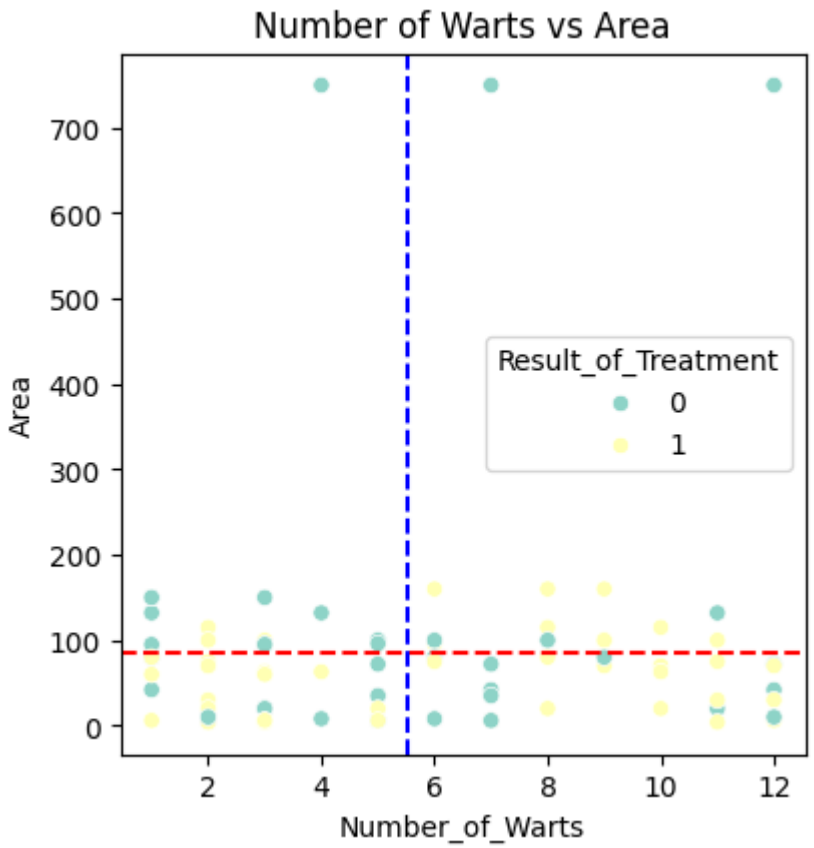



Age vs Number of Warts
Number_of_Warts
age
Result_of_Treatment
0
1
12
10
8
6
4
2
20
30
40
50
60


Age vs Area
Area
age
Result_of_Treatment
0
1
700
600
500
400
300
200
100
0
20
30
40
50
60


Time vs Number of Warts
Number_of_Warts
Time
Result_of_Treatment
0
1
12
10
8
6
4
2
0
2
4
6
8
10
12


Time vs Area
Area
Time
Result_of_Treatment
0
1
700
600
500
400
300
200
100
0
0
2
4
6
8
10
12



https://colab.research.google.com/drive/1400YYexP91MQyQ6XlKjFR8-B2txU--Un?usp=sharing

Age vs Number_of_Warts: Scatter plot ini akan memperlihatkan apakah ada korelasi antara usia pasien dan jumlah kutil yang dimilikinya. Jika ada korelasi positif, maka dapat diasumsikan bahwa semakin tua seseorang, semakin besar kemungkinan dia memiliki lebih banyak kutil.

Time vs Area: Scatter plot ini akan menunjukkan apakah ada hubungan antara waktu (dalam bulan) dan luas area kutil. Dengan visualisasi ini, kita dapat melihat apakah kutil cenderung membesar atau mengecil seiring berjalannya waktu.

Number_of_Warts vs Area: Scatter plot ini akan menunjukkan hubungan antara jumlah kutil dan luas area kutil. Ini dapat memberikan wawasan tentang apakah jumlah kutil berhubungan dengan ukuran kutil.

Age vs Time: Scatter plot ini akan menunjukkan apakah ada hubungan antara usia pasien dan waktu pengobatan. Ini dapat membantu dalam melihat pola umum pengobatan berdasarkan usia pasien.

Age vs Area: Scatter plot ini akan menunjukkan apakah ada hubungan antara usia pasien dan luas area kutil. Ini dapat membantu kita melihat apakah kutil cenderung lebih besar pada pasien tertentu berdasarkan usia mereka.

Dengan memvisualisasikan kombinasi variabel ini, kita dapat mendapatkan pemahaman yang lebih baik tentang hubungan antara berbagai faktor dalam dataset tersebut, yang pada gilirannya dapat membantu dalam analisis lebih lanjut atau pengambilan keputusan.

Berdasarkan kombinasi variabel yang dapat divisualisasikan dari data yang diberikan, kita dapat mencapai beberapa kesimpulan atau observasi:

1. Usia dan Jumlah Kutil: Tidak terlihat pola jelas antara usia pasien dan jumlah kutil yang dimilikinya. Ini menunjukkan bahwa faktor usia mungkin tidak berpengaruh signifikan terhadap jumlah kutil yang dimiliki.

2. Waktu dan Luas Area Kutil: Scatter plot antara waktu dan luas area kutil menunjukkan bahwa ada variasi dalam luas area kutil sepanjang waktu pengamatan. Namun, tidak ada tren yang jelas menunjukkan peningkatan atau penurunan luas area kutil seiring berjalannya waktu.

3. Jumlah Kutil dan Luas Area Kutil: Terlihat adanya hubungan positif antara jumlah kutil dan luas area kutil. Ini menunjukkan bahwa semakin banyak kutil yang dimiliki seseorang, semakin besar luas area kutil tersebut.

4. Usia dan Waktu Pengobatan: Scatter plot antara usia pasien dan waktu pengobatan tidak menunjukkan pola yang jelas. Ini menunjukkan bahwa waktu pengobatan tidak secara signifikan bergantung pada usia pasien.

5. Usia dan Luas Area Kutil: Tidak terlihat pola yang jelas antara usia pasien dan luas area kutil yang dimilikinya. Ini menunjukkan bahwa faktor usia mungkin tidak secara signifikan mempengaruhi ukuran kutil.

Kesimpulannya, visualisasi data memberikan gambaran yang lebih jelas tentang hubungan antara berbagai variabel dalam dataset, meskipun tidak selalu ada pola yang mudah diamati atau korelasi yang kuat antara variabel-variabel tertentu. Analisis lebih lanjut atau metode statistik mungkin diperlukan untuk memahami lebih lanjut pola-pola yang mungkin ada dalam data tersebut.

i. Klasifikasi dengan Logistic Regression

Tujuan utama menggunakan Logistic Regression pada data Cryotherapy adalah untuk memprediksi hasil perawatan (berhasil atau tidak) berdasarkan beberapa variabel input seperti umur, waktu perawatan, jumlah kutil, dan area yang terkena.

Nilai korelasi antar variabel:

	sex	age	Time	Number_of_Warts	Type	Area	Result_of_Treatment
sex	1.000000	-0.115185	0.074417	0.018952	0.219970	0.091213	-0.086203
age	-0.115185	1.000000	0.236305	-0.034797	0.415536	0.080915	-0.542780
Time	0.074417	0.236305	1.000000	-0.074354	0.235056	0.241559	-0.654147
Number_of_Warts	0.018952	-0.034797	-0.074354	1.000000	0.002784	0.108762	0.078273
Type	0.219970	0.415536	0.235056	0.002784	1.000000	0.354398	-0.485030
Area	0.091213	0.080915	0.241559	0.108762	0.354398	1.000000	-0.188886
Result_of_Treatment	-0.086203	-0.542780	-0.654147	0.078273	-0.485030	-0.188886	1.000000

Interpretasi Korelasi

- Korelasi Positif: Nilai positif menunjukkan bahwa kedua variabel bergerak ke arah yang sama.
- Korelasi Negatif: Nilai negatif menunjukkan bahwa kedua variabel bergerak ke arah yang berlawanan.
- Kekuatan Korelasi: Semakin mendekati 1 atau -1, semakin kuat korelasinya. Nilai mendekati 0 menunjukkan tidak ada hubungan linear yang kuat.

Korelasi dengan Result_of_Treatment

- age: Korelasi negatif kuat (-0.542780). Ini berarti bahwa semakin tua pasien, semakin kecil kemungkinan perawatan akan berhasil.
- Time: Korelasi negatif sangat kuat (-0.654147). Ini menunjukkan bahwa semakin lama waktu perawatan, semakin kecil kemungkinan perawatan akan berhasil.
- Number_of_Warts: Korelasi positif sangat lemah (0.078273). Ini menunjukkan hampir tidak ada hubungan antara jumlah kutil dan hasil perawatan.
- Type: Korelasi negatif moderat (-0.485030). Ini menunjukkan bahwa tipe kutil memiliki pengaruh yang cukup signifikan terhadap keberhasilan perawatan.
- Area: Korelasi negatif lemah (-0.188886). Ini menunjukkan bahwa area yang terkena memiliki sedikit pengaruh terhadap keberhasilan perawatan.

- sex: Korelasi negatif sangat lemah (-0.086203). Ini menunjukkan hampir tidak ada hubungan antara jenis kelamin dan hasil perawatan.

Korelasi antar Variabel Lain

- age dan Type: Korelasi positif moderat (0.415536). Ini menunjukkan bahwa ada hubungan yang cukup signifikan antara usia pasien dan tipe kutil.
- age dan Time: Korelasi positif lemah (0.236305). Ini menunjukkan bahwa semakin tua pasien, semakin lama waktu perawatan yang mungkin dibutuhkan.
- Time dan Area: Korelasi positif lemah (0.241559). Ini menunjukkan bahwa semakin lama waktu perawatan, area yang terkena cenderung lebih besar.
- Type dan Area: Korelasi positif lemah (0.354398). Ini menunjukkan bahwa tipe kutil cenderung berhubungan dengan area yang terkena.
- sex dan Type: Korelasi positif lemah (0.219970). Ini menunjukkan bahwa ada hubungan lemah antara jenis kelamin pasien dan tipe kutil.

Klasifikasi

Pengklasifikasian data dilakukan dengan bantuan pemrograman python, dimana data dari dataset dipisahkan terlebih dahulu antara fitur dan target. Dalam dataset Cryotherapy, fitur yang digunakan dalam model ini adalah variabel age, Time, Number_of_Warts, dan Area. Dan target yang diprediksi adalah variabel Result_of_Treatment yang memiliki nilai 0 dan 1.

Kemudian dataset dibagi menjadi dua bagian: data latih dan data uji. Model Logistic Regression dilatih menggunakan data latih dengan maksimum iterasi sebanyak 1000 kali agar mencapai konvergensi. Setelah model dilatih, dilakukan prediksi label target dari data uji.

Sehingga laporan klasifikasi yang meliputi metrik evaluasi seperti precision, recall, f1-score, dan support untuk setiap kelas dalam data uji ditampilkan seperti di bawah ini.

Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.82	1.00	0.90	9
1	1.00	0.78	0.88	9
Accuracy			0.89	18
Macro avg	0.91	0.89	0.89	18
Weighted avg	0.91	0.89	0.89	18

Accuracy: 0.8888888888888888

Interpretasi *Logistic Regression Classification Report:*

- Precision adalah proporsi prediksi positif yang benar. Untuk kelas 0, 82% dari prediksi yang mengatakan "0" benar-benar "0". Untuk kelas 1, semua prediksi yang mengatakan "1" benar-benar "1".
  $Precision = \frac{TP}{TP+FP}$
- Recall adalah proporsi kasus positif yang benar-benar dideteksi. Untuk kelas 0, semua kasus "0" terdeteksi. Untuk kelas 1, 78% dari semua kasus "1" terdeteksi.
  $Recall = \frac{TP}{TP+FN}$
- F1-Score adalah harmonic mean dari precision dan recall, memberikan keseimbangan antara keduanya. Ini memberikan gambaran seberapa baik model menangani trade-off antara precision dan recall. Dan support adalah jumlah kasus sebenarnya dari setiap kelas dalam data uji.
  $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$
- Dengan menggunakan Logistic Regression, didapatkan akurasi atau total prediksi dari keseluruhan model sebesar 0.8888, sehingga menunjukkan bahwa model melakukan prediksi yang benar pada sebagian besar data uji.
  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Model memiliki akurasi keseluruhan sebesar 88.89%. Ini berarti bahwa sekitar 89% dari prediksi model sesuai dengan label sebenarnya dalam data uji. Akurasi yang tinggi ini menunjukkan bahwa model bekerja dengan baik dalam mengklasifikasikan hasil perawatan.

j. Kelas 0 (Hasil Perawatan Negatif):

- **Precision**: 0.82

82% dari prediksi yang menyatakan hasil perawatan negatif benar-benar negatif.

- **Recall**: 1.00

100% dari semua kasus negatif terdeteksi dengan benar.

- **F1-Score**: 0.90

Kombinasi dari precision dan recall, menunjukkan keseimbangan yang baik antara keduanya.

k. Kelas 1 (Hasil Perawatan Positif):

- **Precision**: 1.00

100% dari prediksi yang menyatakan hasil perawatan positif benar-benar positif.

- **Recall**: 0.78

78% dari semua kasus positif terdeteksi dengan benar.

- **F1-Score**: 0.88

Kombinasi dari precision dan recall, menunjukkan keseimbangan yang cukup baik antara keduanya.

l. Nilai Rata-Rata

Macro Avg:

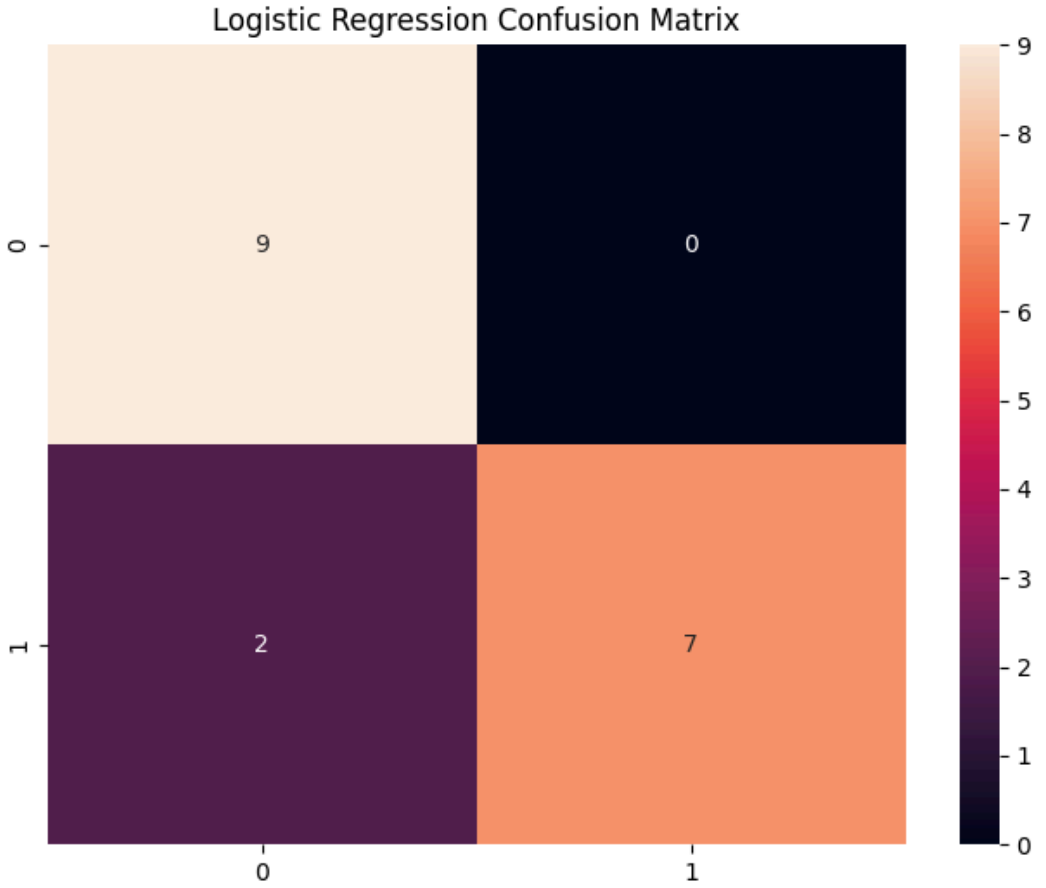
- Precision: 0.91
- Recall: 0.89
- F1-Score: 0.89
- Rata-rata ini menghitung metrik untuk setiap kelas dan kemudian menghitung rata-rata tanpa mempertimbangkan proporsi kelas.

Weighted Avg:

- Precision: 0.91
- Recall: 0.89
- F1-Score: 0.89
- Rata-rata ini menghitung metrik dengan mempertimbangkan proporsi masing-masing kelas dalam dataset.

m. Confusion Matrix

Program membuat serta menampilkan *Confusion Matrix* untuk mengevaluasi kinerja model lebih lanjut.

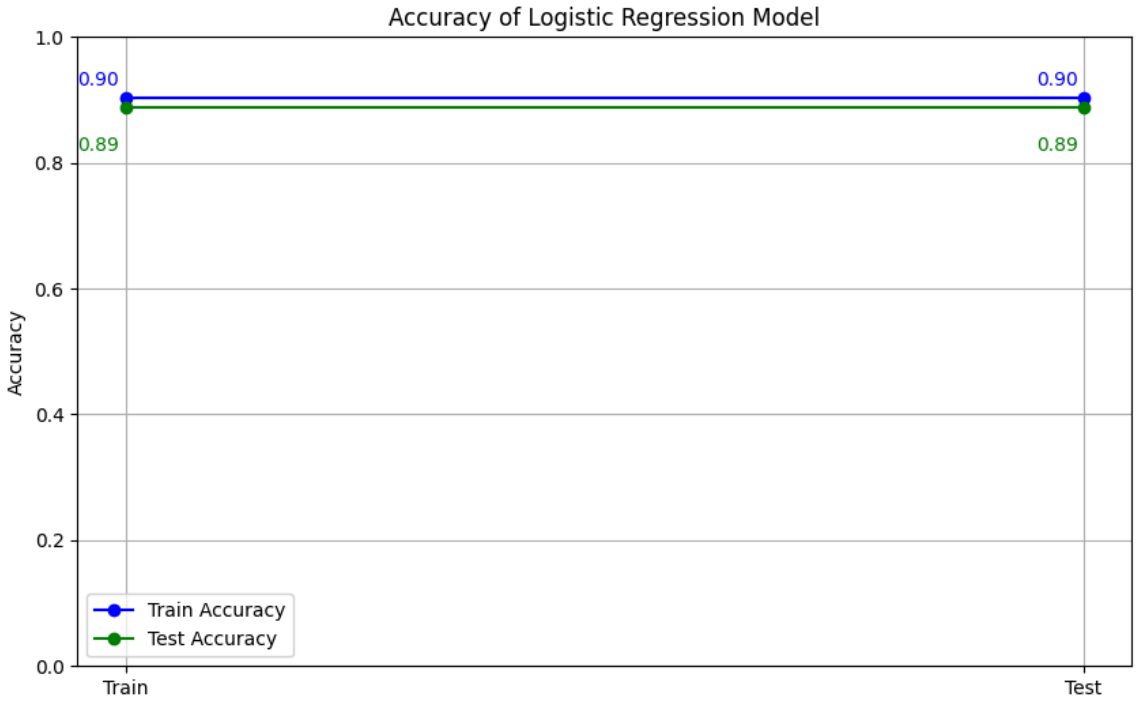


Penjelasan Confusion Matrix

- True Positives (TP):
  Prediksi 1, Sebenarnya 1: 7 (model memprediksi 1 dan benar-benar 1). Ada 7 sampel yang diprediksi sebagai positif dan benar-benar positif.
- True Negatives (TN):
  Prediksi 0, Sebenarnya 0: 9 (model memprediksi 0 dan benar-benar 0). Ada 9 sampel yang diprediksi sebagai negatif dan benar-benar negatif.
- False Positives (FP):
  Prediksi 1, Sebenarnya 0: 0 (model memprediksi 1 tapi sebenarnya 0). Tidak ada sampel yang diprediksi sebagai positif padahal sebenarnya negatif.
- False Negatives (FN):
  Prediksi 0, Sebenarnya 1: 2 (model memprediksi 0 tapi sebenarnya 1). Ada 2 sampel yang diprediksi sebagai negatif padahal sebenarnya positif.

FP dan FN: Memberikan informasi tentang jumlah kesalahan dan jenis kesalahan. Dalam kasus ini, ada 2 FN (model salah memprediksi kelas 0 ketika seharusnya kelas 1). TP dan TN: Memberikan informasi tentang jumlah prediksi benar.

Grafik Akurasi Logistic Regression:



Dalam grafik di atas, bisa dilihat hasil akurasi dari penggunaan *Logistic Regression* adalah 0,89 atau 89%. Sedangkan untuk akurasi dari data latih adalah sebesar 0.90 atau 90%. Yang berarti nilai akurasi dari *Logistic Regression* ini 1% lebih rendah dari data latihnya.

https://colab.research.google.com/drive/1AnBleyKknM9ivx1XwQdP-CtBnwq2Hll5?usp=sharing