

# MINI PORTOFOLIO

Breast Cancer Classification

Presented by Fina Putri



# PENDAHULUAN

Kanker payudara adalah kanker yang paling umum terjadi pada wanita di dunia. Kanker ini menyumbang 25% dari semua kasus kanker, dan mempengaruhi lebih dari 2,1 Juta orang pada tahun 2015. Kanker ini bermula ketika sel-sel di payudara tumbuh secara tidak terkendali membentuk tumor yang dapat dilihat melalui sinar-X atau dirasakan sebagai benjolan di area payudara.

Salah satu tantangan utama dalam mendeteksi kanker payudara adalah mengklasifikasikan tumor sebagai kanker ganas atau jinak. Oleh karena itu, selesaikan analisis klasifikasi tumor ini menggunakan machine learning.

Dataset: Breast Cancer Wisconsin

[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_breast\\_cancer.html#sklearn.datasets.load\\_breast\\_cancer](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer)



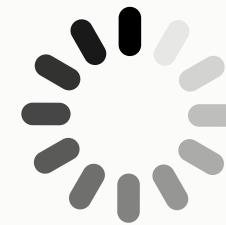
A close-up, microscopic image showing several breast cancer cells. The cells are spherical with distinct nuclei and some internal organelles visible. They are surrounded by a clear, watery fluid containing smaller, more numerous cellular fragments and debris. The overall color palette is a soft, pale blue or lavender.

# TUJUAN ANALISIS

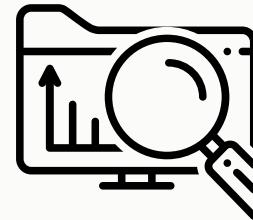
---

1. Membangun dan melatih model machine learning untuk mengklasifikasi kanker payudara berdasarkan dataset yang tersedia.
2. Mengukur akurasi setiap algoritma yang digunakan untuk menentukan model dengan performa terbaik.
3. Menganalisis kelebihan dan kekurangan berbagai algoritma machine learning dalam memprediksi kanker payudara.
4. Menyajikan hasil evaluasi model dalam bentuk diagram batang untuk mempermudah pemahaman dan perbandingan kinerja antar model.

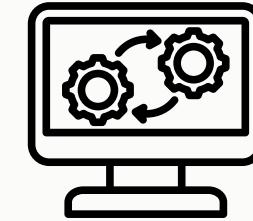
# METODOLOGI



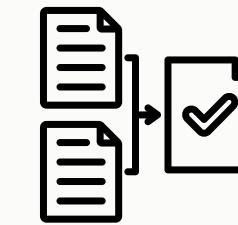
Load Data



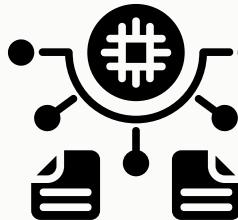
Exploratory Data Analysis



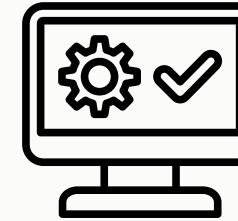
Memisahkan Data Training dan Testing



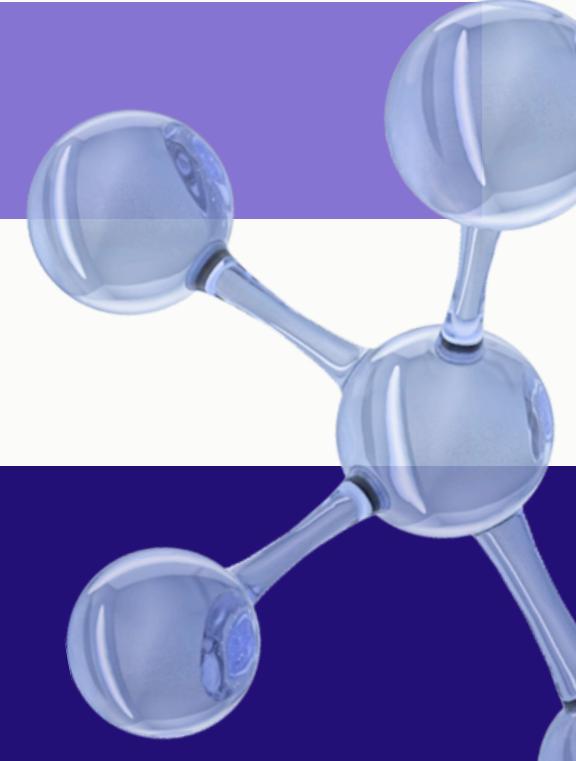
Normalisasi Data



Membangun dan Mengevaluasi Model



Kesimpulan



# LOAD DATA

```
[1] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_breast_cancer
import math
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import MinMaxScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB

# Memuat dataset breast cancer dari scikit-learn
cancer = load_breast_cancer()

# Memisahkan fitur dan target
x = cancer.data      # Fitur (data)
y = cancer.target    # Target (label)

# Mengubah array fitur menjadi DataFrame
df_x = pd.DataFrame(x, columns=cancer.feature_names)
# Mengubah array target menjadi Series
df_y = pd.Series(y, name='target')
```

```
[2] # Menggabungkan fitur dan target dalam satu DataFrame
df = pd.concat([df_x, df_y], axis=1)

# Menampilkan 5 baris pertama data
df.head()
```

INPUT

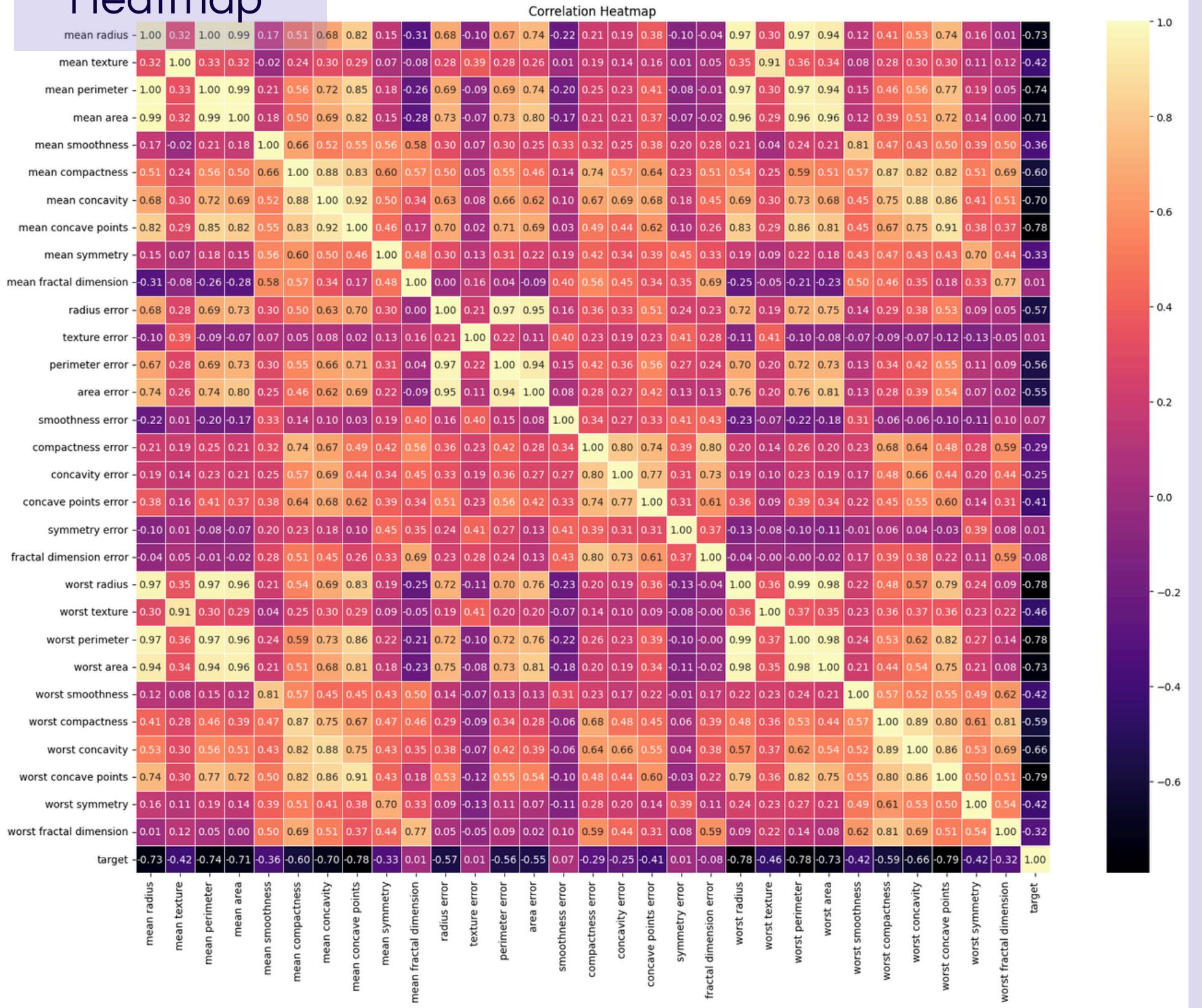
OUTPUT

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625

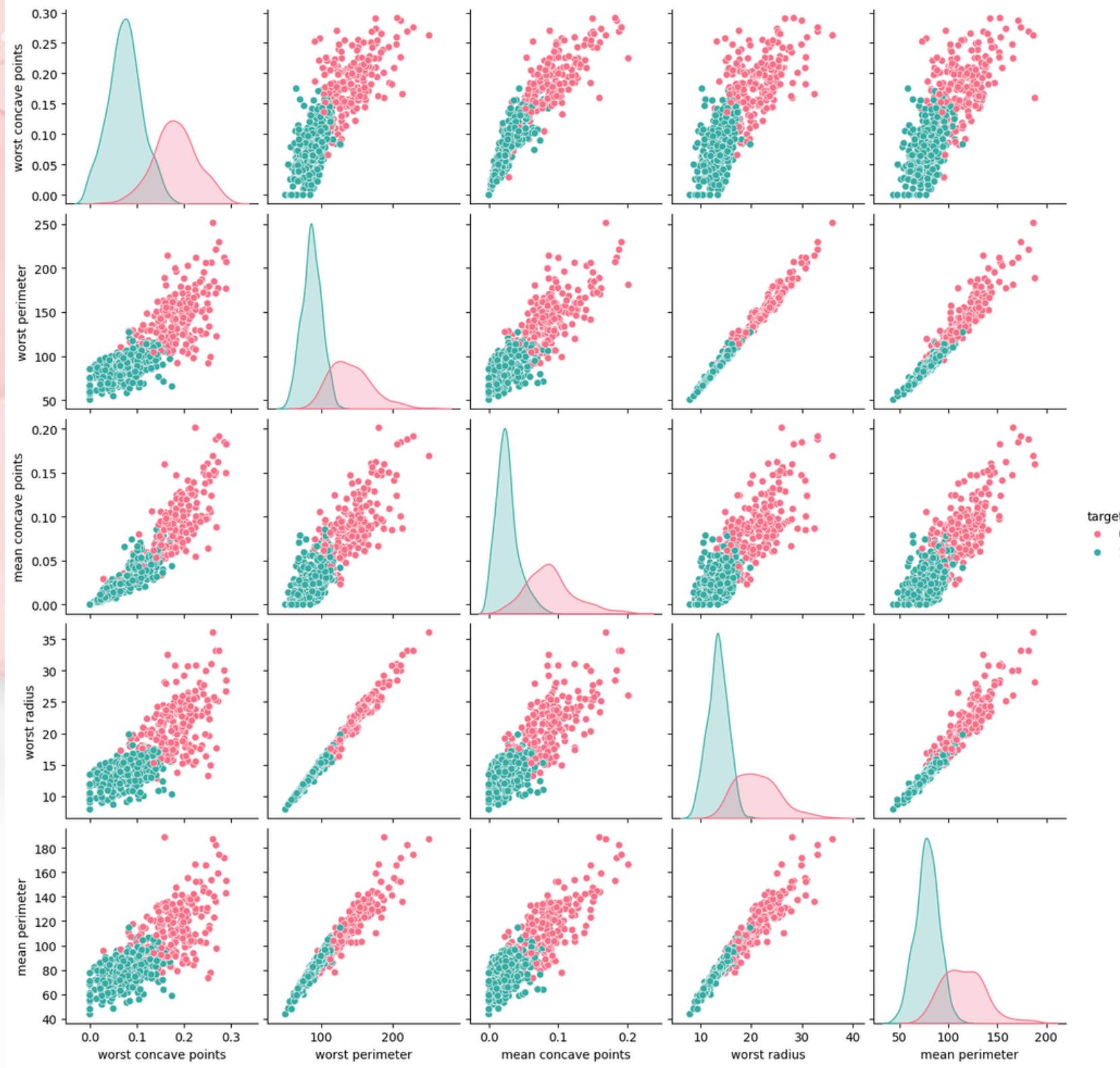
1. Output akan menampilkan 5 baris pertama dari DataFrame df.
2. Setiap baris merepresentasikan satu sampel dari dataset kanker payudara.
3. Kolom-kolom tersebut menunjukkan variabel yang akan di analisis, sementara kolom target menunjukkan klasifikasinya.

# EXPLORATORY DATA ANALYSIS

Heatmap



# EXPLORATORY DATA ANALYSIS

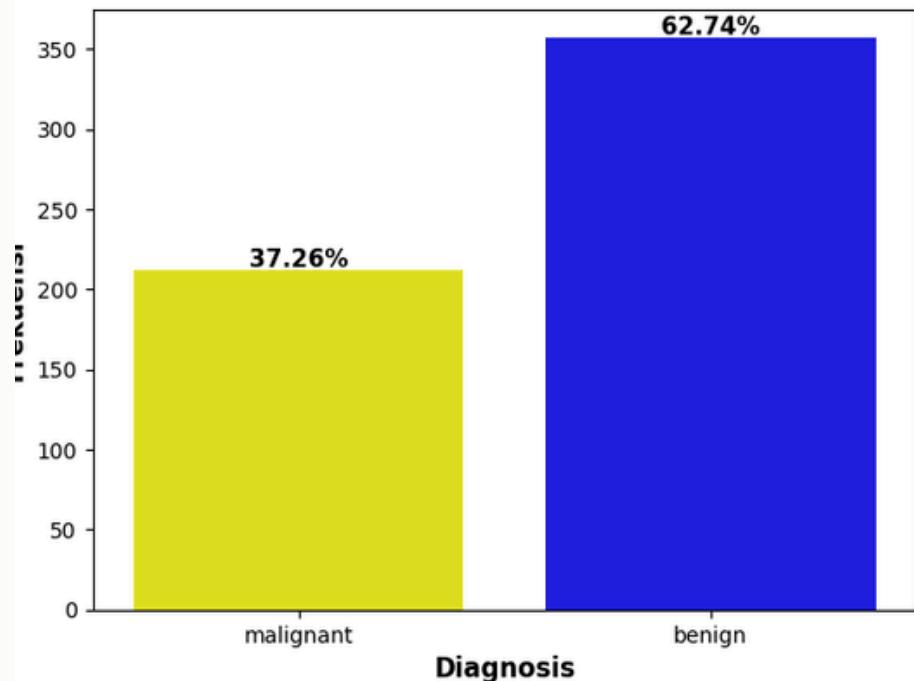


Pairplot

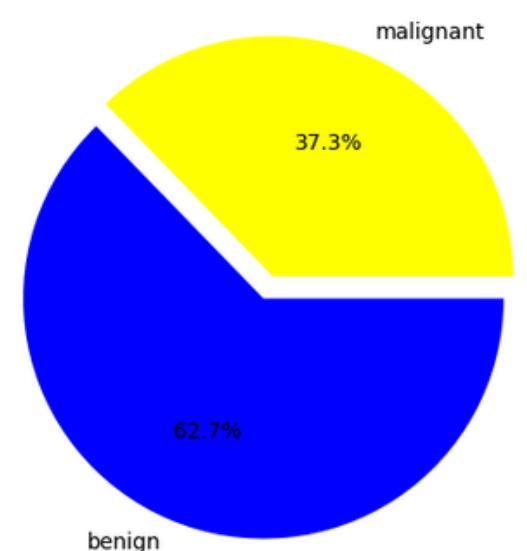
- Terlihat bahwa kelas 1 (tumor jinak) memiliki distribusi nilai yang lebih tinggi dibandingkan kelas 0 (tumor ganas). Hal ini menunjukkan bahwa kelas 1 (tumor jinak) cenderung lebih terkonsentrasi dan memiliki rentang nilai yang lebih besar, sementara kelas 0 (ganis) terlihat lebih menyebar dan memiliki variasi yang lebih luas.
- Dari scatter plot terlihat bahwa sebagian besar variabel menunjukkan korelasi positif yang kuat. Ini berarti bahwa ketika satu variabel meningkat (misalnya mean perimeter) maka variabel lain yang berkaitan (seperti worst perimeter) juga meningkat. Korelasi tinggi menunjukkan bahwa variabel-variabel ini memiliki hubungan linier.

# EXPLORATORY DATA ANALYSIS

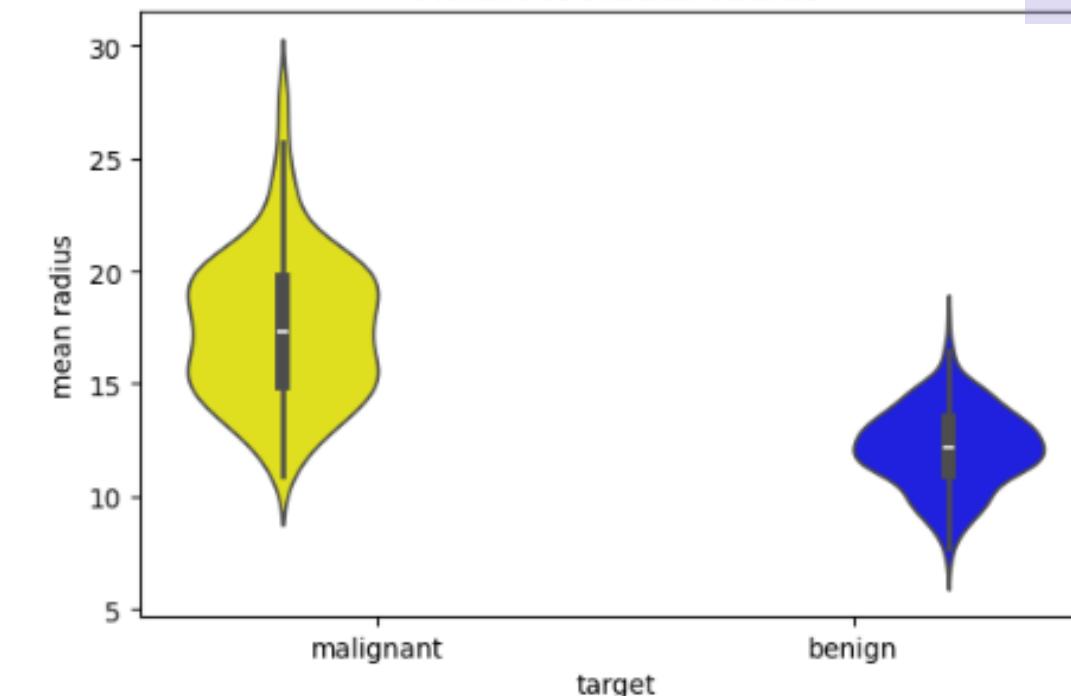
Diagnosis Kanker



Bar Chart & Pie Chart



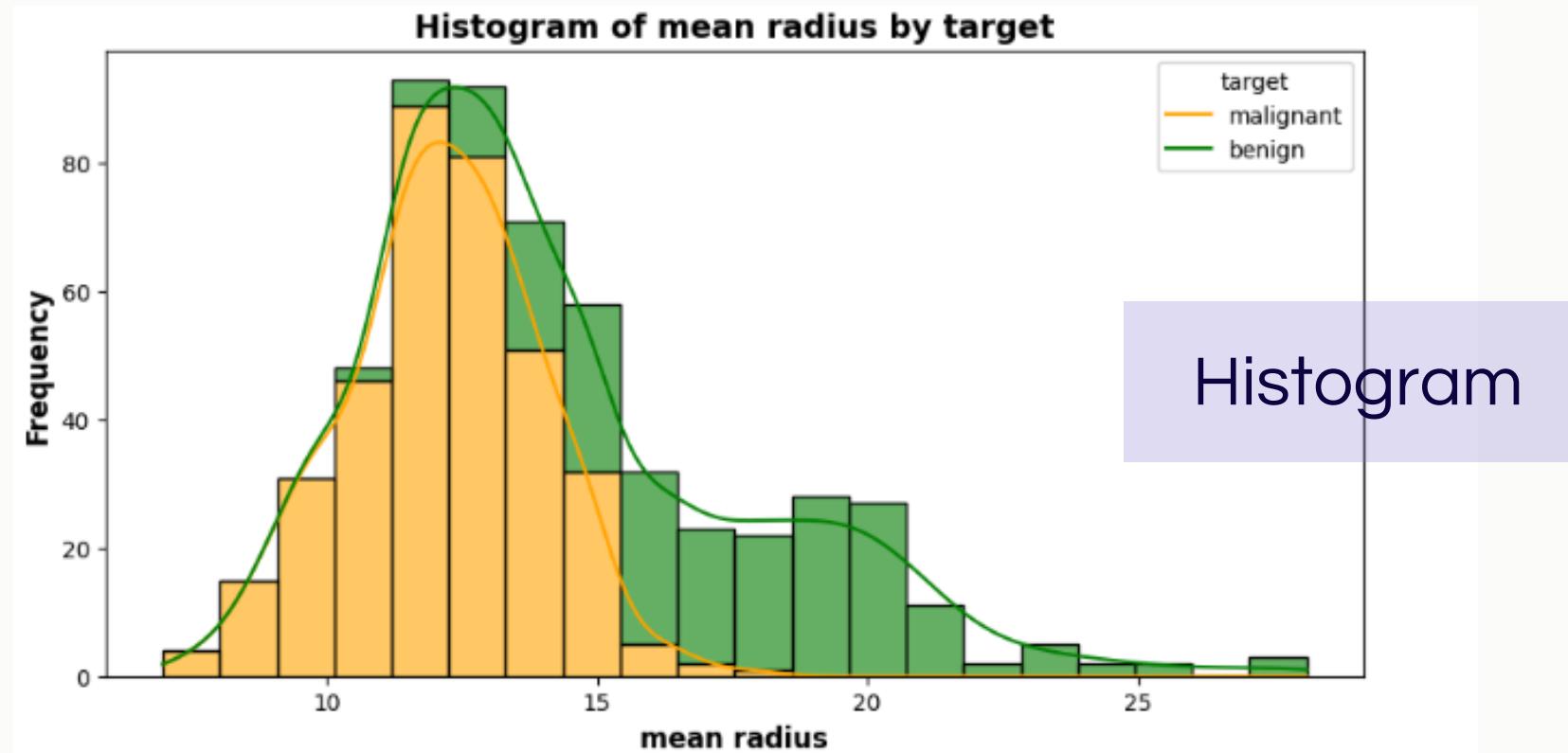
Violin Plot mean radius



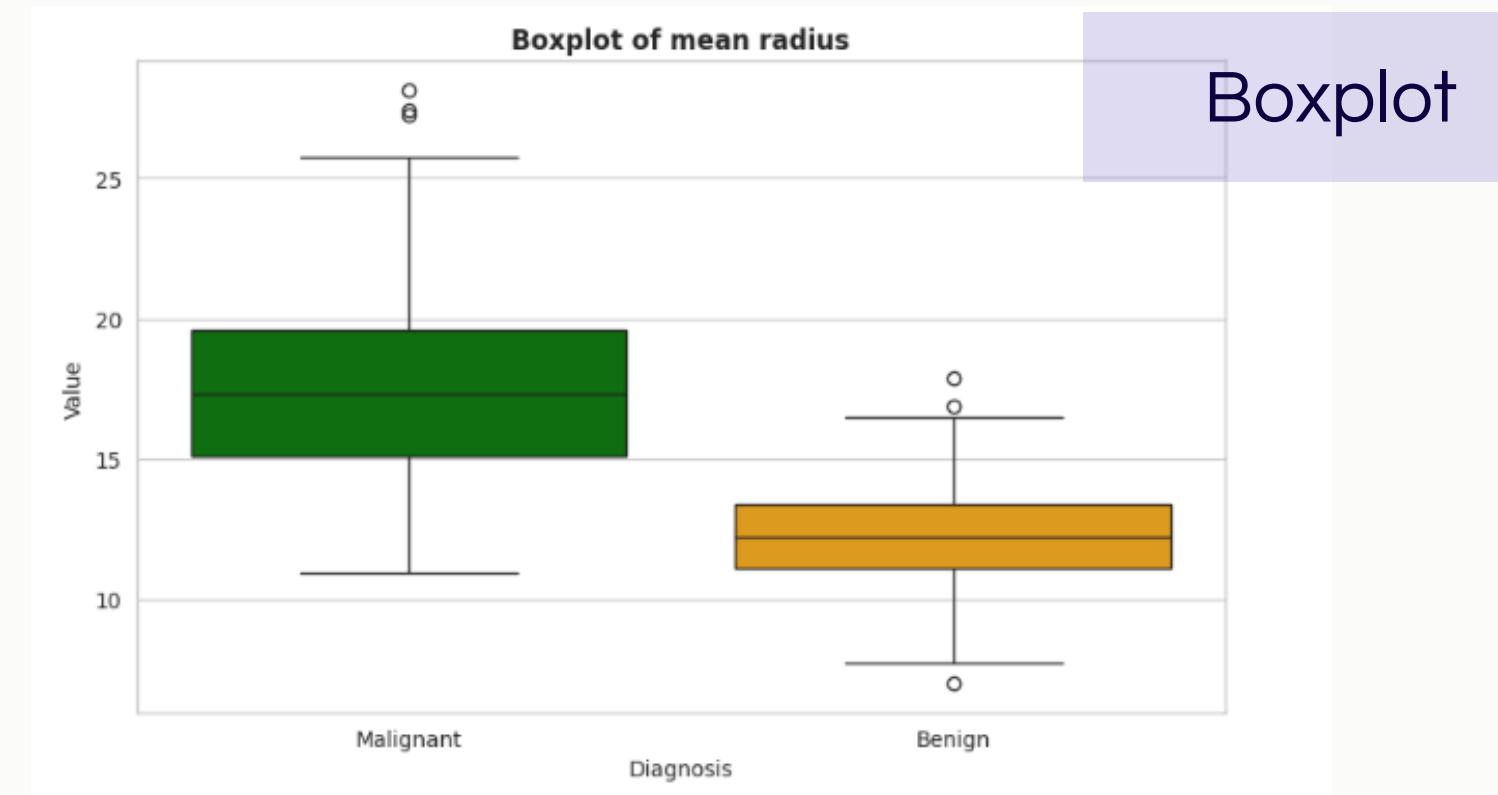
Output tersebut menunjukkan bahwa lebih banyak pasien yang didiagnosis dengan kanker jinak (benign) dibandingkan dengan kanker ganas (malignant). Dari total 569 sampel, sekitar 357 sampel (62.7%) adalah kanker jinak, sedangkan 212 sampel (37.3%) adalah kanker ganas.

Berdasarkan output, kita dapat memahami bahwa bagian yang lebih lebar pada grafik menunjukkan bahwa data lebih sering muncul, sedangkan bagian yang lebih sempit menandakan bahwa data lebih jarang ditemukan. Garis di tengah violin menggambarkan median serta kuartil data. Selain itu, variabel yang berkaitan dengan ukuran tumor (seperti radius, perimeter, dan area) memiliki perbedaan distribusi yang jelas antara kanker ganas dan jinak. Ini menunjukkan bahwa ukuran tumor merupakan faktor penting dalam membedakan jenis kanker payudara.

# EXPLORATORY DATA ANALYSIS



Histogram



Boxplot

- Kanker jinak (hijau) memiliki mean radius lebih besar dibanding kanker ganas (orange).
- Garis hijau dan oranye menunjukkan pola kepadatan data.
- Ada data yang tumpang tindih di sekitar mean radius 12-15, ini menunjukkan bahwa beberapa tumor jinak dan ganas memiliki ukuran yang mirip.

- Beberapa variabel dalam dataset memiliki outlier, menunjukkan adanya nilai ekstrem yang perlu diperhatikan.
- Sebaran data kanker ganas pada beberapa variabel cenderung lebih luas dibandingkan kanker jinak, menandakan bahwa karakteristik kanker ganas lebih beragam. Sedangkan, kanker jinak memiliki sebaran data yang lebih sempit, menunjukkan bahwa karakteristiknya lebih seragam..
- Nilai median pada kelompok kanker ganas lebih tinggi dibandingkan dengan kelompok kanker jinak.

# EXPLORATORY DATA ANALYSIS

## Menghilangkan Outlier

```
[23] # Fungsi untuk mengganti outlier dengan median
def replace_outliers_with_median(df, features):
    for col in features:
        Q1 = np.percentile(df[col], 25)
        Q3 = np.percentile(df[col], 75)
        IQR = Q3 - Q1
        outlier_step = 1.5 * IQR

        # Batas bawah dan atas
        lower_bound = Q1 - outlier_step
        upper_bound = Q3 + outlier_step

        # Ganti outlier dengan median
        median_value = df[col].median()
        df[col] = np.where((df[col] < lower_bound) | (df[col] > upper_bound), median_value, df[col])

    return df

# List fitur numerik tanpa target
numerical_features = df.columns.drop('target')

# Terapkan fungsi
df_cleaned = replace_outliers_with_median(df.copy(), numerical_features)
df_cleaned
```

## Output

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry
0	17.99	10.38	122.80	1001.0	0.11840	0.09263	0.06154	0.14710	0.2419	0.07871	...	17.33	184.60	686.5	0.16220	0.21190	0.7119	0.2654	0.2822
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	23.41	158.80	686.5	0.12380	0.18660	0.2416	0.1860	0.2750
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	25.53	152.50	1709.0	0.14440	0.42450	0.4504	0.2430	0.3613
3	11.42	20.38	77.58	386.1	0.09587	0.09263	0.24140	0.10520	0.1792	0.06154	...	26.50	98.87	567.7	0.13130	0.21190	0.6869	0.2575	0.2822
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	16.67	152.20	1575.0	0.13740	0.20500	0.4000	0.1625	0.2364
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
564	21.56	22.39	142.00	551.1	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	686.5	0.14100	0.21130	0.4107	0.2216	0.2060
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	0.2572
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	0.2218
567	20.60	29.33	140.10	1265.0	0.11780	0.09263	0.06154	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500	0.21190	0.2267	0.2650	0.4087
568	7.76	24.54	47.92	181.0	0.09587	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996	0.06444	0.0000	0.0000	0.2871

569 rows × 31 columns

# MEMISAHKAN DATA TRAINING DAN TESTING



```
# Pisahkan fitur (X) dan target (y)
X = df_cleaned.drop('target', axis=1)
y = df_cleaned['target']
```

```
# Membagi data menjadi training dan testing (misalnya, 80% training dan 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

## 1. Training Set (Data Latihan):

- Biasanya 80% dari data digunakan untuk melatih model.
- Model akan belajar dari data dengan mengenali pola yang ada.

## 2. Testing Set (Data Uji):

- Sisanya 20% dari data digunakan untuk mengukur performa model.
- Model diuji pada data yang tidak pernah dilihat sebelumnya, sehingga bisa diketahui apakah model bekerja dengan baik atau tidak.

# **NORMALISASI DATA**

# Input

```
# Normalisasi fitur menggunakan MinMaxScaler
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Konversi hasil normalisasi ke DataFrame agar lebih mudah dibaca
X_train_scaled_df = pd.DataFrame(X_train_scaled, columns=X_train.columns)
X_train_scaled_df.head()
```

# Output

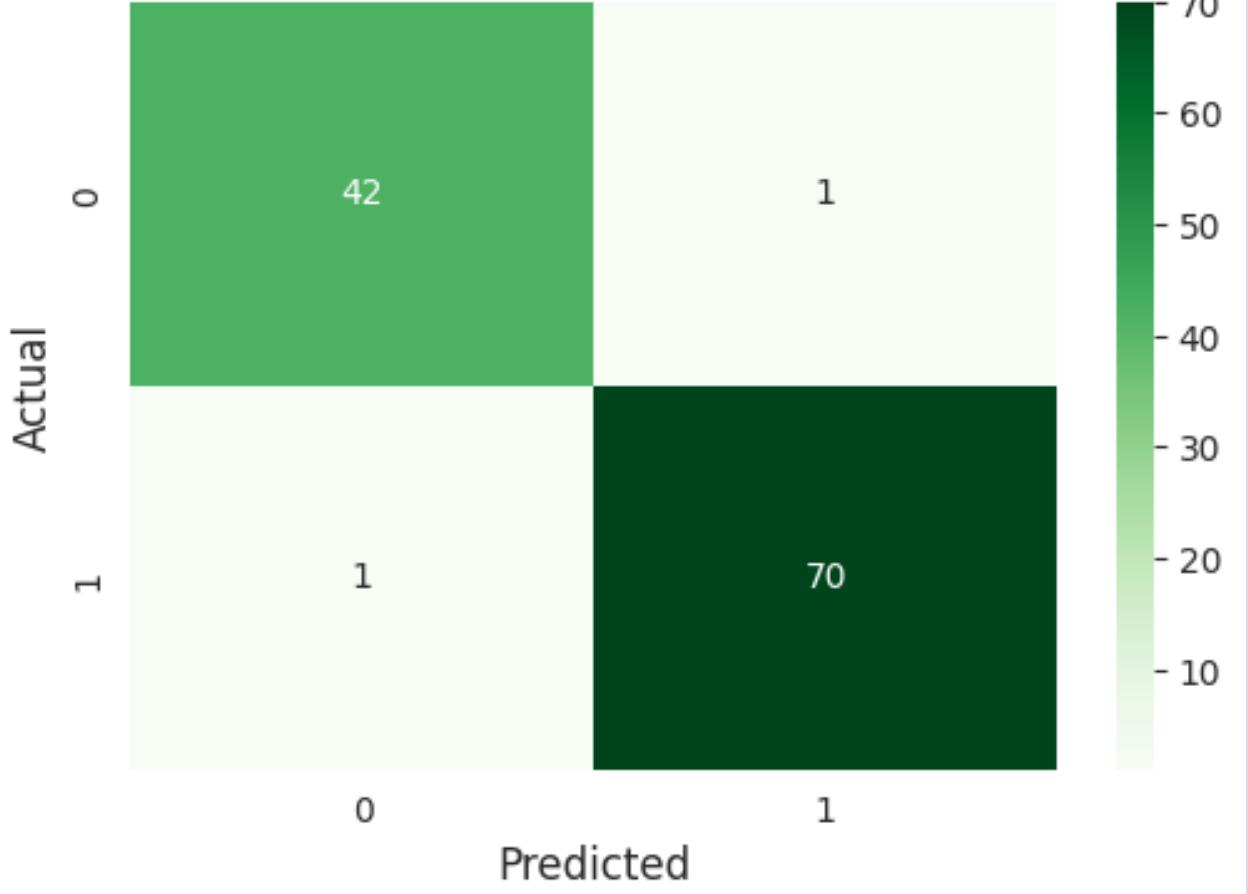
	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	...
0	0.095170	0.376111	0.109378	0.069315	0.629048	0.585535	0.225174	0.290891	0.730650	0.402783	...	0.087544	0.357191	0.083214	0.05
1	0.953055	0.832182	0.953713	0.987020	0.736054	0.351791	0.909989	0.994681	0.950464	0.835478	...	0.965669	0.721102	0.922153	0.21
2	0.105413	0.204837	0.113504	0.078314	0.209730	0.327202	0.219100	0.144947	0.908669	0.684174	...	0.071452	0.242272	0.083894	0.05
3	0.210470	0.765548	0.202153	0.152821	0.485947	0.254346	0.087047	0.107380	0.565015	0.463652	...	0.191610	0.778562	0.177538	0.11
4	0.176328	0.255183	0.167337	0.122447	0.726066	0.294064	0.039663	0.085771	0.483746	0.683130	...	0.125630	0.182460	0.116167	0.08

Output tersebut menunjukkan bahwa dataset berada dalam rentang 0 hingga 1. Ini memastikan bahwa semua variabel memiliki skala yang seragam, sehingga tidak ada variabel yang mendominasi perhitungan akibat perbedaan skala. Dengan demikian, algoritma machine learning dapat bekerja lebih optimal.



# MEMBANGUN DAN MENGEVALUASI MODEL

Confusion Matrix - Logistic Regression



Classification Report - Logistic Regression:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	43
1	0.99	0.99	0.99	71
accuracy			0.98	114
macro avg	0.98	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

Logistic Regression

## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 70 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 42 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 1 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 1 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 42 + 1 = 43$ ), model salah mengklasifikasikan 1 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 1 + 70 = 71$ ), model salah mengklasifikasikan 1 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.98, yang berarti sekitar 98% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.99, yang berarti sekitar 99% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

- Recall untuk kelas 0 (kanker ganas) sebesar 0.98, menunjukkan bahwa model mampu mengidentifikasi sekitar 98% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.99, menunjukkan bahwa model mampu mengidentifikasi sekitar 99% dari seluruh kasus kanker jinak yang ada.

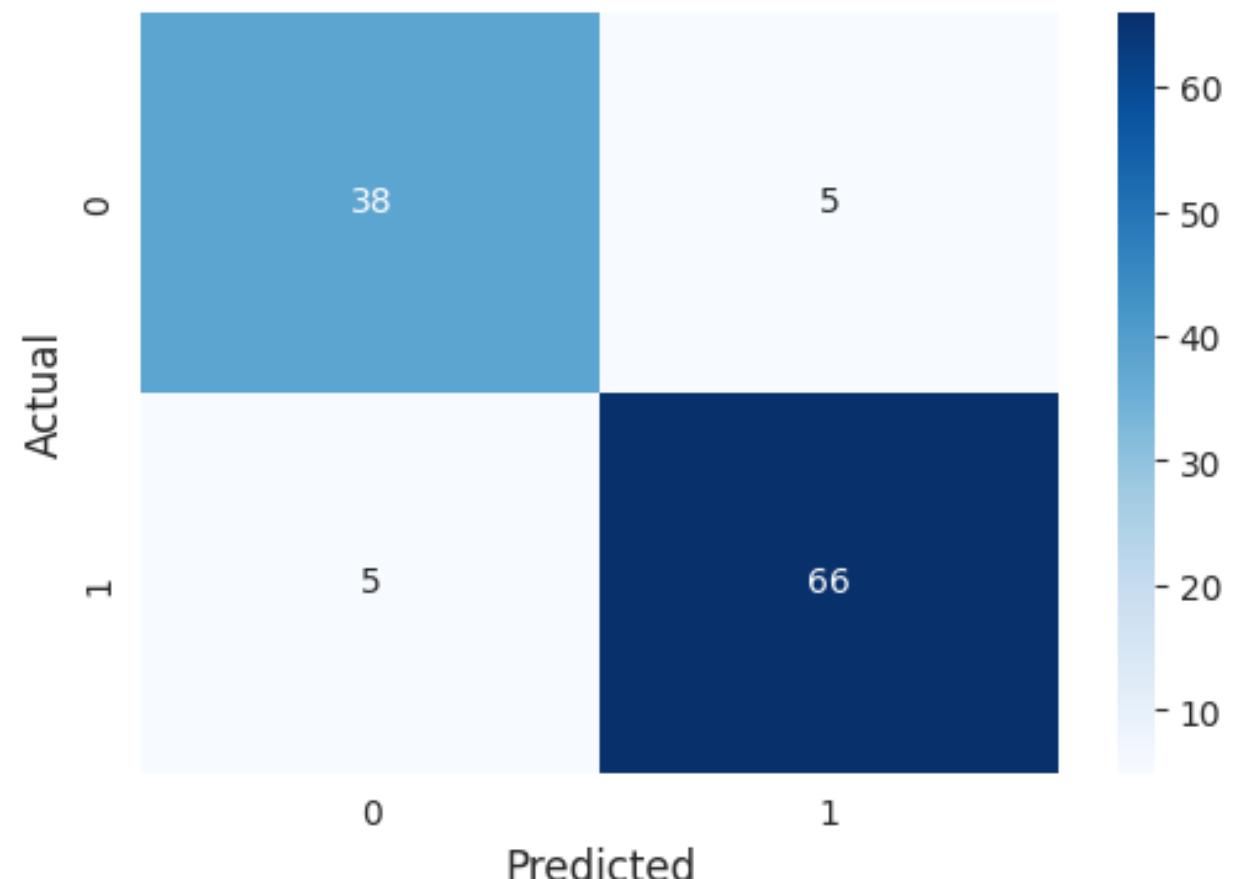
### F1-score:

- F1-score untuk kelas 0 adalah 0.98
- F1-score untuk kelas 1 adalah 0.99.

Accuracy: Model ini memiliki akurasi sebesar 0.98, yang berarti model mampu memberikan prediksi yang benar sebanyak 98% dari seluruh dataset.

# MEMBANGUN DAN MENGEVALUASI MODEL

Confusion Matrix - Decision Tree



Classification Report - Decision Tree:				
	precision	recall	f1-score	support
0	0.88	0.88	0.88	43
1	0.93	0.93	0.93	71
accuracy			0.91	114
macro avg	0.91	0.91	0.91	114
weighted avg	0.91	0.91	0.91	114

Decision Tree

## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 66 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 38 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 5 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 5 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 38 + 5 = 43$ ), model salah mengklasifikasikan 5 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 5 + 66 = 71$ ), model salah mengklasifikasikan 5 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.88, yang berarti sekitar 88% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.93, yang berarti sekitar 93% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

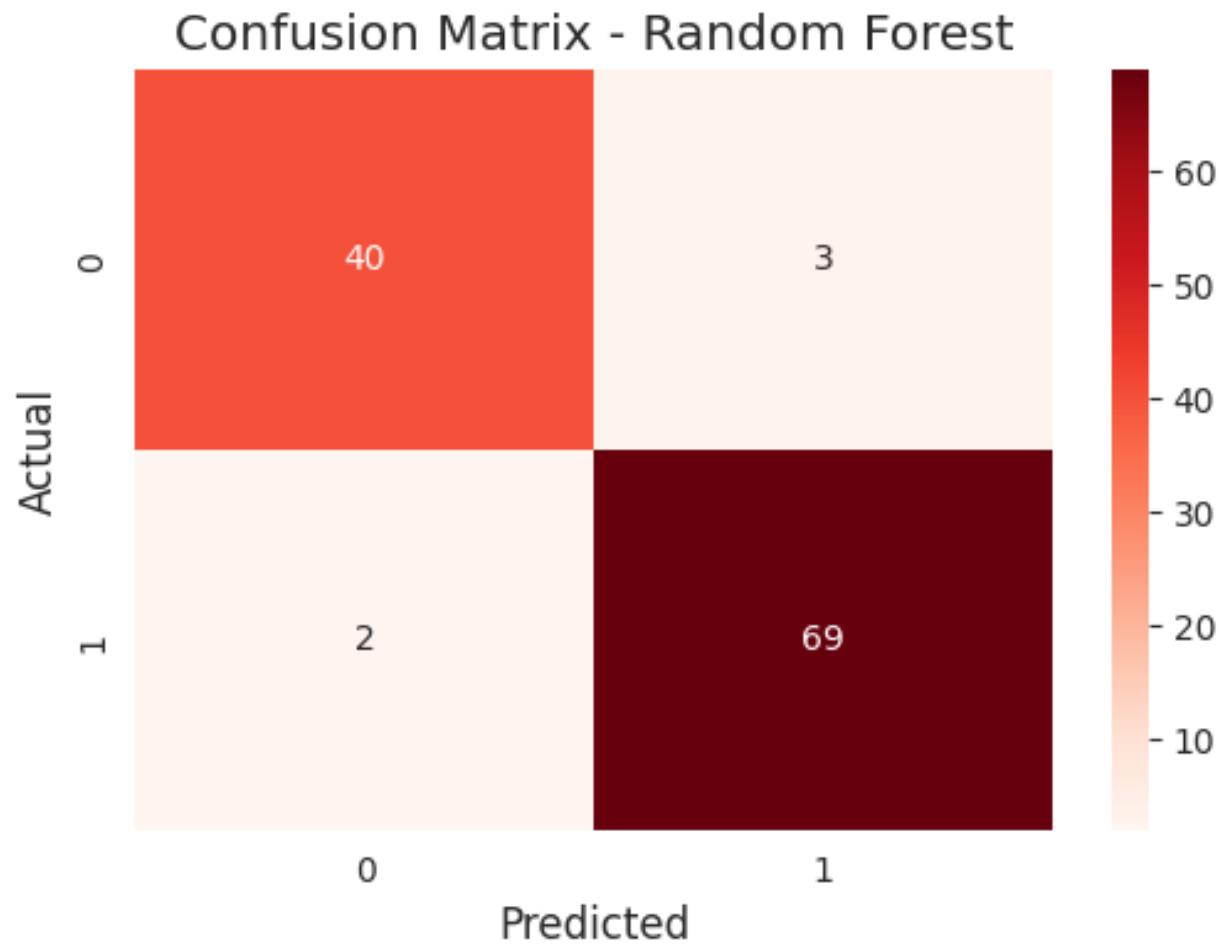
- Recall untuk kelas 0 (kanker ganas) sebesar 0.88, menunjukkan bahwa model mampu mengidentifikasi sekitar 88% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.93, menunjukkan bahwa model mampu mengidentifikasi sekitar 93% dari seluruh kasus kanker jinak yang ada.

### F1-score:

- F1-score untuk kelas 0 adalah 0.88
- F1-score untuk kelas 1 adalah 0.93

Accuracy: Model ini memiliki akurasi sebesar 0.91, yang berarti model mampu memberikan prediksi yang benar sebanyak 91% dari seluruh dataset.

# MEMBANGUN DAN MENGEVALUASI MODEL



Classification Report - Random Forest:				
	precision	recall	f1-score	support
0	0.95	0.93	0.94	43
1	0.96	0.97	0.97	71
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Random Forest

## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 69 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 40 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 3 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 2 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 40 + 3 = 43$ ), model salah mengklasifikasikan 3 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 2 + 69 = 71$ ), model salah mengklasifikasikan 2 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.95, yang berarti sekitar 95% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.96, yang berarti sekitar 96% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

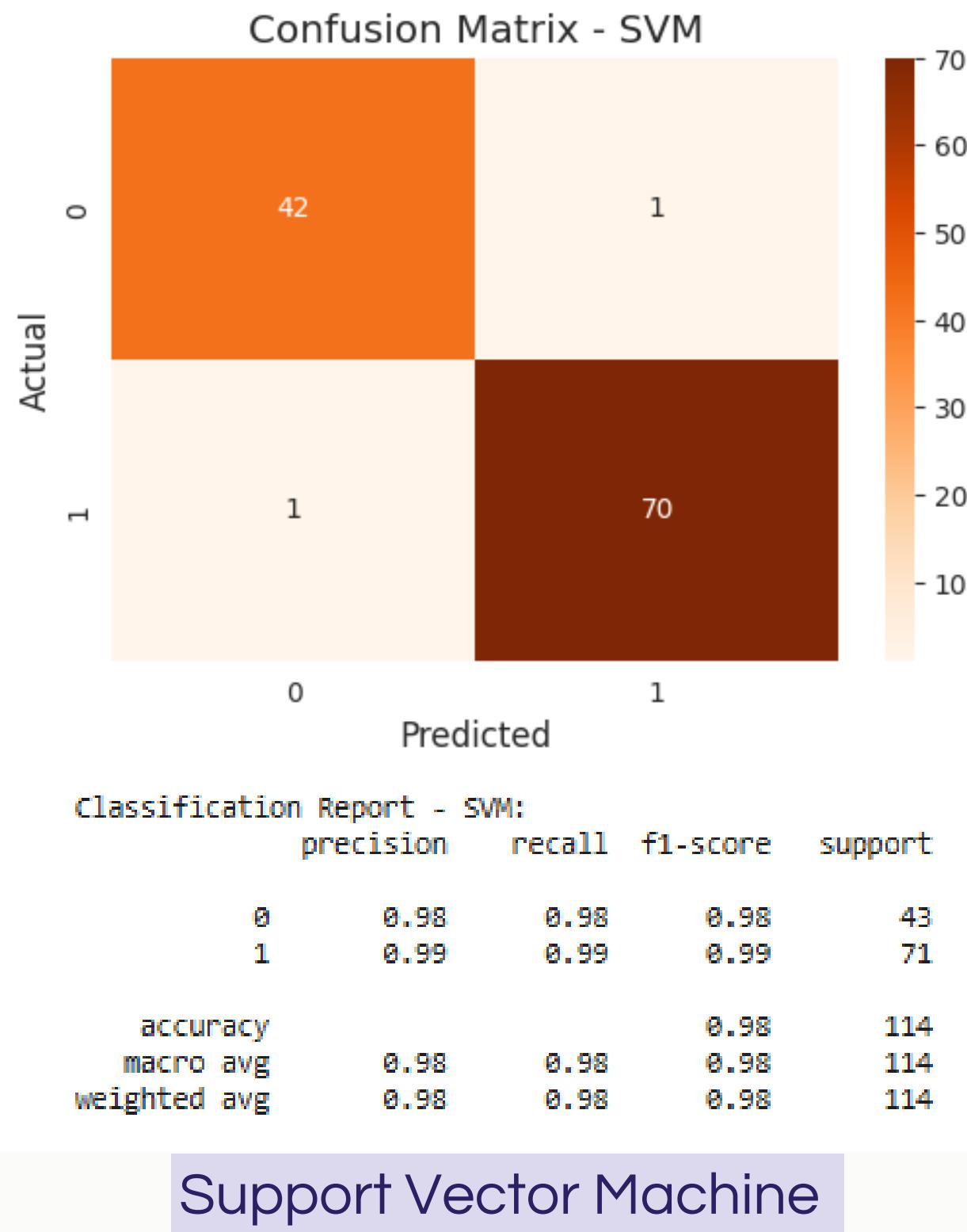
- Recall untuk kelas 0 (kanker ganas) sebesar 0.93, menunjukkan bahwa model mampu mengidentifikasi sekitar 93% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.97, menunjukkan bahwa model mampu mengidentifikasi sekitar 97% dari seluruh kasus kanker jinak yang ada.

### F1-score:

- F1-score untuk kelas 0 adalah 0.94
- F1-score untuk kelas 1 adalah 0.97

Accuracy: Model ini memiliki akurasi sebesar 0.96, yang berarti model mampu memberikan prediksi yang benar sebanyak 96% dari seluruh dataset.

# MEMBANGUN DAN MENGEVALUASI MODEL



## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 70 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 42 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 1 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 1 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 42 + 1 = 43$ ), model salah mengklasifikasikan 1 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 1 + 70 = 71$ ), model salah mengklasifikasikan 1 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.98, yang berarti sekitar 98% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.99, yang berarti sekitar 99% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

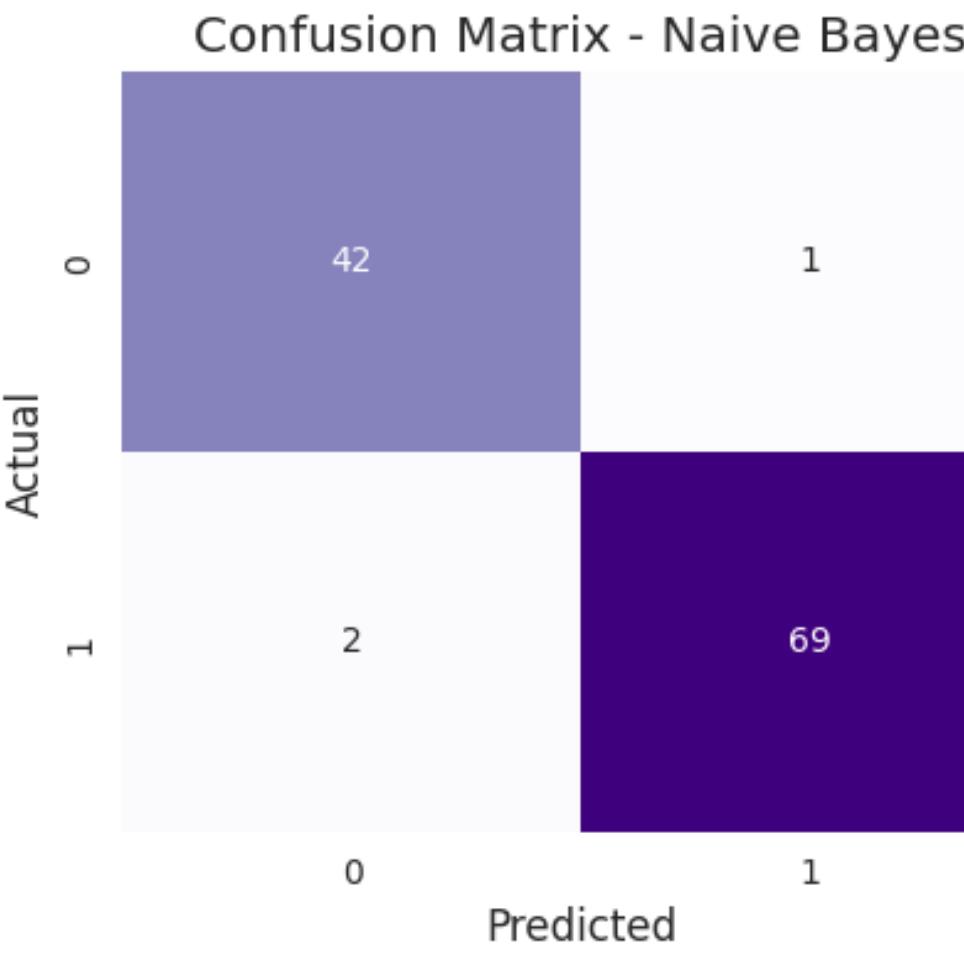
- Recall untuk kelas 0 (kanker ganas) sebesar 0.98, menunjukkan bahwa model mampu mengidentifikasi sekitar 98% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.99, menunjukkan bahwa model mampu mengidentifikasi sekitar 99% dari seluruh kasus kanker jinak yang ada.

### F1-score:

- F1-score untuk kelas 0 adalah 0.98
- F1-score untuk kelas 1 adalah 0.99

Accuracy: Model ini memiliki akurasi sebesar 0.98, yang berarti model mampu memberikan prediksi yang benar sebanyak 98% dari seluruh dataset.

# MEMBANGUN DAN MENGEVALUASI MODEL



Classification Report - Naive Bayes:				
	precision	recall	f1-score	support
0	0.95	0.98	0.97	43
1	0.99	0.97	0.98	71
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Naive Bayes

## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 69 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 42 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 1 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 2 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 42 + 1 = 43$ ), model salah mengklasifikasikan 1 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 2 + 69 = 71$ ), model salah mengklasifikasikan 2 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.95, yang berarti sekitar 95% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.99, yang berarti sekitar 99% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

- Recall untuk kelas 0 (kanker ganas) sebesar 0.98, menunjukkan bahwa model mampu mengidentifikasi sekitar 98% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.97, menunjukkan bahwa model mampu mengidentifikasi sekitar 97% dari seluruh kasus kanker jinak yang ada.

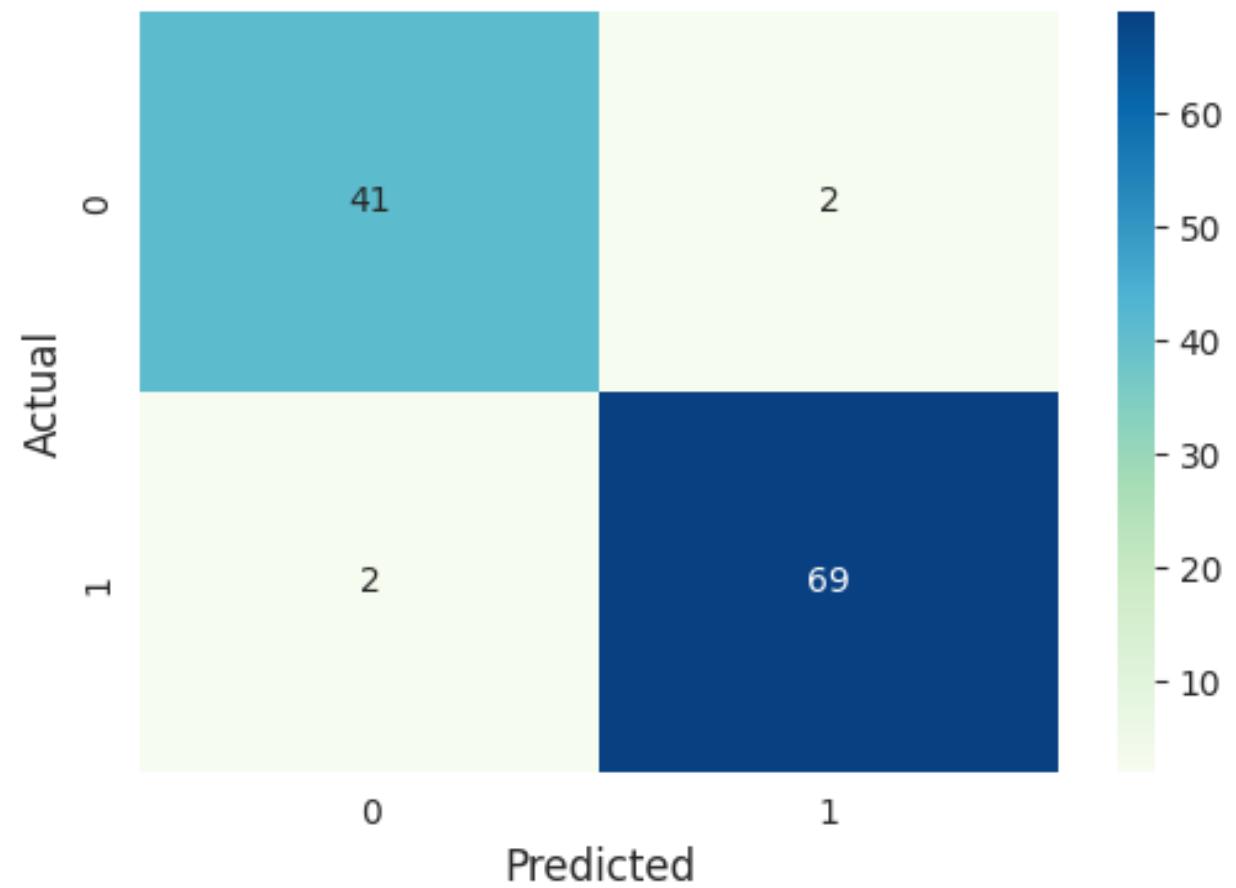
### F1-score:

- F1-score untuk kelas 0 adalah 0.97
- F1-score untuk kelas 1 adalah 0.98

Accuracy: Model ini memiliki akurasi sebesar 0.97, yang berarti model mampu memberikan prediksi yang benar sebanyak 97% dari seluruh dataset.

# MEMBANGUN DAN MENGEVALUASI MODEL

Confusion Matrix - KNN



Classification Report - KNN:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	43
1	0.97	0.97	0.97	71
accuracy			0.96	114
macro avg	0.96	0.96	0.96	114
weighted avg	0.96	0.96	0.96	114

K-Nearest Neighbors

## Confusion Matrix:

- True Positives (TP): Model berhasil mengklasifikasikan 69 kasus kanker jinak dengan benar.
- True Negatives (TN): Model berhasil mengklasifikasikan 41 kasus kanker ganas dengan benar.
- False Positives (FP): Sebanyak 2 kasus kanker ganas salah diklasifikasikan sebagai kanker jinak.
- False Negatives (FN): Sebanyak 2 kasus kanker jinak salah diklasifikasikan sebagai kanker ganas.

Dari total 43 kasus kanker ganas ( $TN + FP = 41 + 2 = 43$ ), model salah mengklasifikasikan 2 kasus sebagai kanker jinak (FP). Sementara itu, dari 71 kasus kanker jinak ( $FN + TP = 2 + 69 = 71$ ), model salah mengklasifikasikan 2 kasus sebagai kanker ganas (FN).

## Classification Report:

### Precision:

- Untuk kelas 0 (kanker ganas), precision sebesar 0.95, yang berarti sekitar 95% dari kasus yang diprediksi sebagai kanker ganas memang benar kanker ganas.
- Untuk kelas 1 (kanker jinak), precision sebesar 0.97, yang berarti sekitar 97% dari kasus yang diprediksi sebagai kanker jinak memang benar kanker jinak.

### Recall:

- Recall untuk kelas 0 (kanker ganas) sebesar 0.95, menunjukkan bahwa model mampu mengidentifikasi sekitar 95% dari seluruh kasus kanker ganas yang ada.
- Recall untuk kelas 1 (kanker jinak) sebesar 0.97, menunjukkan bahwa model mampu mengidentifikasi sekitar 97% dari seluruh kasus kanker jinak yang ada.

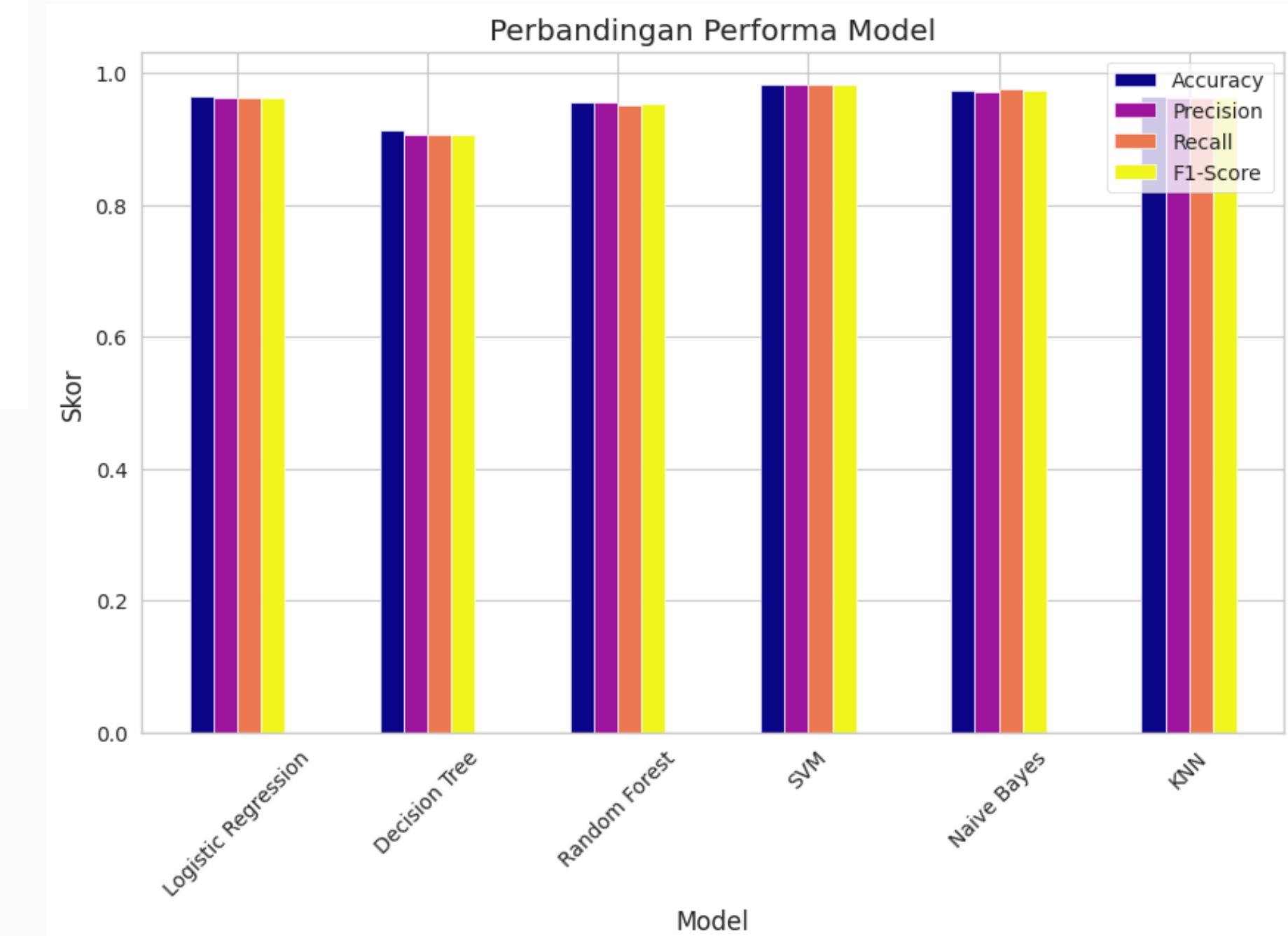
### F1-score:

- F1-score untuk kelas 0 adalah 0.95
- F1-score untuk kelas 1 adalah 0.97

Accuracy: Model ini memiliki akurasi sebesar 0.96, yang berarti model mampu memberikan prediksi yang benar sebanyak 96% dari seluruh dataset.

# PERBANDINGAN HASIL MODEL

	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.9649	0.9627	0.9627	0.9627
1	Decision Tree	0.9123	0.9066	0.9066	0.9066
2	Random Forest	0.9561	0.9554	0.9510	0.9531
3	SVM	0.9825	0.9813	0.9813	0.9813
4	Naive Bayes	0.9737	0.9701	0.9743	0.9721
5	KNN	0.9649	0.9627	0.9627	0.9627



# KESIMPULAN

Berdasarkan hasil analisis, diketahui bahwa distribusi label target tidak seimbang. Oleh karena itu, dalam menentukan model klasifikasi terbaik, kita perlu mempertimbangkan nilai Precision dan Recall. Dari enam model machine learning yang diuji, Support Vector Machine (SVM) terbukti menjadi metode terbaik untuk mengklasifikasikan kasus kanker payudara, dengan nilai Precision dan Recall sebesar 0.9813. Hasil evaluasi menunjukkan bahwa dari 43 kasus kanker ganas (42 benar terdeteksi sebagai ganas dan 1 salah terkласifikasi sebagai jinak), model hanya membuat 1 kesalahan klasifikasi. Sementara itu, dari 71 kasus kanker jinak (70 benar terdeteksi sebagai jinak dan 1 salah terkласifikasi sebagai ganas), model juga hanya melakukan 1 kesalahan. Hal ini membuktikan bahwa model SVM memiliki tingkat akurasi yang sangat baik dalam mendeteksi kedua jenis kanker dengan kesalahan minimal.



# CONTACT



[www.linkedin.com/in/finaputri](https://www.linkedin.com/in/finaputri)



082146303817



[vinaputri412@gmail.com](mailto:vinaputri412@gmail.com)



<https://github.com/ffnn19>

