

Tugas 6: Klasifikasi Data Kanker Payudara Menggunakan Algoritma Support Vector Machine (SVM)

Monika Septiana - 0110222127 ^{1*}

¹ Teknik Informatika, STT Terpadu Nurul Fikri, Depok

*E-mail: moni22127ti@student.nurulfikri.ac.id

Abstract. Penelitian ini bertujuan untuk mengimplementasikan algoritma *Support Vector Machine* (SVM) dalam melakukan klasifikasi terhadap data kanker payudara. Dataset yang digunakan berasal dari platform Kaggle dengan judul *Breast Cancer Dataset*, yang berisi data hasil pengukuran karakteristik sel untuk membedakan antara tumor jinak (*Benign*) dan ganas (*Malignant*). Proses pengolahan data mencakup pembersihan kolom kosong, konversi label diagnosis menjadi numerik, normalisasi data, dan pembagian dataset menjadi data latih serta data uji. Model SVM dengan kernel RBF digunakan untuk melakukan pelatihan dan prediksi. Hasil pengujian menunjukkan bahwa model SVM mampu mencapai akurasi sebesar 95.9%, dengan performa klasifikasi yang sangat baik. Berdasarkan hasil tersebut, dapat disimpulkan bahwa algoritma SVM efektif dalam mendeteksi kanker payudara secara akurat berdasarkan fitur morfologi sel.

1. Tujuan

Melakukan klasifikasi data menggunakan algoritma Support Vector Machine (SVM) pada dataset dari Kaggle, serta menganalisis hasil performa model.

2. Dataset

- Nama Dataset: Breast Cancer Dataset
- Sumber Kaggle: <https://www.kaggle.com/datasets/wasiqaliyasir/breast-cancer-dataset>
- Jumlah Data: 569 baris × 33 kolom
- Tujuan: Memprediksi apakah tumor bersifat *Malignant* (ganas) atau *Benign* (jinak).

Deskripsi Kolom

- diagnosis: Target/label (M = Malignant, B = Benign)
- radius_mean, texture_mean, perimeter_mean, dst.: fitur numerik hasil pengukuran sel kanker.
- Kolom Unnamed: 32 dihapus karena seluruh nilainya NaN.

3. Algoritma: Support Vector Machine (SVM)

SVM bekerja dengan mencari **hyperplane** terbaik untuk memisahkan data antar kelas.

- Kernel yang digunakan: Radial Basis Function (RBF).
- Parameter utama: C (regularization) dan gamma (kernel width).

c. Cocok untuk data non-linear seperti pada kasus klasifikasi kanker.

4. Implementasi Kode (Google Colab)

```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# 1. Load dataset dari Google Drive
df = pd.read_csv('/content/drive/MyDrive/praktikum_ml/praktikum06/data/Breast_cancer_dataset.csv')

# 2. Hapus kolom ID dan kolom kosong
if 'id' in df.columns:
    df.drop(columns=['id'], inplace=True)
if 'Unnamed: 32' in df.columns:
    df.drop(columns=['Unnamed: 32'], inplace=True)

# 3. Ubah label diagnosis ke numerik
df['diagnosis'] = df['diagnosis'].map({'M': 1, 'B': 0})

# 4. Pisahkan fitur dan label
X = df.drop(columns=['diagnosis'])
y = df['diagnosis']

# 5. Split data train/test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=42)

# 6. Pipeline: imputasi, scaling, dan SVM
pipeline = make_pipeline(
    SimpleImputer(strategy='mean'),
    StandardScaler(),
    SVC(kernel='rbf', C=1.0, gamma='scale', random_state=42)
)

# 7. Latih model dan evaluasi
pipeline.fit(X_train, y_train)
```

```
y_pred = pipeline.predict(X_test)
```

```
# 8. Hasil evaluasi
```

```
print(f"Akurasi: {accuracy_score(y_test, y_pred):.4f}")
```

```
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

```
print("\nLaporan Klasifikasi:\n", classification_report(y_test, y_pred))
```

```
# 9. Visualisasi confusion matrix
```

```
plt.figure(figsize=(5,4))
```

```
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Reds')
```

```
plt.title("Confusion Matrix - SVM (Breast Cancer)")
```

```
plt.xlabel("Predicted")
```

```
plt.ylabel("Actual")
```

```
plt.show()
```

5. Hasil Output

Evaluasi Model

Metrik	Nilai
Akurasi	0.9591 (\approx 95.9%)
Precision (Malignant)	1.00
Recall (Malignant)	0.89
F1-Score (Malignant)	0.94

Confusion Matrix

	Pred: Benign (0) Pred: Malignant (1)	
Benign (0)	107	0
Malignant (1)	7	57

Model mampu mengenali sebagian besar kasus dengan benar, meskipun ada sedikit kesalahan pada 7 data malignant.

6. Analisis

- Kolom Unnamed: 32 dihapus karena seluruh nilainya kosong.
- SVM kernel RBF menghasilkan performa tinggi karena mampu menangkap pola non-linear.
- Model cukup efisien tanpa tuning parameter tambahan.
- Akurasi tinggi ($>95\%$) menunjukkan model efektif untuk prediksi kanker payudara.

7. Kesimpulan

- Implementasi algoritma Support Vector Machine (SVM) berhasil dilakukan pada dataset Breast Cancer (Kaggle).
- Model menghasilkan akurasi 95.9% dengan performa klasifikasi yang baik.
- SVM efektif untuk mendeteksi tumor jinak dan ganas berdasarkan fitur morfologi sel.
- Dataset telah dibersihkan dari kolom kosong sebelum pelatihan model.

8. Link Penting

Dataset Kaggle: <https://www.kaggle.com/datasets/wasiqaliyasir/breast-cancer-dataset>

GitHub Repository (isi kode & laporan): [https://github.com/monikasptn/Tugas-Praktikum Machine-Learning/tree/main/Praktikum ML/praktikum06/reports](https://github.com/monikasptn/Tugas-Praktikum-Machine-Learning/tree/main/Praktikum%20ML/praktikum06/reports)