LAPORAN PRAKTIKUM DATA MINING

ANALISA KLASIFIKASI SUPPORT VECTOR MACHINE UNTUK MEMPREDIKSI KANKER PAYUDARA DENGAN WISCONSIN BREAST CANCER DATASET

Binti Fitrothul Khasanah[1)], Dimas Wahyu Saputro[2)], Viyonisa Syafa Sabila[3)], Justin Tigor Hasonangan S[4)], Marhanny Zahra N[5)].

Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera

binti.120450097@student.itera.ac.id [1)], *dimas.120450081*@student.itera.ac.id[2)], *viyonisa.120450027@student.itera.ac.id* [3)], justin.120450061@student.itera.ac.id [4)], marhanny.120450017@student.itera.ac.id [5]

Abstrak

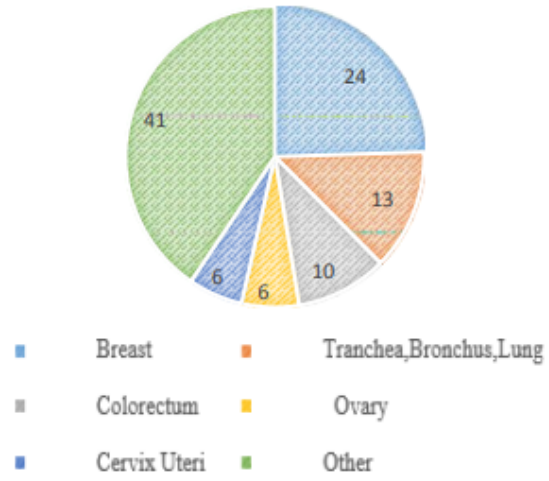
Kanker merupakan salah satu penyakit yang mematikan di dunia. Pada tahun 2018 ada 9,6 juta orang meninggal karena kanker. Dalam catatan medis penyakit kanker payudara adalah salah satu hal yang sensitif dan endemik. Kanker payudara merupakan salah satu penyebab utama kematian wanita di dunia. Kanker payudara membunuh satu diantara sebelas wanita seluruh dunia. "Deteksi dini sama dengan peningkatan peluang hidup," kutipan pepatah kanker yang terkenal. Deteksi dini penting untuk mencegah dan menurunkan resiko penyakit kanker payudara. Telah diprediksi bahwa kanker payudara adalah jenis kanker yang menyerang salah satu masalah paling signifikan yang dihadapi manusia dalam beberapa waktu belakangan ini. Deteksi kanker yang akurat dapat menyelamatkan hidup jutaan nyawa. Teknologi yang efektif untuk mendiagnosis kanker membantu pelayanan kesehatan merawat pasien secara cepat dan akurat. Penelitian ini dilakukan untuk mengkategorikan kanker payudara jinak atau ganas, menggunakan *Wisconsin Diagnosis Breast Cancer* (WDBC) dataset. *Support Vector Machine* (SVM) adalah salah satu metode dalam *supervised learning* yang digunakan untuk pengklasifikasian. Penelitian menunjukkan bahwa kinerja SVM bagus dengan tingkat akurasi 98 persen.

Kata Kunci : Deteksi Dini, Kanker Payudara, *Support Vector Machine* (SVM), Wisconsin Diagnosis Breast Cancer (WDBC).

1. Pendahuluan

Kanker payudara merupakan salah satu penyakit yang paling sering menyerang wanita di seluruh dunia. Kanker payudara menjadi penyebab umum terbanyak kematian pada wanita. Untuk mendeteksi kanker payudara secara tradisional menggunakan X-ray mamografi. Aspirasi Jarum Halus (AJH) dilakukan untuk mengambil sampel cairan atau jaringan pada kanker, AJH dilakukan jika menemukan benjolan yang mencurigakan. Rata-rata

keakurasiannya adalah 90% maka dari itu perlu pendekatan alternatif untuk mendeteksi kanker. Untuk mengurangi data yang salah, data mining sangat cocok digunakan. Deteksi dini tentang kanker dapat meningkatkan harapan hidup sebesar 98%, pada Gambar 1 menunjukkan keganasan penyakit dan kanker payudara memimpin dengan 24%.



Gambar 1 Penyakit di Dunia

Artificial intelligence (AI) dapat digunakan untuk mendeteksi dan mendiagnosis kanker secara akurat. Penggabungan *Artificial intelligence* (AI) dengan *Machine Learning* (ML) memungkinkan mendapat prediksi dan akurasi yang lebih baik. Contohnya untuk mengevaluasi apakah pasien kanker memerlukan operasi atau tidak berdasarkan hasil biopsi deteksi kanker payudara. Menurut *World Health Organization* (WHO) penyakit kanker menjadi kasus terbanyak yang diderita oleh kaum wanita. Penelitian ini bertujuan untuk mendapatkan nilai optimal yang dihasilkan saat diberlakukan metode *Support Vector Machine.*

Penelitian ini menggunakan Metode Support Vector Machine untuk mengklasifikasikan kanker payudara jinak atau ganas. Metode Support Vector Machine merupakan model klasifikasi yang bekerja dengan prinsip Structural Risk Minimization (SRM) yang bertujuan untuk menemukan hyperplane terbaik yang memisahkan dua buah class pada input space dan menghasilkan sebuah model. Model pengklasifikasian yang didapat akan digunakan untuk memprediksi nilai Optimumnya. Sehingga dari hasil tersebut bisa diketahui termasuk ke jenis yang mana kankernya jinak atau ganas.

2. Metode

a. Deskripsi Data Set

Wisconsin Breast Cancer (WBC) adalah dataset kanker payudara yang diambil dari *UCI Machine Learning Repository* [2]. Nilai-nilai dari variabel dihitung dari gambar digital aspirasi jarum halus (FNA) dari massa payudara. Data set terdiri dari 569 baris, dan 33 kolom. Lima data teratas dapat dilihat pada gambar 1.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

5 rows × 33 columns

Gambar 1 Data yang digunakan

b. Proses Pembersihan Data

Data cleaning adalah suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Menggunakan potongan baris pada gambar 2, ditemukan bahwa terdapat satu kolom yang terdapat NaN pada seluruh baris, yaitu "Unnamed: 32". Oleh karena itu, kolom tersebut dihapus.

```
In [7]:
# checking for null values
df.isnull().sum()
```

Gambar 2 Mengetahui data yang NaN

Selanjutnya, "kolom id" turut dihapus karena bukan suatu hal yang penting untuk melakukan klasifikasi pada kali ini.

c. Praproses Data

Kolom "diagnosis" merupakan kolom yang akan menjadi variabel Y yang akan digunakan sebagai hasil dari klasifikasi. Menggunakan potongan baris pada gambar 3, terlihat bahwa pada kolom "diagnosis" terdiri dari 2 nilai, yaitu Diagnosis (M = malignant, B = benign).

```
In [18]:
# counts of unique rows in the 'diagnosis' column
df['diagnosis'].value_counts()

Out[18]:
B    357
M    212
Name: diagnosis, dtype: int64
```

Gambar 3 Menghitung nilai unik pada kolom Diagnosis

Selanjutnya, dilakukan mapping atau pengubahan nama nilai, yaitu "B: menjadi 0, dan "M" menjadi 1 menggunakan potongan kode pada gambar 4.

```
In [19]:
# mapping categorical values to numerical values
df['diagnosis']=df['diagnosis'].map({'B':0,'M':1})

In [20]:
df['diagnosis'].value_counts()

Out[20]:
0    357
1    212
Name: diagnosis, dtype: int64
```

Gambar 4 Mapping nilai

Kemudian, tentukan *data frame* X dan *data frame* y. *Data frame* X merupakan seluruh kolom kecuali kolom "diagnosis", dan *data frame* y merupakan kolom "diagnosis". Karena data yang digunakan merupakan data numerik, maka data langsung bisa digunakan. Selanjutnya, membagi dataset menjadi data train dan data test, dengan proporsi 0.8 dan 0.2, seperti yang terlihat pada gambar 5.

```
In [21]:
from sklearn.model_selection import train_test_split

# splitting data
X_train, X_test, y_train, y_test = train_test_split(
                df.drop('diagnosis', axis=1),
                df['diagnosis'],
                test_size=0.2,
                random_state=42)

print("Shape of training set:", X_train.shape)
print("Shape of test set:", X_test.shape)

Shape of training set: (455, 30)
Shape of test set: (114, 30)
```

Gambar 5 Data latih dan data uji

Penskalaan nilai merupakan langkah penting dalam pemodelan algoritma dengan dataset. Selanjutnya, seperti yang terlihat pada gambar 6 dilakukan normalisasi data dengan menggunakan *StandardScaler*.

```
In [22]:
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_test = ss.fit_transform(X_test)
```

Gambar 6 Penskalaan Nilai

d. Modeling Data

Support Vector Machine (SVM) merupakan salah satu metode dalam supervised learning yang biasanya digunakan untuk

klasifikasi (seperti *Support Vector Classification*) dan regresi (*Support Vector Regression*). Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. SVM digunakan untuk mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas.

Hal yang dilakukan untuk *data set* pada artikel ini, adalah Untuk melakukan modeling, gunakan *library sklearn.svm*, kemudian *import SVC,* seperti pada gambar 7.

```
In [35]:
from sklearn.svm import SVC

svc_model = SVC(kernel="rbf")
svc_model.fit(X_train, y_train)
predictions5 = svc_model.predict(X_test)
```

Gambar 7 Kode untuk melakukan modeling

3. Hasil dan Pembahasan

Saat melakukan modeling, digunakan kernel Radial Basis Function (RBF/Gaussian). Kernel RBF atau juga disebut kernel Gaussian adalah konsep kernel yang paling banyak digunakan untuk memecahkan masalah klasifikasi data yang tidak dapat dipisahkan secara linear. Kernel ini dikenal memiliki performa yang baik dengan parameter tertentu, dan hasil dari pelatihan memiliki nilai error yang kecil dibandingkan dengan kernel lainnya. Berdasarkan perhitungan, didapatkan akurasi sebesar 98%. Nilai lain seperti *Confussion Matrix*, *precision, recall, f1-score, dan support* dapat dilihat pada gambar 8.

```
[ ]  1  from sklearn.metrics import confusion_matrix, classification_report
     2  print("Confusion Matrix: \n", confusion_matrix(y_test, predictionSVM))
     3  print("\n")
     4  print(classification_report(y_test, predictionSVM))
```

```
Confusion Matrix:
 [[71  0]
 [ 2 41]]


              precision    recall  f1-score   support

           0       0.97      1.00      0.99        71
           1       1.00      0.95      0.98        43

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114
```

Gambar 8 Laporan Klasifikasi dan Confusion Matrix

4. Kesimpulan

Kanker payudara merupakan penyakit yang sangat parah yang membunuh banyak wanita di seluruh dunia. Dalam kesehatan, diagnosis dan pengecekan kanker payudara cukup penting. Karena hal itu, dibuatlah model yang dapat melakukan klasifikasi yang dapat memprediksi apakah kanker payudara berada pada tingkat ganas atau jinak menggunakan metode *Support Vector Machine*. Pada pengujian, SVM mencapai akurasi klasifikasi sebesar 98 persen.

Referensi

[1] R. Anuradha, "Support Vector Machine Classifier for Prediction of Breast Malignancy using Wisconsin Breast Cancer Dataset," ASIAN JOURNAL OF CONVERGENCE IN TECHNOLOGY, vol. 7, no. 3. Asian Journal of Convergence in Technology, pp. 57–60, Dec. 20, 2021. doi: 10.33130/ajct.2021v07i03.010.

[2] Pisner, D. A., & Schnyer, D. M. (2019). Support vector machine. Machine Learning: Methods and Applications to Brain Disorders, 101–121. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Lampiran

Kode Pemrograman, terlampir pada .zip