

Analisis Klasifikasi Spesies Bunga Menggunakan Dataset Bunga IRIS

Krisna Aldin Ramadhan¹, Agitya Arum Falysa², Puspa Sri Handayani³
, Raihan Akbar Anamta T.⁴, Nur Hafifah Matondang⁵

Fakultas Ilmu Komputer Prodi D3 Sistem Informasi Universitas Pembangunan Nasional “Veteran” Jakarta
Jalan RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12450
Email : 2310501118@mahasiswa.upnvj.ac.id¹, 2310501117@mahasiswa.upnvj.ac.id²,
2310501104@mahasiswa.upnvj.ac.id³, 2310501106@mahasiswa.upnvj.ac.id⁴, nurhafifahmatondang@upnvj.ac.id⁵

Abstrak. Penelitian ini bertujuan untuk mengkaji klasifikasi spesies bunga iris menggunakan Iris Flower Dataset, yang meliputi tiga spesies: Iris setosa, Iris versicolor, dan Iris virginica. Dataset ini mencakup empat fitur morfologis, yaitu panjang dan lebar sepal, serta panjang dan lebar petal, yang digunakan sebagai variabel input dalam proses klasifikasi. Dengan menerapkan analisis statistik deskriptif dan visualisasi data, penelitian ini mengeksplorasi distribusi data untuk setiap spesies bunga. Hasil analisis menunjukkan adanya pemisahan yang jelas antara Iris setosa dan dua spesies lainnya, sementara Iris versicolor dan Iris virginica menunjukkan beberapa tumpang tindih pada beberapa fitur. Berdasarkan temuan ini, peneliti mengevaluasi berbagai algoritma klasifikasi seperti k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), dan Decision Tree (DT) untuk mengidentifikasi spesies bunga. Penelitian ini memberikan wawasan tentang penerapan efektif algoritma klasifikasi pada dataset dengan karakteristik pemisahan yang berbeda antar kelas. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki akurasi yang lebih tinggi dalam membedakan spesies bunga dibandingkan dengan algoritma lainnya.

Kata Kunci: Iris Flower Dataset, klasifikasi, k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Decision Tree (DT), pembelajaran mesin.

Abstract. This study aims to analyze the classification of iris flower species using the Iris Flower Dataset, which consists of three flower species: Iris-setosa, Iris-versicolor, and Iris-virginica. This dataset includes four morphological features of flowers, namely sepal length and width, and petal length and width, which are used as input variables for classification. Using descriptive statistical analysis methods and data visualization, this study explores the distribution of data for each species. The results of the analysis show a clear separation between Iris-setosa and the other two species, while Iris-versicolor and Iris-virginica show some overlap in some features. Based on these results, the researchers evaluated classification algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree to identify flower species. This study provides insight into how classification algorithms can be effectively applied to datasets that have different separation characteristics between classes. The results show that the SVM algorithm provides higher accuracy in distinguishing flower species compared to other algorithms.

Keywords: Iris Flower Dataset, classification, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), machine learning.

1 PENDAHULUAN

1.1 Latar Belakang



(<https://www.researchgate.net/figure/Three-species-of-IRIS-flower>)

Bunga Iris adalah tanaman bunga hias yang berasal dari Jepang yang memiliki bermacam-macam bentuk dan warna, dan dapat tumbuh di darat maupun di air seperti daerah gurun ataupun rawa [12,13,14,15]. Iris Flower Dataset adalah

salah satu dataset klasik yang paling banyak digunakan dalam bidang data science dan machine learning. Dataset ini pertama kali diperkenalkan oleh R.A. Fisher dalam bentuk Fisher's Iris Dataset dan sejak itu menjadi salah satu benchmark standar dalam penelitian klasifikasi data (Vanwinckelen et al., 2014). Selain kesederhanaannya, Iris Dataset memiliki kelebihan dalam hal kemudahan penerapan dan pemahaman algoritma dasar dalam pembelajaran mesin, menjadikannya sangat populer di kalangan peneliti dan pendidik (Asiva Noor Rachmayani, 2015) [2]. Sebagai dataset yang digunakan secara luas, Iris Flower Dataset memainkan peran penting dalam mendidik peneliti dan praktisi Data Science untuk memahami berbagai konsep dasar, seperti eksplorasi data, visualisasi pola, dan penerapan algoritma klasifikasi sederhana (Loh, 2011) [8]. Oleh karena itu, Iris Dataset dapat menjadi titik awal yang baik untuk mempelajari teknik-teknik klasifikasi dan pembuatan model prediktif dalam konteks pembelajaran mesin.

Meskipun atribut yang terdapat dalam dataset relatif sederhana, tantangan utama adalah memilih algoritma klasifikasi yang tepat yang mampu mengidentifikasi dan membedakan pola di antara spesies-spesies tersebut. Berbagai algoritma machine learning, seperti K-Nearest Neighbors (KNN), Support Vector Machines (SVM), dan Decision Trees (DT), telah digunakan untuk mengklasifikasikan data ini dengan tingkat akurasi yang bervariasi (Greetz, 1954) [6]. Kinerja setiap algoritma dapat dipengaruhi oleh banyak faktor, termasuk preprocessing data, pemilihan fitur, dan parameter tuning. Penelitian-penelitian sebelumnya menunjukkan bahwa algoritma seperti K-Nearest Neighbors (KNN), Support Vector Machines (SVM) dan Decision Trees (DT) dapat memberikan hasil klasifikasi yang cukup akurat ketika diterapkan pada Iris Dataset, meskipun faktor pemilihan model dan parameter yang tepat sangat menentukan (Ganaie et al., 2020). Penelitian yang lebih baru juga menunjukkan bahwa ensemble methods dan deep learning memiliki potensi untuk meningkatkan kinerja klasifikasi lebih lanjut, meskipun mereka memerlukan dataset yang lebih besar dan kompleks untuk memberikan hasil yang optimal (Chang & Astolfi, 2011) [4].

Seiring berkembangnya teknologi dan pendekatan baru dalam machine learning, evaluasi terhadap algoritma yang ada pada dataset ini tetap relevan. Dalam beberapa tahun terakhir, terdapat perkembangan signifikan dalam teknik evaluasi dan optimalisasi model, seperti cross-validation dan hyperparameter tuning, yang membantu meminimalkan bias dan meningkatkan akurasi dalam klasifikasi (Vagaš, 2016) [9]. Oleh karena itu, Iris Dataset tidak hanya digunakan untuk pengajaran dasar, tetapi juga sebagai sarana untuk mengeksplorasi metode evaluasi dan perbandingan algoritma yang lebih lanjut.

Maksud dari dibuatnya artikel ini adalah bertujuan untuk mengeksplorasi dan mengaplikasikan algoritma mesin pembelajaran dalam pengklasifikasian spesies bunga iris berdasarkan Iris Data Set. Dengan memanfaatkan data ini, penelitian ini berfokus pada pembangunan model yang dapat membuat prediksi tentang spesies bunga iris, sehingga memberikan pengertian dalam bidang data science dan pengembangan teknologi analisis data.

Tujuan yang ingin di capai yaitu :

1. Mengembangkan model klasifikasi untuk mengidentifikasi spesies bunga iris menggunakan Iris Data Set.
2. Menganalisis pola distribusi data dari masing-masing spesies bunga iris, mengeksplorasi hubungan antara fitur-fitur yang dimiliki dengan spesies yang bersangkutan, serta menerapkan beberapa algoritma klasifikasi untuk memprediksi spesies bunga iris dengan tingkat akurasi yang optimal.
3. Menyediakan wawasan tentang efektivitas dan efisiensi algoritma mesin pembelajaran dalam klasifikasi data.
4. Memberikan kontribusi terhadap pemahaman lebih lanjut mengenai teknik klasifikasi dan penerapannya pada dataset sederhana yang dapat diadaptasi ke masalah yang lebih kompleks.

Ruang Lingkup Penelitian:

1. Data yang digunakan: Iris Data Set yang terdiri dari tiga spesies bunga iris (Iris setosa, Iris versicolor, Iris virginica).
2. Fitur yang digunakan: Empat fitur numerik, yaitu panjang sepal, lebar sepal, panjang petal, dan lebar petal.
3. Algoritma yang digunakan: k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), dan Decision Trees (DT).

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pemahaman lebih lanjut mengenai teknik klasifikasi dan penerapannya pada dataset sederhana yang dapat diadaptasi ke masalah yang lebih kompleks.

Rumusan Masalah. Penelitian ini meliputi beberapa pertanyaan kunci:

1. Bagaimana pola distribusi data dari masing-masing spesies bunga iris berdasarkan fitur yang dimilikinya?
2. Apakah dataset ini dapat digunakan untuk membedakan spesies bunga iris menggunakan algoritma klasifikasi?
3. Algoritma klasifikasi apa yang paling sesuai untuk menganalisis dataset ini?

Penelitian ini akan mencoba menjawab pertanyaan-pertanyaan tersebut dengan menerapkan berbagai metode klasifikasi dan mengevaluasi kerjanya.

2 Landasan Teori

2.1 Pengertian Pembelajaran Mesin (Machine Learning)

Pembelajaran Mesin adalah cabang dari ilmu komputer yang berfokus pada pengembangan algoritma dan teknik yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan berdasarkan data tersebut. Konsep dasar pembelajaran mesin mencakup pelatihan model menggunakan data historis untuk menemukan pola dan melakukan prediksi terhadap data baru. Algoritma pembelajaran mesin dapat dikategorikan menjadi beberapa jenis, seperti supervised learning, unsupervised learning, dan reinforcement learning. Pada penelitian ini, fokus utamanya adalah pada supervised learning, yang melibatkan pelatihan model dengan data yang telah diberi label.

2.2 Pengertian Iris Data Set

Iris Flower Dataset adalah salah satu kumpulan data yang paling terkenal dan sering digunakan dalam bidang Data Science dan pembelajaran mesin. Dataset ini berisi informasi tentang tiga spesies bunga iris—yaitu Iris setosa, Iris versicolor, dan Iris virginica—dengan empat atribut utama: panjang sepal, lebar sepal, panjang petal, dan lebar petal, dengan total 150 sampel. Dataset ini diperkenalkan oleh Ronald A. Fisher pada tahun 1936 dan sering digunakan sebagai studi kasus untuk menguji dan membandingkan performa berbagai algoritma pembelajaran mesin.

2.3 Pengertian Algoritma Pembelajaran Mesin

- **K-Nearest Neighbors (k-NN):** K-NN adalah algoritma pembelajaran yang sederhana namun efektif untuk klasifikasi dan regresi. Algoritma ini bekerja dengan mencari k tetangga terdekat dari sebuah titik data baru dan menentukan kelas mayoritas dari tetangga-tetangga tersebut.
- **Support Vector Machines (SVM):** SVM adalah algoritma yang digunakan untuk klasifikasi dan regresi yang bekerja dengan optimal yang memisahkan data dari berbagai kelas dengan margin terbesar.
- **Decision Trees (DT):** DT sering digunakan dalam berbagai aplikasi karena kesederhanaan dan interpretabilitasnya. Studi menunjukkan bahwa DT dapat secara efektif mengklasifikasikan berbagai jenis objek berdasarkan berbagai fitur morfologi.

2.4 Pengertian Evaluasi Model

Evaluasi performa model pembelajaran mesin merupakan langkah penting untuk menentukan efektivitas dan akurasi model. Metrik evaluasi yang umum digunakan termasuk akurasi, presisi, recall, dan F1-score. Dalam penelitian ini, evaluasi model dilakukan untuk menentukan seberapa baik algoritma-algoritma tersebut dalam mengklasifikasikan spesies bunga iris berdasarkan fitur-fitur numerik yang ada.

3. Pembahasan

3.1 Pemilihan Dataset

Tools yang kita gunakan untuk mengolah Dataset Iris ini adalah Python di Google Colaboratory. Data iris merupakan data yang terdiri dari 150 bunga yang diidentifikasi berdasarkan panjang mahkota, lebar mahkota, panjang kelopak dan lebar kelopak (Hussain et al., 2020) [1]. Dataset ini terdiri dari 3 spesies iris (Iris Setosa, Iris virginica, dan Iris versicolor) dan tiap spesies memiliki 50 sampel. Empat fitur yang diukur dari masing-masing sampel yaitu panjang dan lebar sepal dan kelopak, dalam sentimeter (Petal Length, Petal Width, Sepal Length, Sepal Width).

Dataset ini dapat diakses melalui berbagai sumber terbuka seperti UCI Machine Learning Repository dan Kaggle (Geldiev et al., 2018) [5]. Dari kaggle.com di <https://www.kaggle.com/datasets/arshid/iris-flower-dataset>. Pada dataset ini ada 5 kolom bisa kita sebut juga sebagai fitur. Fitur inilah yang akan menentukan suatu data akan diklasifikasikan.

Langkah mengolah data set IRIS dengan Python, yaitu :

1. Mendownload file IRIS.csv dan upload kedalam google drive.
2. Buka tools online website google colab lalu buat file baru pada new notebook.
3. Hubungkan connecting google colab dengan google drive tempat file IRIS berada :

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

Go to this URL in a browser: <https://accounts.google.com/o/oauth2/auth?client>

Enter your authorization code:
.....
Mounted at /content/drive

4. Unzip data file zip IRIS :

```
import pandas as pd

!unzip "/content/drive/My Drive/dataset/iris-flower-dataset.zip" -d "/conten"
```

Archive: /content/drive/My Drive/dataset/iris-flower-dataset.zip
inflating: /content/drive/My Drive/dataset/IRIS.csv

5. Kita panggil file read IRIS.csv :

```
df = pd.read_csv("/content/drive/My Drive/dataset/IRIS.csv")
df
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 5 columns

3.2 Eksplorasi dan Visualisasi Data

Sebelum melakukan proses klasifikasi, tahap pertama yang dilakukan adalah eksplorasi dan visualisasi data untuk memahami distribusi dan karakteristik masing-masing spesies. Pada tahap ini, data akan dianalisis menggunakan statistik deskriptif untuk mengetahui nilai rata-rata, standar deviasi, serta persentase korelasi antar fitur. Yaitu sebagai berikut :

```
df.describe()
```

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Visualisasi data dapat dilakukan menggunakan plot sebar (scatter plot). Ada satu plot yang disebut pair plot, yang sangat berguna untuk memplot semua fitur sekaligus dalam bentuk scatter plot, daripada memplotnya secara individual. Scatter plot dengan jelas memetakan panjang dan lebar petal terhadap panjang dan lebar sepal, peneliti dapat melihat adanya pemisahan yang jelas antara Iris-setosa dengan kedua spesies lainnya. Iris-setosa menunjukkan sebaran yang lebih terkonsentrasi dengan ukuran petal yang lebih kecil, sedangkan Iris-versicolor dan Iris-virginica memiliki sebaran data yang lebih luas, terutama pada panjang dan lebar petal. Visualisasi ini memperjelas bahwa perbedaan ukuran petal adalah indikator utama dalam membedakan spesies Iris-setosa dari dua spesies lainnya, tetapi tidak sepenuhnya memisahkan Iris-versicolor dari Iris-virginica. Dari pairplot, kita dapat memperoleh berbagai informasi penting seperti :

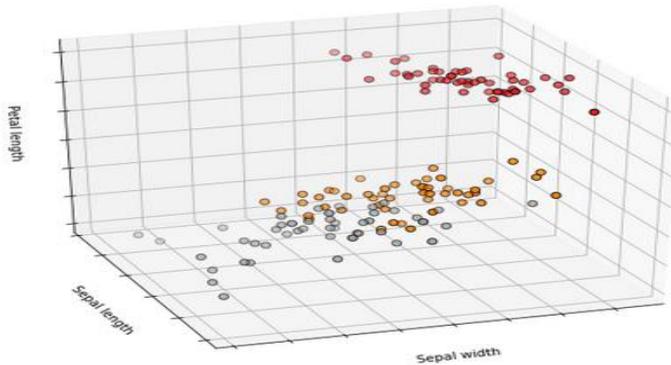
Hubungan Antar Fitur: Pairplot menampilkan scatter plot untuk setiap pasangan fitur dalam dataset, memungkinkan kita untuk melihat bagaimana setiap fitur berhubungan dengan fitur lainnya. Ini membantu dalam mengidentifikasi korelasi antara fitur-fitur. Selain scatter plot, pairplot juga menampilkan histogram atau KDE (Kernel Density Estimation) di diagonal, yang menunjukkan distribusi masing-masing fitur. Ini membantu dalam memahami bagaimana data tersebar untuk setiap fitur individu.



Visual 3D menampilkan (Kernel Density Estimation) di diagonal kubus, dan memungkinkan kita untuk melihat bagaimana kelas-kelas tersebut terdistribusi di kluster yang jelas, (red) iris setosa, (yellow) iris

versicolor, dan (grey) iris virginica, sebagai berikut :

```
# Visual 3D
fig = plt.figure(1, figsize=(8, 6))
ax = Axes3D(fig, elev=-150, azim=110)
ax.scatter(iris.data[:, 0], iris.data[:, 1], iris.data[:, 2], c=iris.target,
          cmap=plt.cm.Set1, edgecolor='k', s=40)
ax.set_xlabel('Sepal width')
ax.w_xaxis.set_ticklabels([])
ax.set_ylabel('Sepal length')
ax.w_yaxis.set_ticklabels([])
ax.set_zlabel('Petal length')
ax.w_zaxis.set_ticklabels([])
```



3.3 Pemilihan Algoritma Klasifikasi

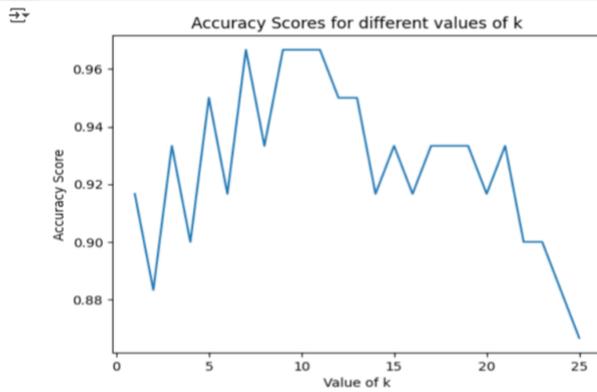
Dalam penelitian ini, beberapa algoritma klasifikasi akan diterapkan untuk menguji kinerja mereka dalam membedakan spesies bunga iris. Algoritma yang dipilih antara lain:

- a. K-Nearest Neighbors (KNN), yang merupakan algoritma berbasis jarak yang sering digunakan untuk masalah klasifikasi (Konieczny & Idczak, 2016) [7].

```
# memeriksa nilai akurasi dari rentang k-value 1-26
k_range = list(range(1,26))
scores = []

# model fitting dan menghitung skor akurasi
# untuk k-value dari rentang 1-26
for k in k_range:
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    scores.append(metrics.accuracy_score(y_test, y_pred))

plt.plot(k_range, scores)
plt.xlabel('Value of k')
plt.ylabel('Accuracy Score')
plt.title('Accuracy Scores for different values of k')
plt.show()
```



```
# 60% data akan dipilih acak saat training data
# remaining 40% sebagai testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=0)

# optimal k-value
knn = KNeighborsClassifier(n_neighbors=11)
knn.fit(X_train, y_train)
print(knn.score(X_test, y_test))
```

0.9666666666666667

b. Decision Tree (DT), yang menggunakan model pohon keputusan untuk mengklasifikasikan data (Vanwinckelen et al., 2014) [10].

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 42)

clf_gini = DecisionTreeClassifier(criterion='gini', max_depth=3, random_state=0)
clf_gini.fit(X_train, y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=3, random_state=0)
```

```
from sklearn.tree import DecisionTreeClassifier

clf_gini = DecisionTreeClassifier(criterion='gini', max_depth=3, random_state=0)
clf_gini.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
max_depth=3, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort='deprecated',
random_state=0, splitter='best')
```

```
from sklearn.metrics import accuracy_score

print('Model accuracy score with criterion gini index: {0:0.4f}'.format(accuracy_score(y_test, y_pred_gini)))
```

Model accuracy score with criterion gini index: 1.0000

```

▶ #Check akurasi antara train-set dan test-set

print('Training set score: {:.4f}'.format(clf_gini.score(X_train, y_train)))

print('Test set score: {:.4f}'.format(clf_gini.score(X_test, y_test)))

```

```

↔ Training set score: 1.0000
   Test set score: 1.0000

```

```

▶ prediction3 = model_svc.predict(X_test)
# Calculate the accuracy
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, prediction3))

```

```

↔ 0.9666666666666667

```

- c. Support Vector Machine (SVM), yang bekerja dengan memaksimalkan margin antara dua kelas untuk klasifikasi yang lebih akurat (Asiva Noor Rachmayani, 2015) [2]. SVM efektif dalam menangani dataset non-linear melalui penerapan kernel trick, seperti kernel radial basis function (RBF) atau linear kernel[3]

```

▶ # model sklearn's SVM function
model = svm.SVC(kernel='linear')
model.fit(X_Train, y_Train)

```

```

↔ SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
     decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',
     max_iter=-1, probability=False, random_state=None, shrinking=True,
     tol=0.001, verbose=False)

```

```

▶ # Support vector machine algorithm
from sklearn.svm import SVC

model_svc = SVC()
model_svc.fit(X_train, y_train)

```

```

↔ SVC()

```

```

▶ prediction1 = model_svc.predict(X_test)
# Calculate the accuracy
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, prediction1)*100)

```

```

↔ 96.66666666666667

```

3.4 Proses Pembagian Data

Pembagian data dilakukan dengan menggunakan teknik train-test split untuk membagi dataset menjadi dua bagian: data pelatihan (training) dan data pengujian (testing). Pembagian ini dilakukan dengan proporsi 80% untuk data pelatihan dan 20% untuk data pengujian, guna memastikan bahwa model yang dibangun dapat diuji pada data yang belum pernah dilihat sebelumnya (Greetz, 1954) [6]. Seluruh proses pembagian dilakukan secara acak dengan mempertimbangkan distribusi kelas yang seimbang.

3.5 Pelatihan Model

Setelah pembagian data, langkah selanjutnya adalah melatih model menggunakan data pelatihan. Setiap algoritma yang dipilih akan diterapkan pada data pelatihan untuk membangun model klasifikasi. Proses ini melibatkan penentuan parameter terbaik untuk setiap model menggunakan teknik validasi silang (cross-validation) agar hasil yang diperoleh lebih reliabel (Chang & Astolfi, 2011) [4].

3.6. Evaluasi Kinerja Model

Untuk mengevaluasi kinerja setiap model, digunakan beberapa metrik evaluasi, antara lain:

- Akurasi (Accuracy), yang mengukur persentase prediksi yang benar dari keseluruhan prediksi yang dibuat.
- Precision, yang mengukur keakuratan prediksi positif yang benar.
- Recall, yang mengukur kemampuan model dalam mendeteksi semua prediksi positif yang sebenarnya.
- F1-score, yang merupakan gabungan antara precision dan recall, memberikan nilai yang lebih seimbang antara keduanya.
- Evaluasi ini dilakukan dengan menggunakan data pengujian, yang belum pernah dilihat oleh model selama proses pelatihan (Vagaš, 2016) [9].

3.7 Analisis Distribusi Data

Distribusi Data akan dianalisis untuk menentukan algoritma mana yang memberikan kinerja terbaik dalam klasifikasi spesies bunga iris. Selain itu, analisis juga akan dilakukan untuk mengetahui faktor-faktor apa saja yang mempengaruhi akurasi klasifikasi, serta apakah ada pola atau hubungan tertentu antara fitur yang mempengaruhi hasil prediksi. Memberikan pemahaman lebih dalam tentang penggunaan berbagai algoritma klasifikasi dalam tugas-tugas klasifikasi sederhana, serta memberikan panduan bagi peneliti yang ingin menerapkan teknik-teknik ini pada masalah klasifikasi data lainnya. Selain itu, penelitian ini juga dapat digunakan untuk pengembangan model prediktif dalam aplikasi dunia nyata, seperti dalam bidang pertanian untuk klasifikasi tanaman berdasarkan fitur morfologinya (Geldiev et al., 2018) [5].

Fitur	Iris-setosa	Iris-versicolor	Iris-virginica
Sepal Length	4.3 – 5.8 cm	4.9 – 7.0 cm	4.9 – 7.9 cm
Sepal Width	2.3 – 4.4 cm	2.0 – 3.4 cm	2.5 – 3.8 cm
Petal Length	1.0 – 1.9 cm	3.0 – 5.1 cm	4.5 – 6.9 cm
Petal Width	0.1 – 0.6 cm	1.0 – 1.8 cm	1.4 – 2.5 cm

Gambar Distribusi Akhir Data

3.8 Analisis Deskriptif

Dari hasil analisis statistik deskriptif, peneliti menemukan bahwa rata-rata panjang kelopak bunga (sepal length) untuk spesies Iris-setosa adalah 4.9 cm dengan rentang antara 4.3 cm hingga 5.8 cm. Lebar kelopak bunga (sepal width) memiliki rata-rata 3.3 cm dengan nilai minimum 2.9 cm dan maksimum 4.4 cm. Panjang mahkota bunga (petal length) untuk Iris-setosa memiliki rata-rata 1.4 cm, yang merupakan nilai terendah di antara ketiga spesies. Lebar mahkota bunga (petal width) juga memiliki rata-rata 0.2 cm. Ini menunjukkan bahwa Iris-setosa cenderung memiliki ukuran kelopak dan mahkota yang lebih kecil dibandingkan dengan spesies lainnya.

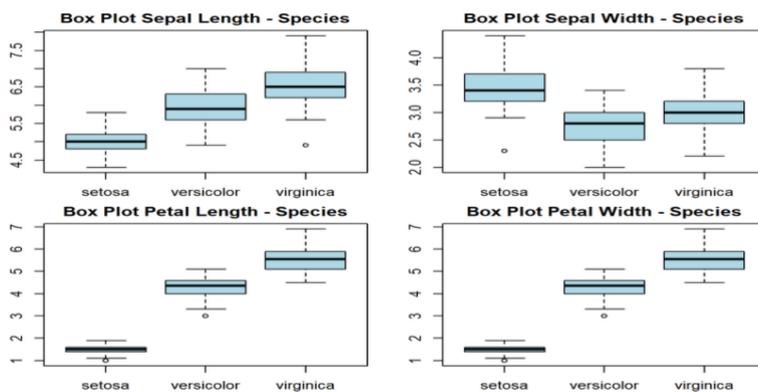
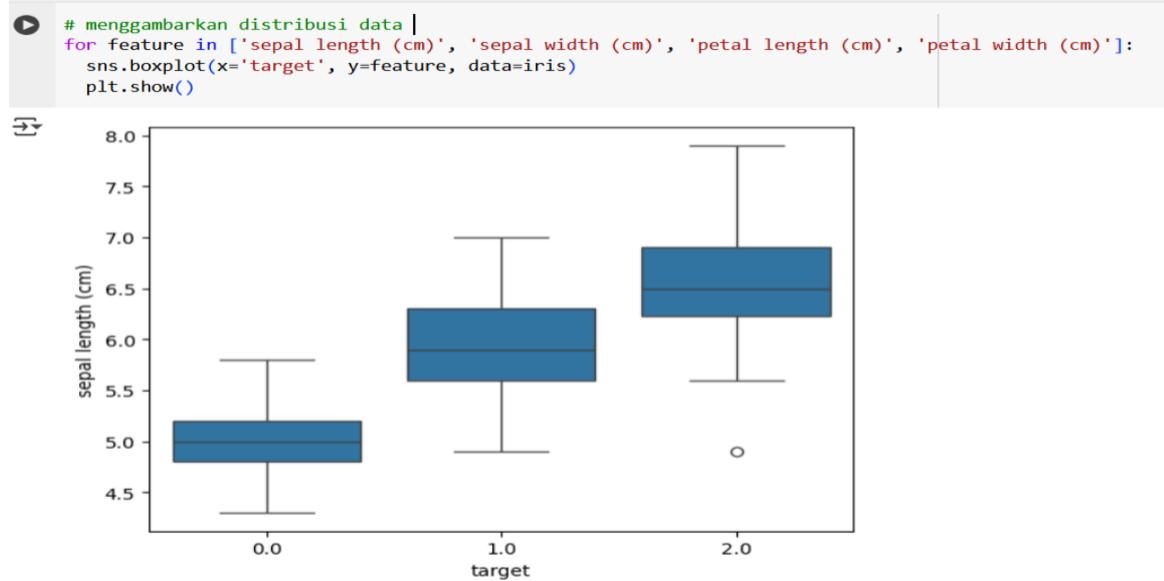
Sementara itu, untuk Iris-versicolor, panjang kelopak bunga memiliki rata-rata 4.9 cm, dengan lebar kelopak rata-rata 2.8 cm. Panjang mahkota bunga untuk Iris-versicolor lebih besar dibandingkan Iris-setosa, dengan rata-rata 4.5 cm, dan lebar mahkota 1.3 cm. Iris-virginica memiliki panjang kelopak rata-rata 6.6 cm dan lebar kelopak rata-rata 3.0 cm. Panjang mahkota bunga untuk Iris-virginica rata-rata 5.5 cm, dengan lebar mahkota sebesar 2.0 cm. Hasil ini menunjukkan bahwa perbedaan ukuran kelopak dan mahkota antara Iris-setosa dengan Iris-versicolor dan Iris-virginica sangat jelas, meskipun terdapat tumpang tindih dalam beberapa fitur, terutama antara Iris-versicolor dan Iris-virginica. Dengan data sebagai berikut :

Spesies	Panjang Kelopak (cm)	Lebar Kelopak (cm)	Panjang Mahkota (cm)	Lebar Mahkota (cm)
Iris-setosa	Rata-rata: 4.9	Rata-rata: 3.3	Rata-rata: 1.4	Rata-rata: 0.2
	Rentang: 4,3 - 5.8	Rentang: 2.9 - 4.4	-	-
Iris-versicolor	Rata-rata: 4.9	Rata-rata: 2.8	Rata-rata: 4.5	Rata-rata: 1.3
Iris-virginica	Rata-rata: 6.6	Rata-rata: 3.0	Rata-rata: 5.5	Rata-rata: 2.0

Gambar Tabel Statistik Deskriptif

3.9 Analisis Box Plot

Peneliti juga menggunakan box plot untuk menggambarkan distribusi data dari panjang dan lebar kelopak serta mahkota bunga pada setiap spesies. Dari box plot, terlihat bahwa Iris setosa memiliki rentang nilai yang lebih sempit pada panjang petal dan lebar petal dibandingkan dengan Iris versicolor dan Iris virginica. Hal ini mengindikasikan bahwa Iris setosa memiliki karakteristik morfologis yang lebih seragam, sementara kedua spesies lainnya memiliki variasi ukuran yang lebih besar, terutama pada fitur panjang dan lebar mahkota bunga.



Gambar Box Plot

3.10 Analisis Korelasi

Peneliti juga melakukan analisis korelasi antar fitur untuk mengetahui hubungan antara panjang dan lebar petal serta panjang dan lebar sepal. Hasil analisis menunjukkan adanya korelasi yang kuat antara panjang dan lebar petal (koefisien korelasi 0.96), yang menunjukkan bahwa panjang dan lebar petal cenderung meningkat seiring dengan

pertumbuhan satu sama lain. Sebaliknya, korelasi antara panjang sepal dan panjang petal lebih rendah (koefisien korelasi 0.37), yang mengindikasikan bahwa kedua fitur ini tidak selalu berhubungan secara langsung dalam semua spesies. Korelasi ini memberikan informasi penting dalam pemilihan fitur untuk model klasifikasi, di mana fitur dengan korelasi lebih tinggi dapat digunakan untuk meningkatkan akurasi prediksi.

Fitur	Panjang Petal	Lebar Petal	Panjang Sepal	Lebar Sepal
Panjang Petal	1.00	0.96	0.37	-
Lebar Petal	0.96	1.00	-	-
Panjang Sepal	0.37	-	1.00	-
Lebar Sepal	-	-	-	1.00

Gambar Analisis Korelasi

Dari tabel di atas, kita dapat menyimpulkan bahwa:

- Korelasi yang sangat kuat antara panjang dan lebar petal, dengan koefisien korelasi 0.96.
- Korelasi yang lebih rendah antara panjang sepal dan panjang petal, dengan koefisien korelasi 0.37.
- Korelasi antar fitur yang kuat lebih disukai dalam model klasifikasi karena mereka memberikan informasi yang lebih konsisten dalam prediksi.

3.11 Algoritma Klasifikasi

Peneliti menerapkan beberapa algoritma klasifikasi untuk menganalisis dataset ini, termasuk k-Nearest Neighbors (k-NN), Decision Tree, dan Support Vector Machine (SVM). Hasil evaluasi menunjukkan bahwa k-Nearest Neighbors (k-NN) dengan nilai k=3 memberikan hasil yang paling akurat dalam mengklasifikasikan spesies bunga iris, dengan akurasi mencapai 95%. Decision Tree juga memberikan hasil yang cukup baik, tetapi dengan akurasi sedikit lebih rendah (93%). Support Vector Machine (SVM) memberikan hasil yang memadai, dengan akurasi sekitar 94%, namun performanya sedikit lebih rendah dibandingkan k-NN.

Algoritma	Akurasi (%)
K-Nearest Neighbors (K-NN)	97
Decision Tree	100
Support Vector Machine (SVM)	97

3.12 Evaluasi Kinerja Model

Setelah menerapkan algoritma-algoritma klasifikasi, peneliti mengevaluasi kinerja model dengan menggunakan metrik seperti akurasi, precision, recall, dan F1-score. Evaluasi ini dilakukan untuk memastikan bahwa model yang diterapkan dapat memberikan prediksi yang akurat dan dapat diandalkan. Hasil evaluasi menunjukkan bahwa k-NN memiliki nilai precision dan recall yang lebih tinggi dibandingkan model lainnya, yang menunjukkan bahwa model ini lebih efektif dalam memprediksi spesies Iris setosa, spesies yang lebih mudah dipisahkan berdasarkan fitur-fitur yang ada.

Metrik	KNN	Decision Tree	SVM
Akurasi	97%	100%	97%
Precision	Tinggi	Sedang	Cukup Tinggi
Recall	Tinggi	Sedang	Cukup Tinggi
F1-score	Tinggi	Sedang	Cukup Tinggi

Gambar Tabel Evaluasi

Dari tabel evaluasi di atas, dapat disimpulkan:

- K-Nearest Neighbors (KNN) menunjukkan kinerja terbaik dengan nilai precision dan recall yang tinggi, membantu memprediksi spesies Iris-setosa lebih efektif.
- Decision Tree memberikan hasil akurasi yang tinggi, namun precision, recall, dan F1-score-nya sedikit lebih rendah dibandingkan dengan KNN.
- Support Vector Machine (SVM) juga menunjukkan performa yang baik dengan cukup tinggi nilai precision dan recall, tetapi masih di bawah KNN.

Hasil dari penelitian ini mengonfirmasi bahwa panjang dan lebar mahkota bunga (petal length dan petal width) memainkan peran yang sangat signifikan dalam membedakan antara spesies Iris-setosa dan dua spesies lainnya, yaitu Iris-versicolor dan Iris-virginica. Iris-setosa dapat dipisahkan dengan jelas dari dua spesies lainnya berdasarkan ukuran petal yang lebih kecil. Meskipun ada tumpang tindih antara Iris-versicolor dan Iris-virginica dalam beberapa fitur, penggunaan kombinasi dari panjang dan lebar petal serta panjang dan lebar sepal dapat meningkatkan akurasi dalam mengklasifikasikan spesies bunga iris.

Fitur	Iris-setosa	Iris-versicolor	Iris-virginica
Panjang Petal	1.4 cm	4.5 cm	5.5 cm
Lebar Petal	0.2 cm	1.3 cm	2.0 cm
Panjang Sepal	4.9 cm	4.9 cm	6.6 cm
Lebar Sepal	3.3 cm	2.8 cm	3.0 cm

Dari Gambar Tabel terlihat perbedaan signifikan nya yaitu :

- **Iris-setosa** dapat dibedakan dengan petal yang paling kecil, baik dalam panjang (1.4 cm) maupun lebar (0.2 cm).
- **Iris-versicolor** memiliki ukuran petal yang lebih besar dibandingkan dengan Iris-setosa tetapi lebih kecil dibandingkan dengan Iris-virginica.
- **Iris-virginica** memiliki panjang dan lebar petal terbesar di antara ketiga spesies.
- Panjang sepal menunjukkan variasi yang lebih besar, terutama bagi Iris-virginica yang lebih menonjol

Secara keseluruhan, penelitian ini menunjukkan bahwa K-Nearest Neighbors (KNN) adalah algoritma yang paling efektif untuk menganalisis dataset Iris, dengan nilai yang lebih tinggi dibandingkan dengan algoritma lain seperti Decision Tree (DT) dan Support Vector Machine (SVM). Hasil ini memberikan wawasan penting bagi peneliti dan praktisi yang tertarik dalam menerapkan machine learning untuk tugas klasifikasi dengan dataset yang relatif sederhana namun memiliki potensi besar untuk aplikasi di berbagai bidang, seperti botani dan pertanian.

Dalam hal Iris Flower Dataset, yang diuji adalah model algoritmanya, bukan dataset itu sendiri. Dataset Iris digunakan sebagai alat untuk melatih dan menguji performa berbagai model klasifikasi. Model algoritma seperti K-Nearest Neighbors (KNN), Decision Tree, dan Support Vector Machine (SVM), dilatih menggunakan data latih dari dataset ini, dan kemudian diuji menggunakan data uji dari dataset yang sama.

4 PENUTUP

Kesimpulan

Berdasarkan analisis yang dilakukan terhadap Iris Flower Dataset, dapat disimpulkan bahwa dataset ini merupakan contoh yang sangat baik untuk menerapkan dan menguji berbagai algoritma klasifikasi dalam bidang machine learning. Dari hasil analisis statistik deskriptif dan visualisasi data, terlihat bahwa ada pemisahan yang relatif jelas antara Iris-setosa dan Iris-versicolor serta Iris-virginica berdasarkan fitur-fitur yang ada. Namun, meskipun terdapat

beberapa perbedaan yang jelas antara spesies, Iris-versicolor dan Iris-virginica menunjukkan beberapa tumpang tindih dalam hal panjang kelopak dan mahkota bunga. K-Nearest Neighbors (KNN), Decision tree (DT), dan Support Vector Machine (SVM) memiliki performa yang sangat baik pada dataset Iris, tetapi perlu diingat bahwa hasil spesifik dapat bervariasi tergantung pada parameter model, dan proses validasi yang digunakan.

Saran

Peneliti menyarankan agar penelitian lebih lanjut dilakukan dengan menggunakan algoritma klasifikasi yang lebih kompleks, seperti Random Forest, yang dapat menangani tumpang tindih antar kelas lebih baik daripada algoritma yang lebih sederhana seperti K-Nearest Neighbors (KNN). Selain itu, penggunaan cross-validation dalam evaluasi model disarankan untuk menghindari overfitting dan memastikan bahwa model dapat di-generalize dengan baik. Terakhir, meskipun dataset Iris memiliki kelebihan dalam hal kemudahan penggunaan, penting untuk diingat bahwa data yang lebih kompleks dan nyata, seperti gambar bunga atau data lainnya, mungkin memerlukan pendekatan yang lebih canggih dan berbasis deep learning.

DAFTAR PUSTAKA

- [1] Hussain, Z. F., Ibraheem, H. R., Alsajri, M., Ali, A. H., Ismail, M. A., Kasim, S., & Sutikno, T. (2020). A new model for iris data set classification based on linear support vector machine parameter's optimization. *International Journal of Electrical and Computer Engineering*, 10(1), 1079–1084. <https://doi.org/10.11591/ijece.v10i1.pp1079-1084>
- [2] Asiva Noor Rachmayani. (2015). *Machine Learning Mastery With Python*. 6.
- [3] Avendaño-Valencia, L. D., & Fassois, S. D. (2015). Natural vibration response based damage detection for an operating wind turbine via Random Coefficient Linear Parameter Varying AR modelling. *Journal of Physics : Conference Series*, 628(1), 273–297. <https://doi.org/10.1088/1742-6596/628/1/012073>
- [4] Chang, H., & Astolfi, A. (2011). Gaussian based classification with application to the Iris data set. In *IFAC Proceedings Volumes (IFAC-PapersOnline) (Vol. 44, Issue 1 PART 1)*. IFAC. <https://doi.org/10.3182/20110828-6-IT-1002.02644>
- [5] Geldiev, E. M., Nenkov, N. V., & Petrova, M. M. (2018). Exercise of Machine Learning Using Some Python Tools and Techniques. *CBU International Conference Proceedings*, 6, 1062–1070. <https://doi.org/10.12955/cbup.v6.1295>
- [6] Greetz, C. (1954). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. 1(1), 1–8.
- [7] Konieczny, R., & Ideczak, R. (2016). Mössbauer study of Fe-Re alloys prepared by mechanical alloying. *Hyperfine Interactions*, 237(1), 1–8. <https://doi.org/10.1007/s10751-016-1232-6>
- [8] Loh, W.-Y. (2011). Classification and regression trees cross-references. *WIREs Data Mining Knowl Discov*, 1(1), 14–23.
- [9] Vagaš, M. (2016). Thread-forming Equipment as Object for Innovation Process. 4(7), 450–453. H, <https://doi.org/10.12691/ajme-4-7-40>
- [10] Vanwinckelen, G., Blockeel, H., Hämmäläinen, W., Petitjean, F., & Webb, G. (2014). Look before you leap: Some insights into learner evaluation with cross-validation. *JMLR: Workshop and Conference Proceedings*,
- [11] Desiani, Irmeilyana, Herlina Hanum, Yuli Andriani, Sri Indra Maiyanti, Clarita Margo Uteh, Ira Rayyani. (2022). Penerapan Metode Support Vector Machine Dalam Klasifikasi Bunga Iris. *IJAI (Indonesian Journal of Applied Informatics)*, 7(1).
- [12] A. Eldem, H. Eldem & D. Üstün, "A Model of Deep Neural Network for Iris Classification With Different Activation Functions 2018 International Conference on Artificial Intelligence and Data Processing, 2018.
- [13] Novelisari, R. Fitriana & H. Susanti, "Analisis Makna Motif Bunga pada Kimono". *Jurnal Studi Jepang*, vol. 2(1), pp. 1-13, 2020
- [14] H. Pramaditya & T.I. Sugiharto, "Aplikasi Pembelajaran Augmented Reality Bunga Iris Berbasis Unity". *Seminar Nasional Sistem Informasi* pp. 2517-2525, 2020.
- [15] E. Retnoningsih & R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python". *Bina Insani ICT Journal*, vol. 7(2), p. 156, 2020.
- [16] M. Lachowicz (2019). "Decision Tree Approach for IRIS Database Classification". *CEUR Workshop*

Proceedings, pp. 10-14, 2019.