

Mini Portofolio

Data Science

Presented by
M.Harris Mulya Bahari



@Dibimbing.id
Digital Skill Fair 33.0

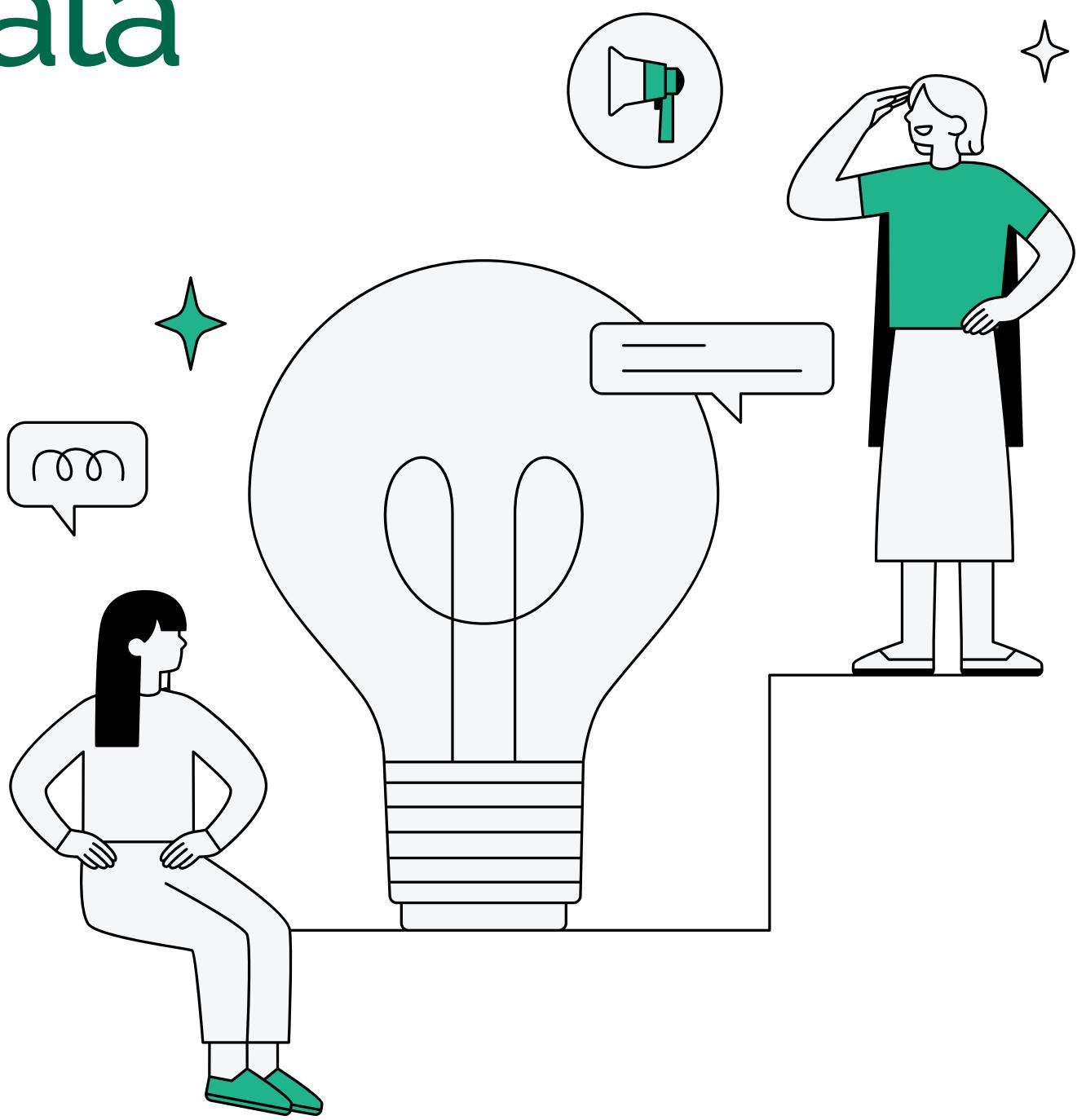
Introduction

Hi, I'm M. Harris Mulya Bahari. graduated with Bachelor of Economics from Airlangga University with GPA 3,68. Enthusiast in data science, macro and micro economics analysis, planning, and business development. Have good analysis, leadership, organizing, communication, and research skills through organizational, internship, and entrepreneurial experience. Can operate Microsoft excel and stata.



Deskripsi Proyek dan Data

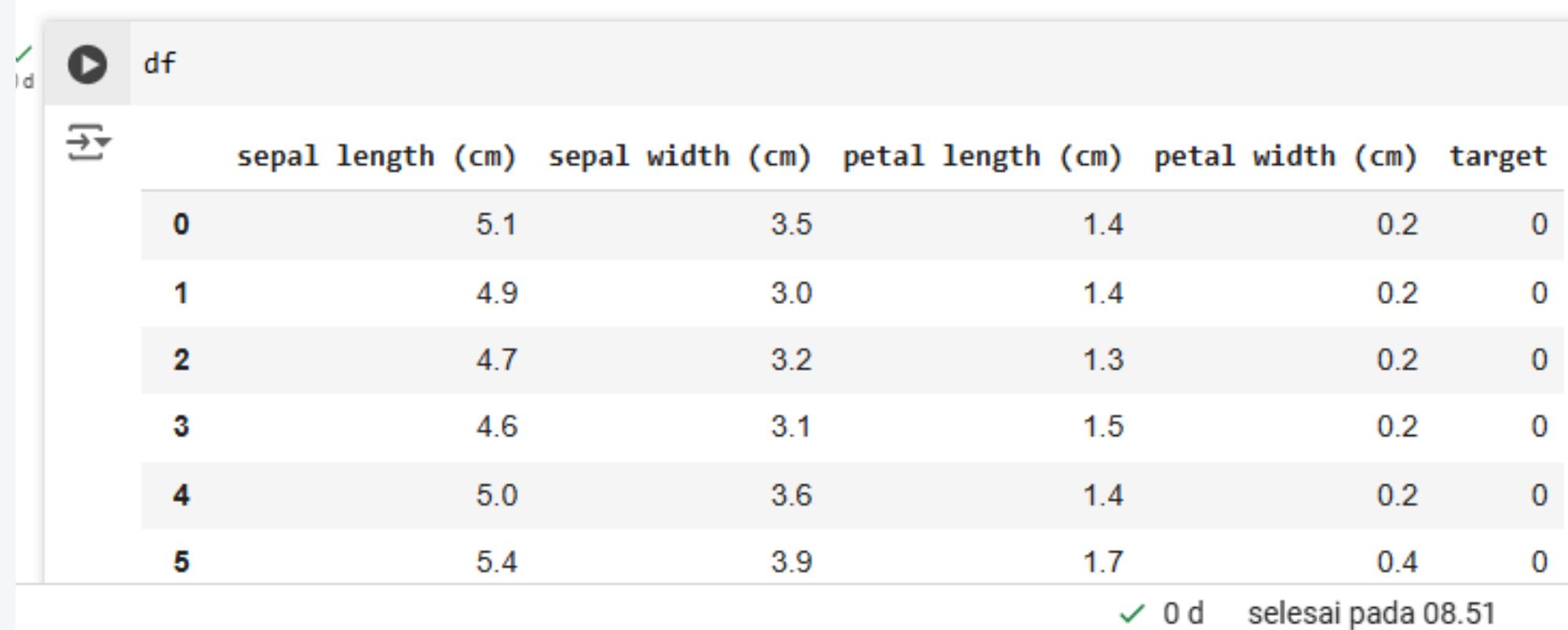
Proyek ini digunakan untuk mengklasifikasikan bunga Iris ke dalam 3 spesies berdasarkan ciri-cirinya. Data yang digunakan berasal dari Scikit dengan jumlah 150 data. Algoritma klasifikasi menggunakan K-Nearest Neighbors (KNN)



Import Library dan Data Overview

```
[ ] # Import Library  
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns
```

```
[40] from sklearn.datasets import load_iris  
data = load_iris()  
df = pd.DataFrame(data.data, columns=data.feature_names)  
df['target'] = data.target
```



The screenshot shows a Jupyter Notebook cell with the code to load the Iris dataset and create a DataFrame. Below the code, the variable `df` is displayed as a Pandas DataFrame. The DataFrame has 150 rows and 5 columns: `sepal length (cm)`, `sepal width (cm)`, `petal length (cm)`, `petal width (cm)`, and `target`. The first six rows of the data are shown:

| | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) | target |
|---|-------------------|------------------|-------------------|------------------|--------|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | 0 |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | 0 |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | 0 |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | 0 |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | 0 |
| 5 | 5.4 | 3.9 | 1.7 | 0.4 | 0 |

At the bottom of the cell, there is a progress bar indicating "0 d selesai pada 08.51".

Import library digunakan untuk memuat dan mevisualisasikan data yang akan digunakan dalam tahapan pemodelan.

Data Overview digunakan untuk menggambarkan informasi data seperti jumlah kolom dan baris.

Jumlah, Tipe, dan Class Dataset

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   sepal length (cm)    150 non-null   float64
 1   sepal width (cm)     150 non-null   float64
 2   petal length (cm)    150 non-null   float64
 3   petal width (cm)     150 non-null   float64
 4   target              150 non-null   int64  
dtypes: float64(4), int64(1)
memory usage: 6.0 KB
```

Data yang digunakan tidak ada yang hilang atau kosong sehingga tidak diperlukan penanganan seperti menghapus data atau interpolasi.

```
[44] print(pd.Series(data['target']).value_counts())
0    50
1    50
2    50
Name: count, dtype: int64

[45] df['target'].unique()
array([0, 1, 2])
```

Jumlah data yang digunakan adalah 150 data dengan klasifikasi target (setosa, versicolor, dan virginia)

Membagi Data

```
# Membagi data menjadi data latih dan data uji (80% latih, 20% uji)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Pembagian dataset menjadi data latih dan data tes menjadi prinsip dasar dari machine learning. Pembagian dataset 80% dari data digunakan untuk melatih model (training set), sementara 20% sisanya digunakan untuk menguji kinerja model (testing set).

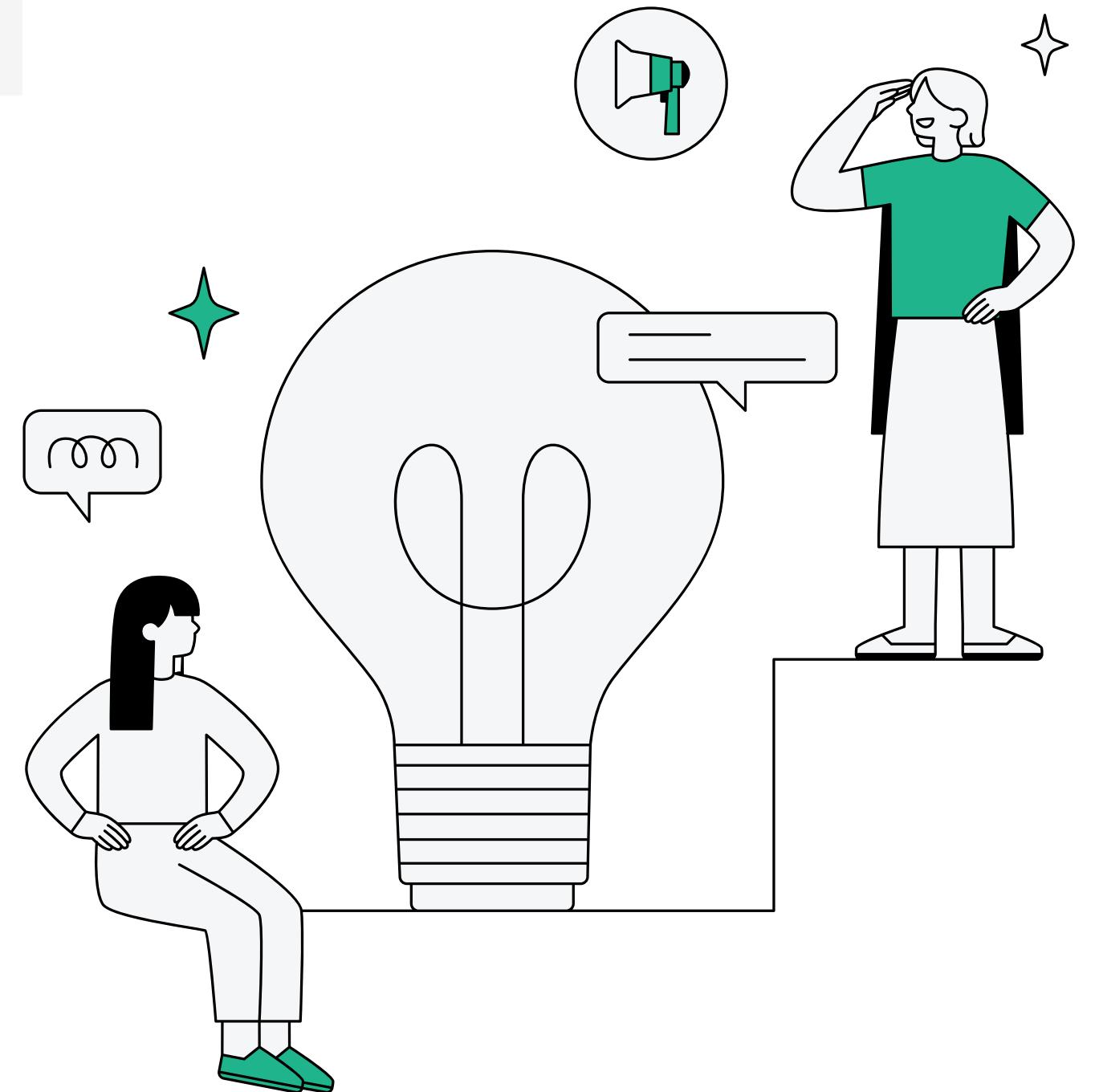
Melakukan modelling

```
# Membuat model KNN dengan K=3
knn = KNeighborsClassifier(n_neighbors=3)
```

```
# Melatih model menggunakan data latih
knn.fit(X_train, Y_train)
```

```
▼ KNeighborsClassifier ① ②
KNeighborsClassifier(n_neighbors=3)
```

Pemodelan menggunakan algoritma K-Nearest Neighbors (KNN)



Model Akurasi

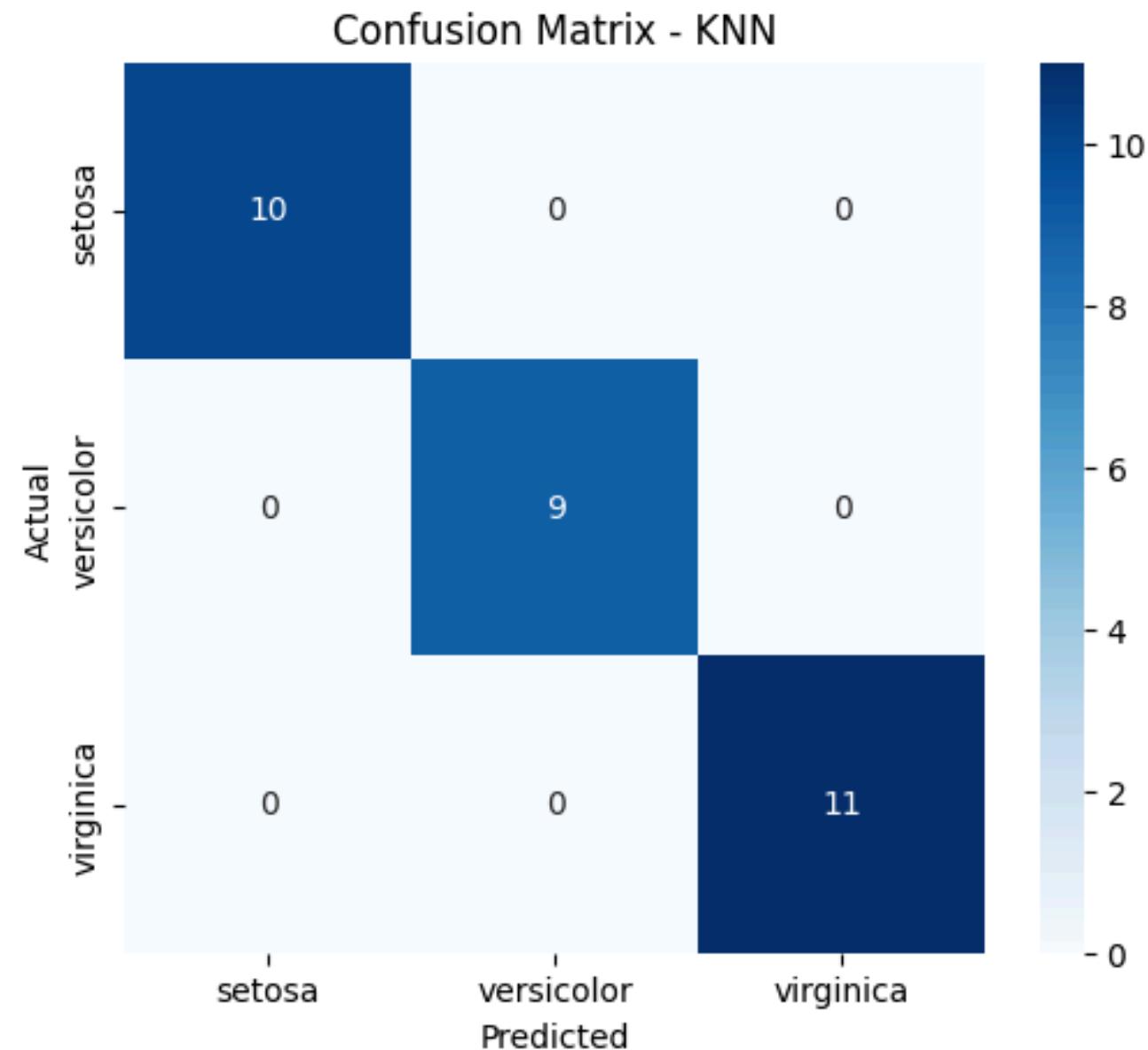
```
# Menghitung akurasi model
accuracy = accuracy_score(Y_test, Y_pred)
print(f"Akurasi model KNN: {accuracy * 100:.2f}%")
print(report)
```

Akurasi model KNN: 100.00%

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 1.00 | 1.00 | 1.00 | 10 |
| 1 | 1.00 | 1.00 | 1.00 | 9 |
| 2 | 1.00 | 1.00 | 1.00 | 11 |
| accuracy | | | 1.00 | 30 |
| macro avg | 1.00 | 1.00 | 1.00 | 30 |
| weighted avg | 1.00 | 1.00 | 1.00 | 30 |

Akurasi model digunakan untuk mengevaluasi performa model klasifikasi dan mengukur sejauh mana model berhasil dalam mengklasifikasikan data dengan benar.

Confusions Matrix

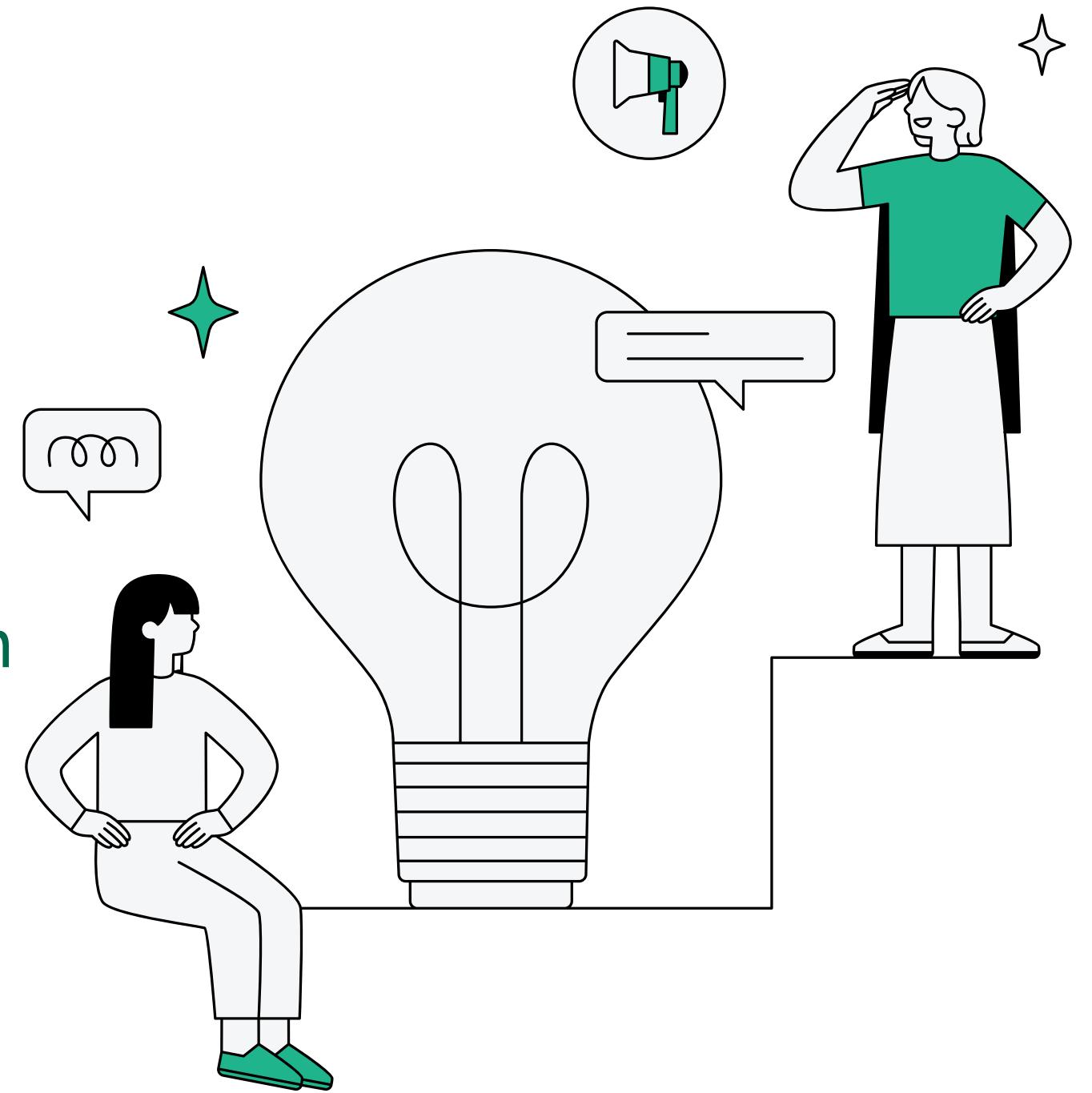


Confusion Matrix memberikan rincian jumlah prediksi yang salah dan benar. Berikut interpretasinya

- Ada **10** data label 'setosa' yang berhasil diprediksi sebagai label 'setosa'. Sementara tidak ada data yang sebenarnya label 'setosa' diprediksi label 'versicolor' dan 'virginica'.
- Ada **9** data label 'versicolor' yang berhasil diprediksi sebagai label 'versicolor'. Sementara tidak ada data yang sebenarnya label 'versicolor' diprediksi sebagai label 'setosa' dan 'virginica'.
- Ada **11** data label 'virginica' yang berhasil diprediksi sebagai label 'virginica'. sementara tidak ada data yang sebenarnya label 'virginica' diprediksi sebagai label 'setosa' dan 'versicolor'.

Simpulan

Algoritma K-Nearest Neighbors (KNN) berhasil mengklasifikasikan jenis bunga iris berdasarkan fitur-fiturnya dengan akurasi tinggi . Model dapat merepresentasikan hasil yang baik dan menjadi pilihan yang efektif.



Presented by M.Harris Mulya Bahari

Thank you very much!

