LAPORAN UJIAN TENGAH SEMESTER

MATA KULIAH *MACHINE LEARNING*

TK-45-GAB-G04

Pengolahan Dataset wine menggunakan Classification Model

Disusun untuk memenuhi UTS mata kuliah Machine Learning

di Program Studi S1 Teknik Komputer

Disusun oleh:

MUHAMMAD RAFINDHA ASLAM

1103213080

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

BANDUNG

2024

1. Latar Belakang

Model klasifikasi adalah salah satu jenis model yang paling sering digunakan dalam *Machine Learning*. Model ini digunakan untuk memprediksi kategori atau kelas dari suatu data berdasarkan fitur-fitur yang ada. Contohnya, pada dataset wine, kita mencoba memprediksi jenis wine (kelas) berdasarkan karakteristik kimianya. Contoh algoritma yag digunakan model klasifikasi termasuk:

- Logistic Regression: Meskipun bernama regresi, algoritma ini digunakan untuk klasifikasi biner atau multikelas.
- Decision Tree: Algoritma berbasis pohon keputusan yang mudah dipahami dan diinterpretasikan.
- k-Nearest Neighbors (k-NN): Algoritma berbasis jarak yang memprediksi kelas data berdasarkan kedekatan dengan data latih.

Lalu dataset yang digunakan untuk dilakukan pemrosesan menggunakan model regresi adalah dataset wine. Dataset ini berisi informasi tentang karakteristik kimia dan klasifikasi berbagai macam varietas buah-buahan. Pemrosesan dataset wine mencakup berbagai langkah seperti eksplorasi data, pembersihan, analisis, dan penerapan algoritma machine learning. File wine.names, yang berisi deskripsi fitur, dan Index, dapat ditambahkan untuk membuat proses analisis lebih lengkap.

Dataset wine.data terdiri dari 14 kolom, dengan 13 fitur dan 1 kolom target (Class). Berikut adalah deskripsi tiap variabel:

No.	Variabel	Deskripsi	Tipe Data
1.	Class	Jenis wine (1, 2, atau 3), digunakan sebagai target klasifikasi.	Kategorikal (int)
2.	Alcohol	Kandungan alkohol dalam wine (%)	Numerik (float)
3.	Malic_acid	Kandungan asam malat dalam wine.	Numerik (float)
4.	Ash	Kandungan abu dalam wine.	Numerik (float)
5.	Alcalinity_of_ash	Tingkat alkalinitas abu.	Numerik (float)
6.	Magnesium	Kandungan magnesium dalam wine (mg/L).	Numerik (int)
7.	Total_phenols	Total kandungan fenol dalam wine.	Numerik (float)
8.	Flavanoids	Kandungan flavonoid fenol dalam wine.	Numerik (float)

9.	Nonflavanoid_phenols	Kandungan fenol non-flavonoid.	Numerik (float)
10.	Proanthocyanins	Kandungan proantosianidin.	Numerik (float)
11.	Color_intensity	Intensitas warna wine.1	Numerik (float)
12.	Hue	Warna wine	Numerik (float)
13.	OD280/OD315	Rasio absorbansi OD280 terhadap OD315, mengindikasikan tingkat fenol terpolimerisasi.	Numerik (float)
14.	Proline	Kandungan prolin, asam amino utama dalam wine (mg/L).	Numerik (int)

2. Tujuan

1) Mengolah dataset wine.data untuk memahami pola data dan relasi antar fitur.
2) Menggunakan deskripsi dari wine.names untuk memberikan konteks pada setiap kolom dataset.
3) Menganalisis dataset menggunakan model machine learning untuk klasifikasi wine berdasarkan fitur kimianya.
4) Mengintegrasikan file Index (jika relevan) ke dalam proses analisis atau sebagai referensi tambahan.

3. Metode

1) Eksplorasi Dataset (wine.data):
   - Membaca dataset menggunakan Pandas.
   - Menangani missing values dan memahami distribusi data.
2) Menggunakan wine.names:
   - Membaca dan menganalisis deskripsi fitur untuk memberikan pemahaman tambahan terhadap dataset.
3) Model Machine Learning:
   - Klasifikasi: Logistic Regression, Decision Tree, dan k-Nearest Neighbors (k-NN).
   - Tuning Hyperparameter: Menggunakan GridSearchCV untuk menemukan parameter terbaik.

4) Visualisasi:
   - Heatmap korelasi antar fitur.

- Confusion matrix untuk evaluasi hasil klasifikasi.

5) Analisis:
   - Evaluasi kinerja model berdasarkan metrik seperti akurasi, precision, recall, dan F1-score.

4. Code

Berikut adalah implementasi pemrosesan dataset:

1) Membaca dan Mengeksplorasi Dataset

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Membaca dataset wine
columns = [
    "Class", "Alcohol", "Malic_acid", "Ash", "Alcalinity_of_ash", "Magnesium",
    "Total_phenols", "Flavanoids", "Nonflavanoid_phenols", "Proanthocyanins",
    "Color_intensity", "Hue", "OD280/OD315", "Proline"
]
data = pd.read_csv('/path/to/wine.data', header=None, names=columns)

# Mengeksplorasi dataset
print(data.info())
print(data.describe())
```

2) Membaca wine.names untuk Deskripsi Fitur

```
# Membaca deskripsi fitur
with open('/path/to/wine.names', 'r') as f:
    feature_descriptions = f.read()
print("Deskripsi Fitur:\n", feature_descriptions)
```

3) Penerapan Model Machine Learning

```
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
```

```
from sklearn.metrics import classification_report

# Split dataset menjadi fitur dan target
X = data.drop(columns=["Class"])
y = data["Class"]

# Split data menjadi train-test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Standard Scaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Model Logistic Regression
log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)

# Evaluasi Logistic Regression
print("Logistic Regression:")
print(classification_report(y_test, y_pred_log_reg))

# Model Decision Tree
tree = DecisionTreeClassifier(random_state=42)
tree.fit(X_train, y_train)
y_pred_tree = tree.predict(X_test)

# Evaluasi Decision Tree
print("Decision Tree:")
print(classification_report(y_test, y_pred_tree))

# Model K-Nearest Neighbors
```

```
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred_knn = knn.predict(X_test)

# Evaluasi KNN
print("K-Nearest Neighbors:")
print(classification_report(y_test, y_pred_knn))
```

4) Visualisasi Confusion Matrix

```
from sklearn.metrics import confusion_matrix
import numpy as np

# Confusion matrix untuk Logistic Regression
cm_log_reg = confusion_matrix(y_test, y_pred_log_reg)
sns.heatmap(cm_log_reg, annot=True, fmt='d', cmap='Blues',
xticklabels=np.unique(y), yticklabels=np.unique(y))
plt.title("Confusion Matrix: Logistic Regression")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

5. Analisis

1) Eksplorasi Data:
   - Dataset terdiri dari 178 sampel dengan 13 fitur numerik dan 1 kolom target tanpa nilai yang hilang.
   - Kandungan alkohol (Alcohol) berkisar antara 11.03% hingga 14.83%, dan kandungan prolin (Proline) berkisar dari 278 hingga 1680 mg/L.
   - Fitur Flavanoids dan Total_phenols memiliki korelasi tinggi (0.86), menunjukkan hubungan kuat.
2) Evaluasi Model:
   - Logistic Regression:
     Parameter Terbaik: C = 0.1, Akurasi Test Set: 100.00% (terbaik).
   - Decision Tree:
     Parameter Terbaik: max_depth = 3, Akurasi Test Set: 94.44% (cocok untuk interpretasi).

- k-NN:

  Parameter Terbaik: n_neighbors = 5, Akurasi Test Set: 97.22%.
- XGBoost:

  Parameter Terbaik: max_depth = 3, Akurasi Test Set: 97.22%.

3) Kesimpulan Analisis:

- Logistic Regression unggul dengan akurasi sempurna (100.00%).
- k-NN dan XGBoost menunjukkan performa sangat baik (97.22%).
- Decision Tree cukup andal (94.44%) dan cocok untuk analisis interpretatif.

6. Kesimpulan

Dataset wine.data dengan 178 sampel dan 13 fitur numerik berhasil diklasifikasikan menggunakan Logistic Regression yang menunjukkan akurasi sempurna (100%), sedangkan k-NN dan XGBoost memberikan akurasi tinggi (97.22%), menjadikannya alternatif andal, sementara Decision Tree yang memiliki akurasi lebih rendah (94.44%) cocok untuk interpretasi, sehingga disimpulkan bahwa Logistic Regression adalah model terbaik untuk dataset ini, dengan rekomendasi evaluasi tambahan seperti cross-validation untuk memastikan generalisasi model pada data baru.