Technical Report of Machine Learning : Breast Cancer Data Exploration

Nama : Harits Maulana Muzakki
NIM : 1103204166

Machine Learning

Machine learning merupakan cabang dari kecerdasan buatan yang melibatkan pembuatan sistem yang dapat belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dengan kata lain, algoritma machine learning dapat belajar dan memperbaiki kinerjanya dari pengalaman, tanpa campur tangan manusia.

Secara umum, machine learning melibatkan memberikan data ke dalam sebuah model, yang kemudian menggunakan data tersebut untuk mempelajari pola dan hubungan dalam data tersebut. Setelah model tersebut mempelajari pola tersebut, model tersebut dapat digunakan untuk membuat prediksi atau keputusan tentang data baru.

Ada tiga jenis utama dari machine learning: supervised learning, unsupervised learning, dan reinforcement learning. Pada supervised learning, model dilatih pada data yang sudah dilabeli, dimana setiap titik data memiliki label atau variabel target yang sesuai. Tujuannya adalah untuk mempelajari pemetaan antara fitur input dan label output, sehingga model dapat memprediksi label untuk data baru yang belum terlihat. Contoh dari tugas supervised learning termasuk klasifikasi gambar, pengenalan suara, dan penyaringan spam.

Pada unsupervised learning, model dilatih pada data tanpa label, dimana tidak ada label target yang dapat membimbing proses pembelajaran. Sebaliknya, tujuannya adalah untuk mempelajari struktur atau pola dalam data, seperti pengelompokan atau reduksi dimensi. Contoh dari tugas unsupervised learning termasuk deteksi anomali, kompresi data, dan sistem rekomendasi.

Pada reinforcement learning, model belajar dengan berinteraksi dengan lingkungan dan menerima umpan balik dalam bentuk hadiah atau hukuman. Tujuannya adalah untuk mempelajari sebuah kebijakan yang memaksimalkan hadiah kumulatif dari waktu ke waktu, seperti dalam bermain game atau robotika.

Machine learning memiliki banyak aplikasi di berbagai industri, termasuk kesehatan, keuangan, e-commerce, dan lain-lain. Beberapa contohnya meliputi diagnosis medis, deteksi penipuan, pemasaran personal, dan mobil otonom.

Pemodelan Machine Learning

Pada umumnya, Machine Learning memiliki beberapa model diantaranya :

1. Regresi Linear: Model sederhana dan banyak digunakan untuk tugas regresi. Ini memodelkan hubungan antara variabel dependen dan satu atau lebih variabel independen menggunakan persamaan linear. Regresi linear sering digunakan untuk tugas seperti memprediksi harga rumah atau harga saham.
2. Regresi Logistik: Model klasifikasi yang memodelkan probabilitas sebuah observasi termasuk ke dalam kelas tertentu. Ini sering digunakan dalam tugas seperti memprediksi apakah pelanggan akan keluar atau apakah transaksi kartu kredit adalah penipuan.
3. Pohon Keputusan: Model yang mempartisi data ke dalam subset berdasarkan nilai fitur, dan membuat model seperti pohon dari keputusan dan konsekuensi yang mungkin. Pohon keputusan dapat digunakan untuk tugas klasifikasi maupun regresi.
4. Random Forest: Jenis model ensemble learning yang menggabungkan beberapa pohon keputusan untuk membuat prediksi yang lebih akurat. Setiap pohon dilatih pada subset acak data, dan prediksi akhir didasarkan pada mayoritas suara dari pohon-pohon individu.
5. Mesin Vector Pendukung (SVM): Model yang menemukan batas atau hipertop yang optimal yang memisahkan data ke dalam kelas-kelas yang berbeda. SVM sering digunakan dalam klasifikasi gambar, klasifikasi teks, dan bioinformatika.

6. Jaringan Saraf: Model yang terinspirasi oleh struktur otak manusia, terdiri dari lapisan-lapisan simpul atau neuron yang saling terhubung untuk memproses informasi. Jaringan saraf dapat digunakan untuk berbagai tugas, termasuk pengenalan gambar dan suara, pemrosesan bahasa alami, dan sistem rekomendasi.
7. K-Nearest Neighbors (KNN): Model yang membuat prediksi berdasarkan kesamaan antara titik data baru dan k-titik data terdekat di set pelatihan. KNN sering digunakan untuk tugas klasifikasi.

Dari ketujuh model Machine Learning yang dijelaskan di atas, tiga model diantaranya populer untuk digunakan dalam tugas klasifikasi, yaitu Support Vector Machines (SVM), Random Forest, dan K-Nearest Neighbors (KNN).

Breast Cancer Dataset

Dataset kanker payudara yang dipublikasikan secara publik diantaranya :

1. Dataset Wisconsin Diagnostic Breast Cancer (WDBC): Dataset ini berisi informasi tentang karakteristik massa payudara, termasuk fitur seperti tekstur, radius, dan perimeter. Dataset ini mencakup 569 instansi dan tersedia di Repositori Pembelajaran Mesin UCI.

2. Dataset Breast Cancer Wisconsin (Original): Dataset ini berisi informasi tentang karakteristik massa payudara, termasuk fitur seperti ketebalan gumpalan, keseragaman ukuran sel, dan inti kosong. Dataset ini mencakup 699 instansi dan tersedia di Repositori Pembelajaran Mesin UCI.

3. Dataset Gambar Histopatologi Payudara: Dataset ini berisi gambar digital dari sampel jaringan payudara, yang dapat digunakan untuk melatih model pembelajaran mesin untuk deteksi dan klasifikasi kanker payudara. Dataset ini mencakup lebih dari 5.000 gambar dan tersedia di Kaggle.

4. Dataset Atlas Genom Kanker (TCGA): Dataset ini berisi data genomik dan klinis untuk beberapa jenis kanker, termasuk kanker payudara. Dataset ini mencakup lebih dari 1.000 kasus kanker payudara dan tersedia di situs web TCGA.

Dari keempat dataset yang tersedia, data eksplorasi kali ini menggunakan Dataset Wisconsin Diagnostic Breast Cancer (WDBC). Pengolahan dataset ini juga dipublikasikan oleh Scikit-Learn yang juga sebagai bahan pembelajaran pada mata kuliah ini. Untuk mengeksplorasi dataset ini, diperlukan beberapa library diantaranya :

```
# Import necessary libraries
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Library berikut digunakan untuk mendefinisikan fungsi-fungsi seperti numpy, pandas, dan seaborn. Selain itu juga ada beberapa dari sklearn untuk memvisualisasikan data trends, diantaranya : Memuat dataset kanker payudara, pemilihan model dengan test split, mempreproses data, menggambarkan regresi menggunakan linear model, dan menghitung nilai akurasi.

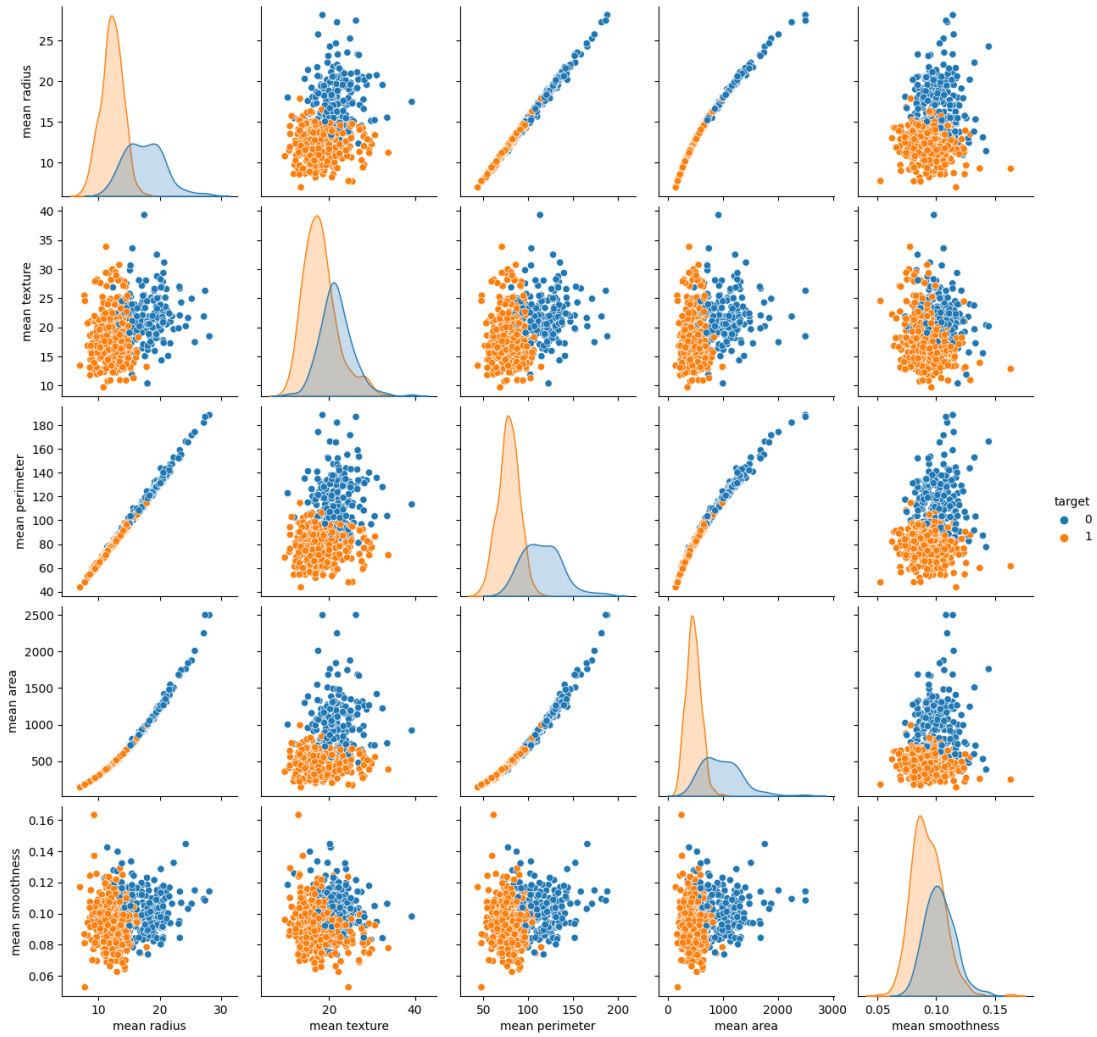
```
from sklearn.datasets import load_breast_cancer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

Sedangkan library ini digunakan untuk menentukan Decision Tree dan Random Forest. Karena Self-Training tidak dapat digunakan pada colab yang diberikan, maka perhitungan Self-Training dilakukan secara manual.

Untuk memvisualisasikan data trends, diperlukan kodingan menggunakan Seaborn untuk menggambarkannya. Tetapi sebelum itu, dataset di-split terlebih dahulu. Berikut kodingannya.

```
# Load the breast cancer dataset
data = load_breast_cancer()

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data.data, data.target, test_size=0.3, random_state=42)

# Scale the data
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Train a logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Convert the dataset into a pandas DataFrame
df = pd.DataFrame(data.data, columns=data.feature_names)

# Add the target variable to the DataFrame
df['target'] = data.target

# Use Seaborn to visualize the data trends
sns.pairplot(df, hue='target', vars=data.feature_names[:5])
```

Hasil dari kodingan diatas tergambarkan dengan tampilan berikut :



Setelah data trends divisualikasikan, data dieksplorasi menggunakan model Random Forest, Decision Tree, dan Self-Training.

```
# Train a decision tree model
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)

# Train a random forest model
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions on the test set using the decision tree and random forest models
dt_pred = dt_model.predict(X_test)
rf_pred = rf_model.predict(X_test)

# Calculate accuracy of the models
dt_accuracy = accuracy_score(y_test, dt_pred)
rf_accuracy = accuracy_score(y_test, rf_pred)
print("Decision tree accuracy:", dt_accuracy)
print("Random forest accuracy:", rf_accuracy)

# Evaluate the performance of the final model on the test set
X_test = data.data[100:]
y_test = data.target[100:]
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Final accuracy: {accuracy} ")
```

Hasil dari kodingan di atas hanya berupa numerik dengan hasil berikut :

Accuracy: 0.9824561403508771

Decision tree accuracy: 0.9415204678362573

Random forest accuracy: 0.9707602339181286

Final accuracy: 0.31343283582089554

Data eksplorasi ini keluarannya berupa angka akurasi dari model yang digunakan. Ternyata hasil akurasi sebelum dan sesudah menggunakan Decision Tree dan Random Forest berbeda karena diolah oleh model tersebut.