Technical Report

Visualisasi dan Eksplorasi Breast Cancer Dataset

Memenuhi Tugas UTS Mata Kuliah Machine Learning

Disusun Oleh :

Puteri Andini Rosmadila

1103204014

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

TELKOM UNIVERSITY

BANDUNG

2023

I. Dataset

Dataset kanker payudara dari sklearn adalah kumpulan data yang menyediakan informasi mengenai tumor payudara yang bersifat jinak (benign) atau ganas (malignant). Dataset ini berasal dari National Institute of Cancer di Amerika Serikat dan telah banyak digunakan sebagai contoh dalam bidang ilmu data dan pembelajaran mesin. Dataset ini terdiri dari 569 contoh, di mana masing-masing contoh terdiri dari 30 fitur numerik yang meliputi ukuran dan bentuk tumor, serta beberapa ukuran sel yang terdapat dalam citra tumor. Setiap contoh juga memiliki label kelas yang menyatakan apakah tumor tersebut bersifat jinak atau ganas. Dataset ini dapat digunakan untuk latihan klasifikasi, yaitu untuk mengembangkan model pembelajaran mesin yang dapat memprediksi apakah sebuah tumor payudara bersifat jinak atau ganas berdasarkan fitur-fitur yang ada. Hal ini sangat berguna dalam bidang kedokteran, di mana diagnosis kanker payudara dapat menjadi sangat kompleks dan membutuhkan bantuan teknologi. Dataset ini dapat diakses menggunakan perintah load_breast_cancer() dari pustaka sklearn.datasets.

II. Import Library dan Load Data

Pada tahap ini, dilakukan import beberapa library yang dibutuhkan untuk pengolahan data seperti

NumPy: library untuk komputasi numerik, termasuk operasi array dan matriks.

Pandas: library untuk analisis data, termasuk manipulasi data, agregasi, dan pengolahan data dalam format tabel.

Matplotlib: library untuk visualisasi data, termasuk pembuatan grafik dan plot.

Scikit-learn: library untuk pembelajaran mesin, termasuk klasifikasi, regresi, clustering, dan preprocessing data.

Seaborn: digunakan untuk visualisasi data pada tingkat tinggi dan dirancang untuk bekerja dengan baik dengan library Pandas.

Setelah itu, dilakukan load dataset menggunakan library sklearn.datasets.load_breast_cancer

III. Exploratory Data Analysis

EDA (Exploratory Data Analysis) atau Analisis Data Eksploratif adalah proses mengidentifikasi pola dan hubungan dalam data menggunakan metode visual dan deskriptif.

Tujuan utama dari EDA adalah untuk memahami karakteristik dan kecenderungan data sebelum dilakukan analisis lebih lanjut atau pembuatan model prediksi. Seperti melihat informasi umum dari data seperti jumlah baris dan kolom, tipe data dari masing-masing kolom, serta statistik deskriptif dari masing-masing kolom. Selanjutnya, dilakukan visualisasi data menggunakan library seaborn untuk melihat distribusi data dari masing-masing kolom serta hubungan antar kolom.

IV. Splitting Data atau Pemisah Data

Splitting data atau pemisahan data adalah proses memisahkan dataset menjadi dua bagian: data latih (training data) dan data uji (testing data). Data latih digunakan untuk membangun model prediksi atau pembelajaran mesin (machine learning) dan data uji digunakan untuk mengevaluasi kinerja model atau untuk menguji keakuratan model prediksi pada data yang belum pernah dilihat sebelumnya. Pada bagian ini, dilakukan pemisahan data menjadi data train dan data test menggunakan perintah train_test_split dengan rasio data train 70% dan data test sebesar 30%.

V. Modeling dan Evaluasi data

Pemodelan adalah proses membangun model prediksi atau pembelajaran mesin (machine learning) menggunakan algoritma tertentu untuk mempelajari pola dan hubungan dalam data. Pada percobaan ini, dilakukan pemodelan menggunakan Decision Tree, Random Forest, dan Self-Training dengan menggunakan library scikit-learn di Python.

Pertama, dilakukan pemodelan menggunakan DecisionTreeClassifier. Decision Tree adalah algoritma pembelajaran mesin yang populer digunakan untuk klasifikasi dan regresi. DecisionTreeClassifier digunakan untuk membangun model Decision Tree pada data yang diberikan. Setelah membangun model, evaluasi dilakukan menggunakan accuracy_score. Accuracy_score adalah metrik evaluasi yang digunakan untuk mengukur seberapa akurat model dalam memprediksi label pada data uji. Semakin tinggi nilai accuracy_score, semakin baik performa model dalam melakukan prediksi.

Selanjutnya, dilakukan pemodelan menggunakan RandomForestClassifier. Random Forest adalah algoritma ensemble dari Decision Tree yang digunakan untuk klasifikasi, regresi, dan tugas-tugas lainnya. RandomForestClassifier digunakan untuk membangun model Random

Forest pada data yang diberikan. Setelah membangun model, evaluasi dilakukan menggunakan accuracy_score, sama seperti pada pemodelan Decision Tree sebelumnya.

Terakhir, dilakukan pemodelan Self-Training menggunakan SelfTrainingClassifier dengan DecisionTreeClassifier sebagai model classifier. Self-Training adalah teknik pembelajaran mesin semi-supervised yang digunakan untuk memanfaatkan data yang tidak memiliki label untuk memperbaiki performa model pada data yang memiliki label. SelfTrainingClassifier digunakan untuk membangun model Self-Training pada data yang diberikan. Setelah membangun model, evaluasi dilakukan menggunakan accuracy_score, sama seperti pada pemodelan sebelumnya. Dalam kesimpulan, pada pemodelan menggunakan Decision Tree, Random Forest, dan Self-Training dilakukan pemodelan dan evaluasi performa model menggunakan accuracy_score. Metrik evaluasi ini digunakan untuk mengevaluasi seberapa akurat model dalam melakukan prediksi pada data uji. Semakin tinggi nilai accuracy_score, semakin baik performa model dalam melakukan prediksi.

VI. Visualisasi data

Pada bagian ini, dilakukan visualisasi model Decision Tree dan Random Forest menggunakan library graphviz. Selain itu, dilakukan visualisasi fitur terpenting dari model Random Forest menggunakan library seaborn.

VII.Kesimpulan

Berdasarkan hasil evaluasi dari model, ditemukan bahwa model Random Forest memiliki akurasi yang lebih tinggi dibandingkan dengan Decision Tree dan Self-Training. Dalam percobaan ini di peroleh hasil akurasi dari DecisionTreeClassifier sebesar 90,64%, kemudian model RandomForestClassifier dengan akurasi 95,90%, dan Self-Training sebesar 90,64%.