Nama : Ardhien Fahillah Suhartono

NIM : 1103204137

Dataset : https://www.kaggle.com/datasets/danielgrijalvas/movies

ChatGPT : https://chat.openai.com/share/b5180bf5-89c5-4c53-b743-c1336293ea70

Script UTS Machine Learning

Assalamualaikum semuanya, perkenalkan saya (perkenalan diri) akan menyampaikan tugas uts machine learning dengan data set movie yang bisa dilihat dari link berikut serta dengan prompt chatgpt saya yang bisa dilihat dari link berikut

Random Forest

Pertama tama saya akan menjelaskan random forest terlebih dahulu dimana random forest merupakan salah satu metode klasifikasi data yang menggunakan ensemble learning. Ensemble learning adalah metode pembelajaran mesin yang menggabungkan hasil dari beberapa model untuk mendapatkan hasil yang lebih baik daripada model tunggal.

Dalam klasifikasi menggunakan random forest, ensemble learning dilakukan dengan membangun kumpulan pohon keputusan (decision tree) yang saling independen. Setiap pohon keputusan dibangun menggunakan sampel data yang berbeda dan atribut yang dipilih secara acak.

Exploratory Data Analysis

Selanjutnya kita akan membahas mengenai eda yang merupakan langkah penting dalam proses analisis data. Ini melibatkan pemeriksaan dan ringkasan kumpulan data untuk memahami karakteristik utamanya, mengidentifikasi pola, dan mendeteksi anomali. EDA membantu ilmuwan dan analis data mendapatkan pemahaman yang lebih mendalam tentang data sebelum mendalami pemodelan atau pengujian hipotesis yang lebih kompleks.

Data Visualization

adalah representasi grafis dari data dan informasi. Ini melibatkan perancangan dan pembuatan elemen visual, seperti bagan, grafik, dan peta, untuk mengkomunikasikan data secara efektif. Visualisasi data membantu orang memahami tren, pola, dan hubungan dalam data dengan lebih mudah daripada sekadar membaca teks atau angka.

Dimana kali ini kita akan menggunakan pustaka seaborn dan matplotlib untuk membuat subplot yang menampilkan visualisasi distribusi beberapa fitur dalam dataset film, seperti distribusi

rating, genre, tahun rilis, skor, jumlah suara (votes), dan durasi (runtime). Setiap subplot disusun dalam tata letak 3x2 untuk memberikan gambaran yang lebih komprehensif tentang karakteristik dataset film.

Training Data

Dalam konteks machine learning dan data mining, data pelatihan mengacu pada contoh berlabel yang digunakan untuk melatih model pembelajaran mesin. Data ini penting untuk mengajarkan model cara mengidentifikasi pola dan membuat prediksi. Kualitas dan kuantitas data pelatihan berdampak signifikan terhadap performa model yang dihasilkan.

1. **Label Encoding (LabelEncoder):** Menggunakan LabelEncoder dari scikit-learn untuk mengonversi nilai pada kolom 'rating' ke dalam bentuk bilangan bulat. Ini berguna saat kita bekerja dengan algoritma pembelajaran mesin yang memerlukan input berupa bilangan bulat, seperti model klasifikasi.

2. **Pengisian Nilai Kosong (fillna):** Menggantikan nilai yang hilang (NaN) pada kolom 'score' dengan nilai rata-rata dari kolom tersebut. Langkah ini membantu mengatasi masalah nilai yang hilang pada dataset, sehingga data menjadi lebih lengkap dan dapat digunakan untuk analisis lebih lanjut atau pelatihan model.

Evaluating Data

adalah proses menilai kualitas, kelengkapan, dan keakuratan data sebelum digunakan untuk analisis atau pengambilan keputusan. Ini merupakan langkah penting dalam setiap proyek berbasis data, karena membantu memastikan bahwa data dapat diandalkan dan dipercaya.

Kode tersebut menghitung Mean Squared Error (MSE), sebuah metrik evaluasi untuk mengevaluasi sejauh mana model regresi numerik cocok dengan data yang diuji. MSE mengukur rata-rata kuadrat perbedaan antara nilai sebenarnya (y_test) dan nilai yang diprediksi oleh model (y_pred). Semakin rendah MSE, semakin baik model dalam memprediksi data yang diuji.

Mean Squared Error (MSE):

Definisi: MSE adalah rata-rata dari kuadrat selisih antara nilai sebenarnya (y) dan nilai yang diprediksi ($\hat{y}$). Formula: $MSE = (1/n) \Sigma(y - \hat{y})^2$, di mana n adalah jumlah sampel. Interpretasi: Semakin rendah nilai MSE, semakin baik modelnya. MSE memberikan gambaran tentang seberapa dekat prediksi model dengan nilai sebenarnya. Namun, MSE cenderung memberikan bobot lebih pada kesalahan yang besar karena selisihnya dikuadrat.