Nama :Dery Hidayat
NIM :1103228181

Link YT : https://youtu.be/Ze13jAXSAOU

YT tutorial script:

Tutorial

Pertama kita Mount Google Drive yaitu memuat Dataset

1. **Buka Google Colab**: Kunjungi Google Colab di https://colab.research.google.com/.
2. **Buat Notebook Baru atau Buka Notebook yang Ada**: Anda dapat membuat notebook baru atau membuka notebook yang sudah ada.
3. **Mount Google Drive**: Untuk melakukan mounting Google Drive, jalankan kode berikut pada cell pertama di notebook Anda:

Dengan melakukan mounting Google Drive, Anda bisa dengan mudah mengakses file dan dataset yang ada di Google Drive Anda dan menggunakannya dalam proses analisis data atau project machine learning di Google Colab.

Kedua apa itu EDA?

EDA adalah singkatan dari Exploratory Data Analysis, yang merupakan proses awal dalam analisis data. Tujuannya adalah untuk memahami isi dari dataset yang telah diperoleh sebelum menjalankan analisis yang lebih mendalam atau membangun model. EDA melibatkan serangkaian teknik dan pendekatan untuk mengeksplorasi, meringkas, dan menganalisis data untuk mendapatkan wawasan yang bermanfaat.

Langkah-langkah dalam proses EDA umumnya meliputi:

1. **Pemahaman Data**: Mengetahui jumlah kolom, jenis data (numerik, kategorikal, dll.), serta struktur umum dataset.
2. **Handling Missing Values**: Mengidentifikasi dan menangani nilai-nilai yang hilang dalam dataset, jika ada.
3. **Statistical Summary**: Mendapatkan ringkasan statistik seperti mean, median, nilai maksimum, minimum, dan deviasi standar untuk setiap fitur numerik.
4. **Visualisasi Data**: Membuat grafik atau plot untuk memahami distribusi, korelasi, dan pola dalam data. Ini meliputi histogram, scatter plot, box plot, heatmap, dan visualisasi lainnya.
5. **Mengetahui Distribusi Data**: Melihat distribusi setiap fitur untuk memahami bagaimana nilai-nilai tersebar dalam setiap variabel.
6. **Korelasi antar Fitur**: Menganalisis korelasi antar fitur untuk memahami hubungan di antara variabel-variabel tersebut.
7. **Pengecekan Outlier**: Mengidentifikasi dan memahami titik-titik data yang berbeda atau tidak biasa yang mungkin mempengaruhi analisis.

8. **Feature Engineering**: Menyusun fitur atau membuat fitur baru yang mungkin diperlukan untuk analisis atau model selanjutnya.
9. **Eksplorasi Melalui Grup atau Segmen**: Jika ada, mengeksplorasi perbedaan dalam data berdasarkan kelompok atau segmen tertentu.
10. **Validasi Asumsi**: Melakukan validasi atas asumsi-asumsi yang mungkin Anda miliki tentang data sebelum melangkah ke langkah-langkah analisis atau modelling yang lebih kompleks.

EDA sangat penting karena membantu dalam merumuskan pertanyaan-pertanyaan analisis yang lebih mendalam, mengidentifikasi pola yang menarik, memahami karakteristik data, dan mempersiapkan data dengan cara yang diperlukan untuk analisis atau pemodelan yang akan datang. Ini merupakan langkah awal yang krusial dalam menggali wawasan dari dataset sebelum proses analisis atau pemodelan lebih lanjut.

Ketiga Data Visualization

Visualisasi data adalah representasi grafis dari informasi dan data yang memungkinkan kita untuk memahami pola, hubungan, dan tren dalam dataset dengan lebih baik melalui penggunaan grafik, plot, dan diagram. Tujuannya adalah untuk menyajikan informasi yang kompleks dalam bentuk yang lebih mudah dipahami dan digunakan untuk pengambilan keputusan.

Beberapa alasan penting mengapa visualisasi data sangat penting:

1. **Mengungkap Pola dan Tren**: Grafik dan plot memungkinkan kita untuk melihat pola atau tren yang mungkin tidak terlihat saat melihat data mentah.
2. **Membandingkan Data**: Grafik memudahkan perbandingan antara berbagai aspek data, memungkinkan identifikasi perbedaan atau kesamaan yang signifikan.
3. **Mengidentifikasi Outlier**: Visualisasi membantu mengidentifikasi data outlier atau titik data yang tidak biasa yang dapat mempengaruhi hasil analisis.
4. **Menguji Asumsi**: Dengan visualisasi, kita dapat menguji asumsi-asumsi tentang data yang mungkin kita miliki.
5. **Komunikasi yang Efektif**: Grafik dan plot dapat dengan mudah dikomunikasikan kepada orang lain, membuatnya menjadi alat yang kuat untuk berbagi temuan atau wawasan.

Beberapa jenis visualisasi data yang umum digunakan meliputi:

- **Histogram**: Menampilkan distribusi frekuensi dari data numerik.
- **Scatter Plot**: Menunjukkan hubungan antara dua variabel numerik.
- **Line Chart**: Menunjukkan tren dari data seiring waktu atau urutan.
- **Bar Chart**: Membandingkan nilai-nilai kategori atau variabel.
- **Pie Chart**: Menunjukkan proporsi atau persentase dari kategori-kategori yang berbeda dalam suatu dataset.

Ada banyak alat dan library yang tersedia untuk membuat visualisasi data dalam bahasa pemrograman seperti Python (Matplotlib, Seaborn, Plotly), R (ggplot2), dan alat lainnya seperti Tableau, Power BI, dan Excel.

Visualisasi data membantu dalam menggali wawasan dari dataset dengan cara yang intuitif dan memudahkan untuk memahami pola dan tren yang mungkin tersembunyi dalam data.

Keempat yaitu Training Dataset

Training dataset adalah bagian dari dataset yang digunakan untuk melatih atau mengajarkan model machine learning. Data dalam training dataset digunakan oleh algoritma machine learning untuk menyesuaikan parameter-parameter internal model agar dapat membuat prediksi atau melakukan tugas tertentu.

Training dataset biasanya terdiri dari dua komponen utama:

1. **Fitur (Features)**: Fitur-fitur atau atribut-atribut dari dataset yang digunakan sebagai input atau variabel independen dalam proses pembelajaran. Misalnya, jika Anda ingin memprediksi harga rumah berdasarkan ukuran, lokasi, dan jumlah kamar, fitur-fitur tersebut akan menjadi bagian dari dataset.
2. **Label atau Target (Target)**: Label atau target adalah variabel yang ingin diprediksi oleh model. Misalnya, dalam masalah klasifikasi, label adalah kategori atau kelas yang ingin diprediksi oleh model. Dalam regresi, label adalah nilai yang ingin diprediksi.

Tujuan dari menggunakan training dataset adalah untuk mengajarkan model algoritma machine learning tentang pola dan hubungan antara fitur-fitur dengan label atau target yang sesuai. Model menggunakan training dataset untuk menyesuaikan parameter-parameter internalnya sehingga dapat membuat prediksi yang akurat atau melakukan tugas tertentu pada data yang belum pernah dilihat sebelumnya (data uji).

Kelima Evaluating Model

Evaluating model dataset adalah proses untuk menilai atau mengukur kinerja suatu model machine learning menggunakan dataset yang telah dipecah menjadi training set dan test set. Evaluasi model dilakukan untuk memahami seberapa baik model dapat menggeneralisasi pada data yang belum pernah dilihat sebelumnya (test set) setelah melalui proses pelatihan pada data yang telah dikenal (training set).

Beberapa metrik evaluasi umum untuk mengukur kinerja model machine learning termasuk:

1. **Akurasi (Accuracy)**: Persentase prediksi yang benar dari total prediksi yang dibuat oleh model.
2. **Precision**: Ukuran dari keakuratan prediksi positif. Precision memberi tahu seberapa banyak dari prediksi positif yang sebenarnya benar.
3. **Recall (Sensitivity)**: Ukuran dari seberapa banyak dari kelas positif yang terdeteksi. Recall memberi tahu seberapa banyak dari kelas positif yang sebenarnya terdeteksi oleh model.
4. **F1-Score**: Pengukuran rata-rata harmonis antara precision dan recall. Berguna ketika kelas-kelas tidak seimbang.
5. **ROC-AUC Score**: Area di bawah kurva ROC. Berguna untuk evaluasi klasifikasi biner.
6. **MSE (Mean Squared Error)**: Pengukuran kesalahan rata-rata kuadrat dari prediksi model pada data regresi.

7. **R-squared (Coefficient of Determination)**: Mengukur seberapa baik model sesuai dengan data pada regresi.

Langkah-langkah dalam melakukan evaluasi model dataset meliputi:

1. **Prediksi pada Data Uji**: Gunakan model untuk membuat prediksi pada data uji (test set) yang belum pernah dilihat sebelumnya.
2. **Perhitungan Metrik Evaluasi**: Hitung metrik evaluasi yang relevan seperti akurasi, precision, recall, atau metrik lainnya sesuai dengan jenis masalah dan model yang digunakan.
3. **Analisis Hasil**: Analisis dan interpretasikan hasil metrik evaluasi untuk memahami seberapa baik model berkinerja pada data yang belum pernah dilihat sebelumnya.
4. **Fine-tuning Model**: Jika hasil evaluasi menunjukkan performa yang kurang baik, mungkin diperlukan fine-tuning pada model, seperti mengubah parameter, melakukan feature engineering, atau menggunakan model yang berbeda.

Evaluasi model dataset membantu untuk memahami seberapa baik model dapat melakukan prediksi pada data yang belum pernah dilihat sebelumnya, dan memungkinkan pengembang atau pemodel untuk meningkatkan atau memperbaiki model sesuai dengan hasil evaluasi yang diperoleh.

Keenam Input Data Baru

Input data baru dalam konteks dataset merujuk pada data yang tidak pernah dilihat oleh model pada tahap pelatihan atau evaluasi sebelumnya. Data ini digunakan untuk menguji kinerja model atau untuk membuat prediksi baru setelah model selesai dilatih.

Saat model machine learning telah dilatih menggunakan training dataset dan diuji menggunakan test dataset, input data baru biasanya digunakan untuk menguji kemampuan prediktif model. Data ini tidak terlibat dalam proses pembelajaran atau evaluasi model sebelumnya.

Proses penggunaan input data baru dalam dataset melibatkan langkah-langkah berikut:

1. **Pemisahan Data**: Pastikan data baru yang ingin Anda uji tidak termasuk dalam training atau test dataset yang telah digunakan sebelumnya untuk melatih atau menguji model.
2. **Preprocessing**: Lakukan preprocessing data yang diperlukan pada data baru agar sesuai dengan format dan standar yang sama dengan data yang digunakan dalam pelatihan model. Ini mungkin melibatkan langkah-langkah seperti normalisasi, pengubahan tipe data, atau penghapusan nilai yang hilang.
3. **Penggunaan Model**: Gunakan model yang telah dilatih sebelumnya untuk melakukan prediksi atau analisis pada data baru. Input data ini akan melewati model dan menghasilkan prediksi atau output berdasarkan apa yang telah dipelajari oleh model dari data training sebelumnya.
4. **Evaluasi Prediksi**: Jika tujuan penggunaan data baru adalah untuk menguji kinerja model, hasil prediksi dari data baru dapat dievaluasi menggunakan metrik evaluasi yang relevan seperti akurasi, precision, recall, dll.

Input data baru adalah langkah penting dalam pengujian keandalan dan kinerja model machine learning. Ini memungkinkan kita untuk melihat bagaimana model berkinerja pada data yang belum pernah dilihat

sebelumnya dan memverifikasi apakah model mampu menggeneralisasi dengan baik pada data baru tersebut.

Penutup

Setelah mengetahui tentang dataset, analisis, EDA, Data Visualization, Training model, dan Input data baru. Semoga dataset ini dapat memberikan wawasan yang bermanfaat dan memicu penelitian yang lebih lanjut dalam Zoo Animals Extended Dataset ini. Terima kasih atas perhatian 😊