NIM : 442023611012
NAME : RIZKY CAHYONO PUTRA
CLASS : TI 4 / A2

Laporan Singkat tentang Pra-pemrosesan Data

1. Pentingnya Pra-pemrosesan Data

Pra-pemrosesan data adalah langkah krusial dalam analisis data dan pembelajaran mesin. Langkah ini melibatkan berbagai teknik untuk membersihkan dan mempersiapkan data sebelum digunakan dalam model analisis atau algoritma pembelajaran mesin. Tanpa pra-pemrosesan yang tepat, data yang tidak bersih dapat menyebabkan hasil analisis yang tidak akurat dan interpretasi yang salah. Misalnya, nilai yang hilang, outlier, dan format data yang tidak konsisten dapat mengganggu performa model dan mengurangi keakuratan prediksi. Oleh karena itu, memastikan bahwa data dalam kondisi baik adalah langkah pertama yang harus diambil sebelum melanjutkan ke analisis lebih lanjut.

Dalam konteks pembelajaran mesin, pra-pemrosesan data tidak hanya meningkatkan kualitas model tetapi juga mempercepat waktu pelatihan. Dengan menerapkan teknik normalisasi, pengkodean kategori, dan penghapusan data yang tidak relevan, kita dapat mengurangi kompleksitas model dan meningkatkan efisiensi. Proses ini juga membantu dalam meningkatkan generalisasi model, yang penting agar model tidak hanya bekerja baik pada data pelatihan tetapi juga pada data yang tidak terlihat sebelumnya. Oleh karena itu, investasi waktu dan upaya dalam pra-pemrosesan data akan membuahkan hasil yang signifikan pada tahap selanjutnya.

2. Pembelajaran tentang Dataset

Dataset yang digunakan dalam analisis ini adalah dataset harga perumahan di California. Dataset ini memiliki beberapa atribut, termasuk lokasi geografis (longitude dan latitude), usia median rumah, jumlah total kamar dan tidur, populasi, serta pendapatan median. Melalui eksplorasi dataset ini, saya belajar bahwa harga rumah dipengaruhi oleh berbagai faktor, termasuk lokasi dan demografi populasi. Misalnya, daerah dengan pendapatan median yang lebih tinggi cenderung memiliki harga rumah yang lebih tinggi. Selain itu, analisis statistik dasar menunjukkan adanya distribusi yang berbeda di antara fitur-fitur tersebut, yang memerlukan perhatian khusus dalam pra-pemrosesan, seperti menangani outlier dan nilai yang hilang.

Selama eksplorasi, saya juga menemukan bahwa terdapat beberapa kolom yang memiliki nilai null. Penanganan nilai null ini sangat penting karena keberadaan nilai yang hilang dapat menyebabkan model tidak dapat berfungsi dengan baik. Saya belajar bahwa kita dapat menghapus baris dengan nilai null atau menggantinya dengan nilai rata-rata atau median. Pemahaman tentang bagaimana masing-masing fitur saling berhubungan dan pengaruhnya terhadap harga rumah adalah bagian penting dari analisis ini, yang membantu dalam pengambilan keputusan yang lebih baik di masa depan.

3. Tantangan yang Dihadapi Selama Langkah Pra-pemrosesan

Selama langkah pra-pemrosesan, saya menghadapi beberapa tantangan yang perlu diatasi agar analisis dapat dilakukan dengan baik. Salah satu tantangan utama adalah menangani nilai yang hilang. Dataset ini mengandung beberapa baris yang memiliki nilai null pada atribut penting seperti jumlah total kamar dan pendapatan median. Menghapus baris-baris ini berpotensi mengurangi jumlah data yang tersedia untuk analisis. Oleh karena itu, saya harus memutuskan antara menghapus data tersebut atau mengisi nilai yang hilang dengan estimasi yang tepat. Saya memilih untuk mengganti nilai null dengan nilai rata-rata, tetapi tantangan ini memberikan wawasan tentang pentingnya pemilihan metode yang tepat untuk menangani nilai yang hilang dalam dataset.

Tantangan lainnya adalah menangani outlier yang dapat mempengaruhi analisis dan model yang dihasilkan. Dalam visualisasi awal, saya melihat bahwa ada beberapa outlier yang sangat berbeda dari data lainnya, terutama dalam kolom harga rumah. Outlier ini bisa disebabkan oleh kesalahan pengukuran atau kondisi khusus yang perlu dipertimbangkan. Saya harus melakukan analisis lebih dalam untuk memutuskan apakah outlier tersebut perlu dihapus atau tetap dipertahankan untuk memahami distribusi data dengan lebih baik. Proses ini mengajarkan saya pentingnya melakukan analisis yang menyeluruh terhadap data untuk memastikan bahwa setiap langkah pra-pemrosesan diambil berdasarkan pemahaman yang baik tentang data itu sendiri.