Technical Test

Analyst Development Program 2023

Rizky Ardianto – Data Scientist

Latihan 1

Sebagai seorang Data Scientist, berikut adalah beberapa cara yang dapat dilakukan jika data hilang atau tidak sesuai :

1. Imputasi data

Hal ini bermaksud untuk mengisi nilai-nilai yang hilang atau tidak sesuai dengan perkiraan yang masuk akal, contohnya menggunakan teknik statistik seperti nilai rata-rata (Mean), nilai tengah (Median), dan nilai yang sering muncul (Modus).

2. Menghapus data yang hilang

Jika hanya sebagian kecil data yang hilang, dan tidak mengganggu proses pengujian Analisa, maka seorang Data Scientist dapat menghapus data yang hilang atau tidak sesuai.

3. Pengambilan ulang data

Jika data yang hilang atau tidak sesuai dalam jumlah besar ataupun penting, seorang Data Scientist dapat melakukan pengambilan ulang data, meskipun membutuhkan waktu dan biaya tambahan.

4. Menggunakan Machine Learning

Melalui algoritma machine learning, seorang Data Scientist dapat dibantu untuk memprediksi nilai yang hilang, sama halnya dengan imputasi data, hanya saja jauh lebih kompleks.

5. Menganalisa data yang hilang

Sebagai seorang data scientist, data-data yang ada dapat menjadi petunjuk dan pola tertentu. Ada kemungkinan bahwa data yang hilang dapat memberikan tambahan informasi dan membantu analisis menjadi lebih baik.

Jika terdapat duplikasi data, berikut adalah beberapa hal yang dapat dilakukan :

1. Pengecekan melalui aplikasi

Melalui berbagai alat bantu seperti Excel, dan menggunakan bahasa pemrograman seperti SQL dan Python.

2. Menganalisa ulang

Setelah menemukan duplikasi, seorang data scientist, dapat melakukan analisa dan pengecekan ulang. Melalui hal tersebut, ada kemungkinan bahwa data yang terduplikasi dapat memberikan tambahan informasi dan membantu analisis menjadi lebih jauh.

Latihan 2

Sebagai seorang Data Scientist, berikut adalah beberapa visualisasi efektif yang dapat digunakan untuk mengeksplorasi pola dan hubungan :

1. **Line Chart**. Melalui visualisasi ini, dapat menampilkan tren selama perjalanan.
2. **Bar Chart**. Melalui visualisasi ini, dapat membandingkan data perjalanan di berbagai kategori.
3. **Heatmap**. Melalui visualisasi ini, dapat membantu melihat pola hubungan seperti korelasi.
4. **Map**. Melalui visualisasi ini, dapat membantu melihat secara interaktif atau statis perjalanan dalam konteks geografis.
5. **Histogram**. Melalui visualisasi ini, dapat menunjukkan distribusi data, seperti frekuensi perjalanan, durasi, harga, dan lainnya.
6. **Scatterplot**. Melalui visualisasi ini, dapat melihat hubungan antara dua variabel dan menentukan apakah ada korelasinya.

Latihan 3

Berikut adalah statistik deskriptif dan inferensial yang dapat membantu menganalisa dan memahami pola dan hubungan antar fitur :

1. Melalui kolom umur, pendidikan, dan lama bekerja, dapat ditemukan hubungan seberapa lama waktu yang dibutuhkan karyawan untuk lulus dan mendapatkan posisi sekarang ini.
2. Melalui kolom lama bekerja dan gaji, dapat ditemukan hubungan kenaikan dan total gaji karyawan. Hal ini sangat berguna untuk sampel karyawan dengan posisi yang sama.
3. Melalui gabungan dari keseluruhan kolom, yaitu umur, jenis kelamin, pendidikan, lama bekerja, dan gaji, dapat ditemukan hubungan karyawan dengan kemampuan perusahaan.
4. Melalui beberapa sampel karyawan dengan posisi dan jabatan yang sama, maka dapat diketahui kondisi internal perusahaan.

Latihan 4

Berikut adalah model Machine Learning yang dapat dikembangkan :

Pengembangan Model :

1. Memahami jenis dan tipe data yang akan digunakan untuk pelatihan, uji, dan validasi data.
2. Menstandarisasi format data, menghapus duplikasi dan ambiguitas data, serta menambah dan meningkatkan data.
3. Salah satu model yang dapat digunakan untuk kasus ini adalah regresi logistic untuk menghasilkan probabilitas pelanggan menerima atau menolak penawaran

Pemilihan Fitur :

1. Memilih algoritma dan data yang sesuai dengan kebutuhan.
2. Menggunakan informasi data pada dataset sebagai fitur yang akan dikembangkan.
3. Menambahkan pengkategorian data.

Pelatihan :

1. Menyesuaikan parameter yang ada, seperti membagi data menjadi data pelatihan dan data validasi.
2. Memilih algoritma pelatihan yang tepat, sesuai dengan model machine learning.

Validasi :

1. Melakukan evaluasi model pada data validasi untuk melihat seberapa baik model dapat memprediksi pada data baru.
2. Menggunakan metrik untuk mengetahui kinerja model, salah satunya akurasi, yaitu jumlah prediksi benar dibagi jumlah total prediksi.

Evaluasi :

1. Melakukan uji model pada data tes, yaitu data yang tidak digunakan sama sekali dalam proses pelatihan dan validasi.
2. Menggunakan metrik untuk mengetahui kinerja model, seperti metrik presisi dan recall.

Latihan 5

Berikut adalah beberapa wawasan yang dapat diambil :

1. Menemukan produk unggulan berdasarkan jumlah *review* dan *love*.
2. Menemukan produk teratas dan terbawah berdasarkan *rating*, *review*, dan *love*.
3. Menemukan *category* dan *brand* yang paling diminati berdasarkan *rating*, *review*, dan *love*.
4. Menemukan hubungan antara jumlah *review* dengan tipe *online only*, *exclusive*, *limited _edition*, dan *limited_time_offer*.
5. Menemukan apakah harga jual (*price*) memiliki hubungan dan pengaruh terhadap *size*.

Berikut adalah beberapa fitur dan pengembangan untuk machine learning berdasarkan data yang ada :

1. Menangani data yang salah, hilang, atau duplikasi.
2. Menambahkan fitur pengkategorian.
3. Menambahkan fitur dasar statistik, seperti nilai rata-rata, nilai terbesar/terkecil, dan lainnya.
4. Menambahkan fitur pengubah data numerik menjadi tipe tertentu.
5. Melakukan validasi dan evaluasi model machine learning.

Berikut adalah beberapa grafik yang dapat digunakan untuk melakukan visualisasi berdasarkan data yang ada :

1. Menggunakan histogram terhadap data numerik.
2. Menggunakan scatter plots untuk mencari hubungan antar kolom data.
3. Menggunakan bar plot untuk mencari hubungan antara tipe data.