|  |  |
| --- | --- |
| **FR.IA.02.** | **LEMBAR KERJA** |

|  |  |  |  |
| --- | --- | --- | --- |
| Skema Sertifikasi (~~KKNI~~/Okupasi/~~Klaster~~) | Judul | : | **Ilmuwan Data Madya (*Associate* *Data Scientist*)** |
| Nomor | : | RT-032 |
| TUK | | : | Sewaktu/~~Tempat Kerja/Mandiri~~\* |
| Nama Asesor | | : |  |
| Nama Asesi | | : |  |
| Tanggal | | : |  |

\*Coret yang tidak perlu

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.004.1** |
| **Judul Unit** | **:** | **Menentukan kebutuhan data** |
|  |  | **(1.1) (1.2) (1.3)** |
|  |  | **Mengambil data** |
|  |  | **(2.1) (2.2) (2.3) (2.4)** |
|  |  | **Mengintegrasikan data** |
|  |  | **(3.1) (3.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1.** |
| Tentukan dan tuliskan domain permasalahan yang akan Anda selesaikan! |
|  |
| **Langkah Kerja 2.** |
| Tuliskan kebutuhan data yang akan Anda perlukan untuk menyelesaikan permasalahan! |
|  |
| **Langkah Kerja 3.** |
| Tentukan sumber data yang akan Anda gunakan serta jelaskan metode pengambilan data yang akan Anda terapkan! |
|  |
| **Langkah Kerja 4** |
| 4.1 Hal-hal yang perlu diperhatikan jika Anda akan menggunakan lebih dari 1 sumber data! |
|  |
| 4.2 Strategi atau langkah-langkah yang perlu Anda lakukan dalam mengintegrasikan data yang berasal dari sumber yang berbeda? |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.005.1** |
| **Judul Unit** | **:** | **Menganalisis tipe dan relasi data** |
|  |  | **(1.1) (1.2) (1.3)** |
|  |  | **Menganalisis karakteristik data** |
|  |  | **(2.1) (2.2) (2.3)** |
|  |  | **Membuat laporan telaah data** |
|  |  | **(3.1) (3.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| 1.1. Tuliskan fitur apa saja yang dimiliki oleh dataset serta jelaskan tipe data dari dataset tersebut. |
|  |
| 1.2 Buatlah diagram histogram dari data yang bertipe numerikal dan kategorikal kemudian tulislah hasil kesimpulan yang dapat Anda simpulkan dari hasil visualisasi distribusi data dengan diagram histogram tersebut |
|  |
| 1.3 Nilai atribut data yang terkumpul diuraikan sesuai dengan batasan konteks bisnisnya |
| Gambar di atas adalah representasi visual sederhana dari dataset dalam bentuk dataframe. Setiap baris mewakili satu entri dalam dataset, dan kolom-kolom mewakili atribut yang telah Anda sebutkan sebelumnya  Berikut adalah uraian nilai-nilai atribut dalam dataset berdasarkan batasan konteks bisnis yang telah Anda jelaskan:  1. \*\*ID\*\*: Nilai-nilai ini adalah identifikasi unik untuk setiap entri dalam dataset. Digunakan untuk mengidentifikasi dan membedakan setiap entri dari yang lainnya.  2. \*\*Warehouse\_block\*\*: Nilai-nilai ini mewakili blok gudang di mana produk disimpan. Ini membantu dalam mengorganisasi dan mengelola inventaris di gudang yang berbeda.  3. \*\*Mode\_of\_Shipment\*\*: Nilai-nilai ini menunjukkan metode pengiriman yang digunakan untuk mengirimkan produk kepada pelanggan. Ini penting untuk mengatur proses pengiriman dan menentukan biaya serta waktu pengiriman.  4. \*\*Customer\_care\_calls\*\*: Nilai-nilai ini mencerminkan jumlah panggilan yang dilakukan oleh pelanggan ke layanan pelanggan. Ini membantu dalam mengevaluasi seberapa sering pelanggan membutuhkan bantuan atau memiliki pertanyaan.  5. \*\*Customer\_rating\*\*: Nilai-nilai ini adalah peringkat yang diberikan oleh pelanggan terhadap produk atau layanan. Ini membantu dalam menilai kepuasan pelanggan dan kualitas produk atau layanan.  6. \*\*Cost\_of\_the\_Product\*\*: Nilai-nilai ini adalah biaya produk atau barang. Ini digunakan untuk menghitung margin keuntungan dan menentukan harga jual.  7. \*\*Prior\_purchases\*\*: Nilai-nilai ini menunjukkan jumlah pembelian sebelumnya yang telah dilakukan oleh pelanggan. Ini dapat membantu dalam mengidentifikasi pelanggan yang sering berbelanja.  8. \*\*Product\_importance\*\*: Nilai-nilai ini menggambarkan tingkat pentingnya produk dalam konteks bisnis. Ini dapat membantu dalam mengelola persediaan dengan lebih efektif.  9. \*\*Gender\*\*: Nilai-nilai ini adalah jenis kelamin pelanggan. Ini dapat digunakan untuk analisis demografi dan pemasaran yang lebih terarah.  10. \*\*Discount\_offered\*\*: Nilai-nilai ini adalah diskon yang ditawarkan pada produk. Ini mempengaruhi harga jual dan dapat mempengaruhi daya tarik produk bagi pelanggan.  11. \*\*Weight\_in\_gms\*\*: Nilai-nilai ini adalah berat produk dalam gram. Ini penting untuk menghitung biaya pengiriman dan mengatur metode pengiriman yang sesuai.  12. \*\*Reached.on.Time\_Y.N\*\*: Nilai-nilai ini menunjukkan apakah produk mencapai pelanggan tepat waktu atau tidak. Ini berhubungan dengan efisiensi proses pengiriman. |
| 1.4 Relasi antar data yang terkumpul diidentifikasi sesuai dengan tujuan teknis |
|  |
| **Langkah Kerja 2** |
| 2.1 Karakteristik data yang terkumpul disajikan dengan deskripsi statistik dasar (meliputi nilai maksimum, minimum, rerata, median, modus, *skewing*, persentil, distribusi, *outliers* dan lain sejenisnya) dan visualisasi grafik |
| A screenshot of a computer  Description automatically generated |
| 2.2 Tuliskan nilai unique dari setiap fitur kategori yang ada pada dataset. |
| A screen shot of a computer code  Description automatically generated |
| 2.3 Lakukan analisis kuartil data dan identifikasi pencilan dengan menggunakan visualisasi boxplot. Tuliskan fitur apa yang memiliki pencilan! |
| A screenshot of a graph  Description automatically generated  Pada gambar diatas terdapat kesimpulan bahwa fitur yang memiliki pencilan(outlier) adalah ‘Discount\_offerend’ dan ‘Prior\_purchases’ |
| **Langkah Kerja 3** |
| Tuliskan hasil analisis dalam bentuk laporan dokumentasi sesuai dengan tujuan teknis (Data Understanding Documentation), dengan hipotesis berdasarkan hasil analisis sesuai tujuan teknis data science |
| Data Understanding Documentation  \*Tujuan Teknis Data Science: Menganalisis Distribusi dan Identifikasi Pencilan dalam Dataset Pengiriman Produk\*  ---  1. Pendahuluan  Dokumentasi ini bertujuan pada langkah-langkah Data Understanding dalam proses Data Science. Tujuan dari analisis ini adalah untuk memahami distribusi data dalam dataset pengiriman produk, serta mengidentifikasi potensi pencilan (outliers) dalam atribut yang relevan. Analisis ini dilakukan untuk memberikan wawasan awal tentang karakteristik data, yang nantinya dapat digunakan dalam pengambilan keputusan bisnis.  2. Analisis Distribusi Data  Analisis dimulai dengan memeriksa distribusi data dalam atribut dengan tipe data selain ‘object’. Visualisasi boxplot digunakan untuk mengidentifikasi adanya nilai-nilai pencilan.  A screenshot of a graph  Description automatically generated  Dari visualisasi boxplot, terlihat bahwa terdapat beberapa pencilan yang berada di atas batas atas. Pencilan ini mungkin mengindikasikan bahwa terdapat yang berada di luar kisaran nilai yang diharapkan atau normal dalam distribusi data.  4. Identifikasi Pencilan  Pencilan dalam atribut ‘Discount\_offered’ dan ‘Prior\_purchases’ diidentifikasi sebagai nilai-nilai yang berada di atas batas atas dalam boxplot. Pencilan ini dapat merujuk pada produk-produk yang memiliki biaya yang sangat tinggi dibandingkan dengan produk lainnya. Identifikasi ini memberikan peluang untuk lebih memahami faktor-faktor yang memengaruhi biaya produk yang tidak biasa tersebut.  \*\*5. Hipotesis Berdasarkan Hasil Analisis\*\*  Berdasarkan hasil analisis, beberapa hipotesis yang mungkin dapat diajukan adalah:  - Produk dengan biaya yang sangat tinggi mungkin merupakan produk khusus atau langka yang memiliki atribut atau fitur unik.  - Adanya perbedaan biaya produk yang signifikan mungkin disebabkan oleh variasi dalam bahan, produksi, atau metode pengiriman.  \*\*6. Kesimpulan\*\*  Analisis Data Understanding telah memberikan wawasan tentang distribusi dan karakteristik data dalam dataset pengiriman produk. Pencilan dalam biaya produk dapat menjadi titik fokus analisis lanjutan untuk memahami lebih dalam faktor-faktor yang memengaruhi biaya yang tidak biasa tersebut. Hasil analisis ini dapat membantu dalam pengambilan keputusan bisnis yang lebih terinformasi dalam mengelola biaya produk dan strategi pengiriman. |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.006.1** |
| **Judul Unit** | **:** | **Melakukan pengecekan kelengkapan data** |
|  |  | **(1.1) (1.2)** |
|  |  | **Membuat rekomendasi kelengkapan data** |
|  |  | **(2.1) (2.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| 1.1. Tuliskan kolom mana saja yang masih memiliki nilai missing value (null, NaN, ataupun duplikasi dan ambigu) |
| A screenshot of a computer  Description automatically generated  Dari gambar diatas dapat disimpulkan bahwa data yang digunakan tidak ada nilai null dan duplikasi dalam datanya. |
| 1.2 Tuliskan jumlah persentase nilai missing value pada kolom fitur tersebut. |
| A screenshot of a computer  Description automatically generated  Dari gambar diatas diketahui tidak ada missing value untuk setiap kolom sehingga persentase missing value untuk setiap kolom adalah 0 |
| 1.3 Jelaskan apakah terdapat fitur yang memiliki nilai benar atau tidak sesuai dengan fakta (misal: gender berisi lebih dari dua nilai, dlsb) |
| A screen shot of a computer code  Description automatically generated  Dari fitur kategori diatas didapatkan kesimpulan bahwa semua fitur tersebut tidak ada data yang tidak sesuai. Contoh nya kolom gender dimana hanya terdapat 2 gender yaitu perempuan dan laki laki. |
| 1.4 Tuliskan nilai unique dari masing-masing kolom fitur yang berjenis kategori! Apakah terdapat nilai redundan? Jika Ya, fitur apa saja yang memiliki nilai redundan? Contoh: misalnya terdapat kolom kategori pendidikan yang mana ditemukan nilai fitur SMA dan SLTA. Nilai tersebut dituliskan dalam kodefikasi berbeda, namun menunjukkan informasi yang sama. |
| A screen shot of a computer code  Description automatically generated  Dari fitur kategori diatas tidak ada nilai redudan atau biasa disebut fitur yang memiliki informasi hampir sama |
| **Langkah Kerja 2** |
| 2.1 Rekomendasi hasil penilaian kualitas data terkait kebenaran, kelengkapan, dan konsistensi data |
| Dari hasil analisis data sebelumnya dapat disimpulkan bahwa data sudah tergolong bagus dikarenakan data yang digunakan sudah lengkap(tidak ada duplikasi dan missing value), juga tidak ada data yang menyimpang setiap kolomnya. |
| 2.2 Rekomendasi hasil penilaian kecukupan data terkait cukup atau perlu diulang pengumpulan data ulang ataupun proses resampling. |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.007.1** |
| **Judul Unit** | **:** | **Memutuskan kriteria dan teknik pemilihan data** |
|  |  | **(1.1) (1.2)** |
|  |  | **Menentukan attributes (columns) dan records (row) data** |
|  |  | **(2.1) (2.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| 1.1. Lakukan analisis kriteria pemilihan data mencakup kuantitas data (volume data yang menggambarkan ukuran data misalkan dalam terabyte, petabyte, atau jumlah record) dan kualitas data ( nilai null, NaN, duplikasi, ataupun ambigu), berdasarkan analisis kelengkapan data yang telah Anda lakukan sebelumnya. Tuliskan hasil temuan Anda. |
| A screenshot of a computer  Description automatically generated  Dari gambar diatas bisa didapatkan informasi terkait data yang digunakan. Pada data tersebut terdapat 10999 jumlah record data dengan 12 kolom atau fitur, pada data tersebut tidak ditumukan nilai null atau duplikasi, dan juga ukuran memori yang digunakan pada data tersebut sebesar +- 1MB |
| 1.2 Lakukan analisis Kriteria bisa berbentuk ketentuan mengenai pencilan, korelasi antara atribut, dan lain sebagainya. Tuliskan hasil temuan Anda. |
| A screenshot of a graph  Description automatically generated  Pada gambar diatas terdapat kesimpulan bahwa fitur yang memiliki pencilan(outlier) adalah ‘Discount\_offerend’ dan ‘Prior\_purchases’.  A screenshot of a graph  Description automatically generated  Pada gambar diatas merupakan kerelasi menggunakan heatmap, dimana dapat disimpulkan bahwa kolom atau fitur yang mempunyai nilai korelasi terbesar adalah ‘discount\_offered’ dan ‘reacher.on.Time\_Y.N’ dengan nilai sebesar 0.4 |
| 1.3 Tentukan teknik pemilihan data dalam pengambilan sampel yang akan Anda gunakan, secara umum dibagi menjadi dua teknik yaitu probability sampling atau random sampling dan non probability sampling. |
|  |
| **Langkah Kerja 2** |
| 2.1 Tentukan attributes/fitur (columns) mana yang akhirnya akan Anda gunakan dalam membangun model machine learning untuk menyelesaikan permasalahan yang tertera pada skenario! |
|  |
| 2.2 Tentukan jumlah data (rows) mana yang akhirnya akan Anda gunakan dalam membangun model machine learning untuk menyelesaikan permasalahan yang tertera pada skenario! |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.008.1** |
| **Judul Unit** | **:** | **Melakukan pembersihan data yang kotor** |
|  |  | **(1.1) (1.2)** |
|  |  | **Membuat laporan dan rekomendasi hasil membersihkan data** |
|  |  | **(2.1) (2.2) (2.3)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| 1.1. Strategi pembersihan data ditentukan berdasarkan hasil telaah data, yaitu dapat berupa pengisian dengan nilai yang tepat (mean, median, min/max, mode, dan lain-lain), koreksi nilai standar, diisi dengan konstanta, menghapus baris kosong dan lain-lain |
|  |
| 1.2 Data yang kotor dikoreksi berdasarkan strategi pembersihan data, yaitu dapat berupa data terstruktur maupun tidak terstruktur berupa missing value, data yang salah, dan data outlier. |
|  |
| **Langkah Kerja 2** |
| 2.1 Rekomendasi adalah tindak lanjut dari proses pembersihan data, dimana permintaan atas kebutuhan disesuaikan dengan standard di organisasi tersebut |
|  |
| 2.2 Masalah dan teknis koreksi data dideskripsikan sesuai dengan kondisi data dan strategi pembersihan data |
|  |
| 2.3 Evaluasi dihasilkan dari proses berdasarkan analisis koreksi yang telah dilakukan dan hasilnya kemudian di dokumentasikan. |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.009.1** |
| **Judul Unit** | **:** | **Menganalisis teknik transformasi data** |
|  |  | **(1.1) (1.2)** |
|  |  | **Melakukan transformasi data** |
|  |  | **(2.1) (2.2)** |
|  |  | **Membuat dokumentasi konstruksi data** |
|  |  | **(3.1) (3.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| 1.1. Analisis data untuk menentukan representasi fitur data awal, yaitu dapat berupa kolom data atau fitur tipe data yang dapat digunakan algoritma machine learning sesuai dengan tipe data. |
|  |
| 1.2 Analisis representasi fitur data dilakukan untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science |
|  |
| **Langkah Kerja 2** |
| 2.1 Transformasi dilakukan untuk mendapatkan fitur data awal. |
|  |
| 2.2 Rekayasa fitur data dilakukan untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science, yaitu dapat berupa normalisasi, pemilihan fitur tipe data baru, menambahkan kolom data baru |
|  |
| **Langkah Kerja 3** |
| 3.1 Dokumentasi berisikan teknis transformasi data dijabarkan dalam bentuk tertulis. |
|  |
| 3.2 Dokumentasi berisikan hasil transformasi data dan rekomendasi hasil transformasi dituangkan dalam bentuk tertulis |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.010.1** |
| **Judul Unit** | **:** | **Melakukan pelabelan data** |
|  |  | **(1.1) (1.2)** |
|  |  | **Membuat laporan hasil pelabelan data** |
|  |  | **(2.1) (2.2)** |

**LEMBAR KERJA**

| **Langkah Kerja 1** |
| --- |
| Lakukan proses pelabelan data atau proses encoding pada kolom data kategori. |
|  |
| **Langkah Kerja 2** |
| Buatlah laporan hasil pelabelan data atau proses encoding yang telah Anda lakukan, mencakup statistik hasil pelabelan dan evaluasi proses pelabelan. |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.013.1** |
| **Judul Unit** | **:** | **Menyiapkan parameter model** |
|  |  | **(1.1) (1.2)** |
|  |  | **Menggunakan tools pemodelan** |
|  |  | **(2.1) (2.2)(2.3) (2.4)** |

**LEMBAR KERJA**

| **Langkah Kerja 1** |
| --- |
| Lakukan identifikasi parameter yang dibutuhkan dalam membangun model, dengan nilai toleransi parameter evaluasi pengujian sesuai tujuan teknis. |
|  |
| **Langkah Kerja 2** |
| Tentukan tools yang akan digunakan dan algoritma yang akan Anda gunakan untuk membangun model, sesuai dengan skenario yang telah ditetapkan (termasuk di dalamnya sesuai dengan skenario pengujian)! |
|  |
| **Langkah Kerja 3** |
| Lakukan tahapan cross-validation untuk menentukan konfigurasi parameter model yang terbaik |
|  |
| **Langkah Kerja 4** |
| Tuliskan hasil konfigurasi yang paling optimal yang didapatkan dari hasil cross-validation tersebut. |
|  |

|  |  |  |
| --- | --- | --- |
| **Kode Unit** | **:** | **J.62DMI00.014.1** |
| **Judul Unit** | **:** | **Menggunakan model dengan data riil** |
|  |  | **(1.1) (1.2)** |
|  |  | **Menilai hasil pemodelan** |
|  |  | **(2.1) (2.2)** |

**LEMBAR KERJA**

|  |
| --- |
| **Langkah Kerja 1** |
| Lakukan uji model dengan data testing yang telah Anda persiapkan |
|  |
| **Langkah Kerja 2** |
| Lakukan pengukuran model yang dibangun dengan menggunakan metrik performance yang telah Anda pilih sebelumnya! |
|  |
| **Langkah Kerja 3** |
| Tuliskan hasil pengukuran yang Anda dapatkan sebagai dokumentasi sesuai standar yang berlaku! |
|  |