

3.3 Penyediaan Data (P3)

3.3.1 Pengenalan

Peringkat ini melibatkan aktiviti perolehan, pengumpulan, eksplorasi dan transformasi data bagi mendapat pemahaman yang lebih baik, mengenal pasti keperluan bagi pembersihan data, peluang bagi pemerkayaan data iaitu mengenal pasti data lain yang dapat meningkatkan kualiti analitis, dan mengenal pasti analisis yang boleh dibuat ke atas data tersebut. Setelah data diperolehi (sama ada keseluruhan atau subset data), eksplorasi data iaitu kajian terhadap data perlu dilakukan untuk mendapatkan pemahaman tentang kualiti data, corak data, kelainan dan kekurangan yang wujud dalam data. Terma yang sering digunakan bagi aktiviti ini ialah Analisis Penerokaan Data. Rajah 3.5 menunjukkan kedudukan Peringkat 3 dalam metodologi DRSA.



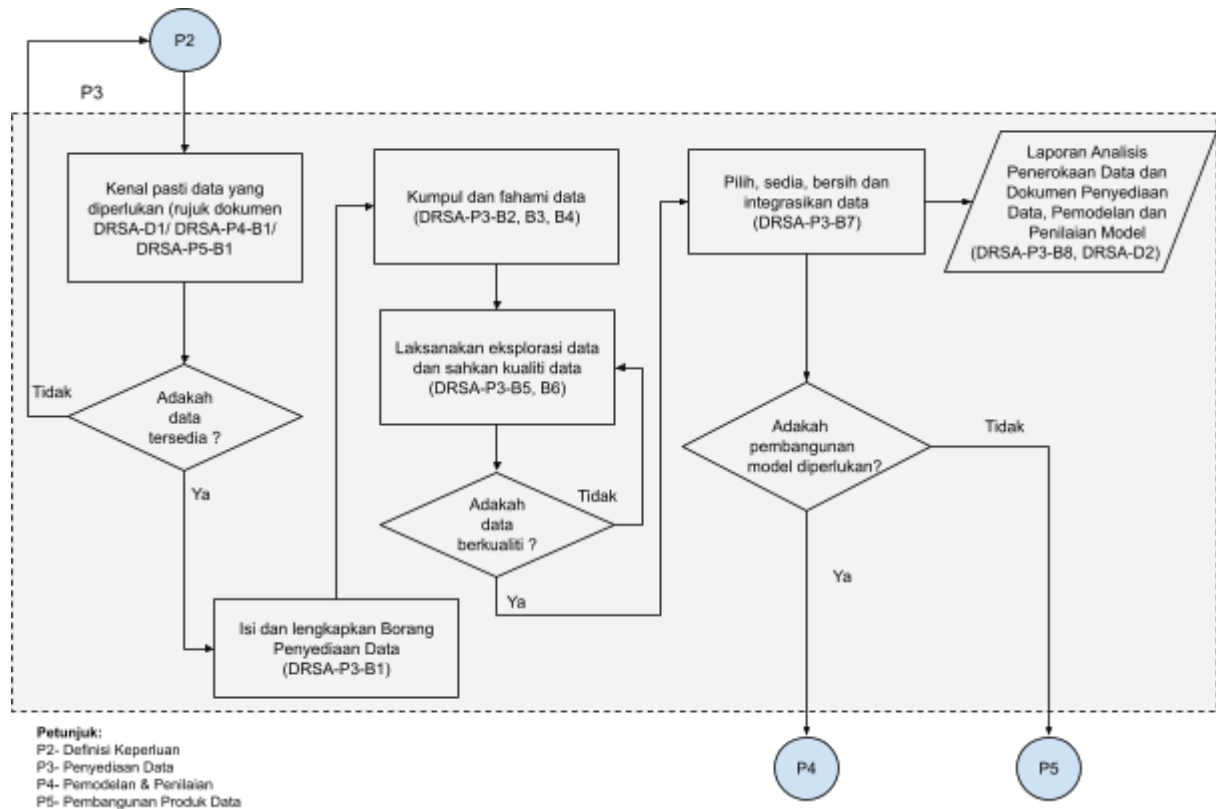
Rajah 3.5: Kedudukan Peringkat 3 Dalam Metodologi DRSA

3.3.2 Hasil Serahan

Lampiran C: Dokumen Penyediaan Data, Pemodelan dan Penilaian Model (DRSA-D2).

3.3.3 Carta Alir

Aktiviti Penyediaan Data boleh dirujuk pada Rajah 3.6.



Rajah 3.6: Carta Alir Penyediaan Data

3.3.4 Borang dan Panduan

3.3.4.1 DRSA-P3-B1: Borang Penyediaan Data

Borang DRSA-P3-B1 bertujuan untuk mengenal pasti data yang diperlukan dari pelbagai sumber.

DRSA-P3-B1: Borang Penyediaan Data

Nama Set Data	Atribut	Cadangan Sumber Data							Tempoh Masa	Garis Masa Data			Transformasi Data Yang Diperlukan (jika ada)
		Manual		Pangkalan Data	Sumber lain (termasuk data tidak berstruktur)					Data Sejarah	Data Semasa	Data Masa-Nyata	
		Jenis	Nama dan Lokasi Fail	Sumber Pangkalan Data	Data Diperlukan	Kata Kunci	Cadangan Sumber	Pembekal Data (cth:Beli data di luar)					
Data kesalahan trafik	IdKesalahan JenisKesalahan	-	-	myJPJ	-	-	-	-	5 tahun terkini	√	√		
Data media sosial	Timestamp TwitterId Text	-	-	-	FB Post Tweet	Carian: - Kemalangan - JPJ	Facebook Twitter	-	7 hari terkini		√	√	

1
Sumber data selain pangkalan data seperti csv, pdf, txt, xls dan lain-lain

2
Proses transformasi yang dilakukan terhadap data (imputasi, kejuruteraan fitur, sample data)

3.3.4.2 DRSA-P3-B2: Borang Kamus Data

Borang DRSA-P3-B2-1 sehingga DRSA-P3-B2-4 bertujuan untuk mewujudkan kamus data sekiranya kamus data tidak tersedia. Ia juga bergantung kepada jenis sumber data dan data yang diperlukan. Agensi disarankan merujuk *Data Dictionary* Sektor Awam (DDSA) untuk mereka bentuk kamus data yang mempunyai standard/piawaian kod data.

i. Kamus Data Bagi Pangkalan Data

Borang ini digunakan untuk data yang diambil daripada pangkalan data atau fail.

DRSA-P3-B2-1: Borang Kamus Data Bagi Pangkalan Data

No.	Jadual	Atribut	Jenis Data	Penerangan	Format Data (Jika Berkenaan)	Saiz Atribut	Kunci Utama	Kunci Asing		Data Terbitan		Data Agregat		
							(Y/T)	(Y/T)	Jika (Y), Masukkan Nama Jadual	(Y/T)	Fungsi Terbitan (Formula)	(Y/T)	Fungsi Agregat (sum/min/max/count/avg)	Dikumpul mengikut (baris/lajur)
1	Kesalahan Trafik	IdKesalahan	integer	Id untuk setiap kesalahan		10	Y	T	-	T	-	T	-	-
2	Kesalahan Trafik	JenisKesalahan	char	Jenis kesalahan trafik		30	T	T	-	T	-	Y	sum	lajur
3	Kesalahan Trafik	Tahun Lahir	date	Tahun lahir pesalah	dd/mm/yyyy	-	T	T	-	T	-	T	-	-
4	Kesalahan Trafik	umur	integer	Umur pesalah trafik		2	T	T	-	Y	Tahun Semasa-Tahun Lahir	T	-	-

ii. Kamus Data Bagi Gudang Data

Borang ini digunakan untuk data yang diambil daripada gudang data.

a. Jadual Fakta

DRSA-P3-B2-2: Borang Kamus Data Bagi Gudang Data - Jadual Fakta

No.	Atribut	Jenis Data	Penerangan	Kunci Asing		Data Agregat		Data Terbitan
				(Y/T)	Jika (Y), masukkan nama Jadual Dimensi	(Y/T)	Fungsi Agregat (sum/min/max/count/avg)	
1	Tarikh	Date	Tarikh kesalahan	Y	Jadual Tarikh	T		
2	Lokasi	Char	Lokasi kesalahan	Y	Jadual Lokasi	T		
3	JenisKesalahan	integer	Jenis kesalahan	Y	Jadual Kesalahan	T		

No.	Atribut	Jenis Data	Penerangan	Kunci Asing		Data Agregat		Data Terbitan
				(Y/T)	Jika (Y), masukkan nama Jadual Dimensi	(Y/T)	Fungsi Agregat (sum/min/max/count/avg)	
4	JumlahKesalahan	integer	Jumlah jenis kesalahan pada tarikh dan lokasi	T	-	Y	sum	1

Data yang diterbitkan daripada data-data lain

b. Jadual Dimensi

DRSA-P3-B2-3: Borang Kamus Data Bagi Gudang Data - Jadual Dimensi

No.	Nama Jadual Dimensi	Atribut	Jenis Data	Penerangan	Kunci Utama/Asing	Jika Kunci Asing, masukkan nama jadual
1	KesalahanTrafik	IdKesalahan	integer	Id untuk setiap kesalahan	Kunci Utama	-
2		JenisKesalahan	char	Jenis kesalahan trafik	-	-

iii. Kamus Data Bagi Tasik Data

Borang ini digunakan untuk data yang diambil daripada tasik data.

DRSA-P3-B2-4: Borang Kamus Data Bagi Tasik Data

No.	Nama Koleksi	ID Dokumen	Kunci
1	Maklumat Kemalangan Dari Media Sosial Aplikasi X	123456	Timestamp Id Text

3.3.4.3 DRSA-P3-B3: Borang Metadata Data Tidak Berstruktur

Borang DRSA-P3-B3 bertujuan untuk merekodkan data tidak berstruktur seperti teks, video, imej dan audio.

DRSA-P3-B3: Borang Metadata Data Tidak Berstruktur

No.	ID Video/Imej/Audio	URL>Nama Fail	Tajuk	Jenis/Format Fail	Durasi Video/Audio	Timestamp	Kata Kunci/Tag	Penerangan/Kapsyen	Sumber/Pembekal
1	V1	JPJ/cctv/kemalangan	Kemalangan Jalan Duta	MP4	2 minit	20042023 1130	Jalan Duta LanggarLari	Sebuah kereta jenama tempatan telah melanggar lari sebuah motosikal di Jalan Duta	CCTV Jalan Duta (DBKL)

3.3.4.4 DRSA-P3-B4: Borang Katalog API

Borang DRSA-P3-B4 bertujuan untuk mewujudkan katalog API yang diperlukan untuk penyediaan data.

DRSA-P3-B4: Borang Katalog API

No.	Nama API	Penerangan	Kategori	Agensi Pembekal	Agensi Penerima	Kategori Data (Sulit/ Terbuka/ Terhad)	URL API	Kunci Pengesahan (Auth)	Parameter (Nama Medan)	Keterangan Parameter	Jenis Data	Kaedah (POST/GET/DELETE/ PUT/PATCH)	Jenis Data Dikembalikan (Return Data) cth:JSON, SML,CSV
1	GetData Asas	Untuk mendapatkan data asas bagi kesalahan trafik	Kesalahan trafik	JPN	JPJ	Sulit	https://api.example.com/GetDataAsas	XYZ-4662	no_kp, nama	No Kad Pengenalan, Nama	varchar	POST, GET	JSON
			1					2	3				

Kategori API (berdasarkan kategori sistem, contoh: Sumber Manusia, Kewangan dll)

Kunci untuk mengesahkan transaksi boleh dilakukan

Parameter yang digunakan untuk menghantar dan/atau menerima API tersebut

3.3.4.5 DRSA-P3-B5: Borang Tahap Kualiti Data

Pembekal data hendaklah menyediakan data yang berkualiti. Data yang berkualiti adalah penting bagi mengelakkan masalah data hilang, duplikasi data atau kemasukan data yang silap. Ini kerana kos untuk membersihkan data dan sumber data adalah tinggi. Borang DRSA-P3-B5 bertujuan untuk memastikan data yang digunakan boleh dipercayai dan berkualiti.

Bagi memastikan data adalah berkualiti, panduan untuk mencapai standard data yang minimum perlu disediakan. Panduan hendaklah merangkumi standard kualiti data minimum dan kriteria kualiti data yang diperlukan. Semakan kualiti data perlu dilaksanakan setiap kali data diimport ke dalam sumber penyimpanan data (pangkalan data, gudang data atau tasik data) atau data diubah menjadi format baharu.

DRSA-P3-B5: Borang Tahap Kualiti Data

No.	ID	Tajuk	Penerangan	Indikator	Dimensi Kualiti Data	Unit Ukuran/Formula	Nilai Sasaran	Nilai Ambang	Kekerapan Semakan	Punca Masalah	Tindakan Pembaikan Diperlukan?
1	NoKP	Duplikasi No KP, Tarikh, Masa, Jenis Kesalahan	Duplikasi No KP, Tarikh, Masa, Jenis Kesalahan	Bilangan duplikasi No KP, Tarikh, Masa, Jenis Kesalahan mesti 0%	Unik	Peratusan Duplikasi	0%	99% - 100% (hijau) 98% - 50% (kuning) <50% (merah)	Mingguan	Praktis bisnes yang tidak tepat dan kesilapan memasukkan data	☐ Ya ☐ Tidak Jika Ya, nyatakan
				1	2	3	4	5			

Peratusan yang dibenarkan untuk atribut yang tidak tepat/lengkap

Dimensi kualiti data yang diperlukan untuk atribut (Rujuk Jadual 3.1)

Nilai terendah yang dibenarkan sebagai indikator kualiti atribut

Nilai ambang untuk menunjukkan status kualiti atribut (Rujuk DRSA-P3-B6)

Formula yang digunakan untuk mengukur tahap kualiti data (Rujuk Rajah 3.7)

3.3.4.6 Panduan Kriteria Kualiti Data

Kriteria kualiti data adalah set standard yang ditetapkan untuk mengukur tahap kualiti data dalam sesuatu set data atau pangkalan data. Jadual 3.1 menerangkan kriteria kualiti data yang sering diukur.

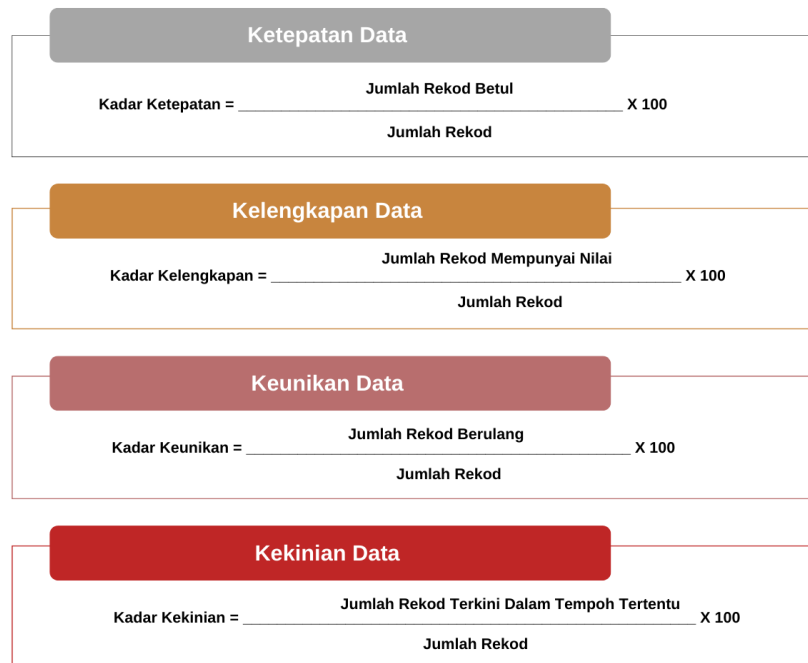
Jadual 3.1: Kriteria Kualiti Data

Kriteria Kualiti Data	Keterangan	Contoh Situasi Tidak Memenuhi Kriteria	Amalan Baik
Kelengkapan	Mengukur tahap data yang hilang atau tidak boleh digunakan	Dalam pangkalan data pelanggan, beberapa rekod tidak mempunyai alamat e-mel	Memastikan semua data yang diperlukan dikumpulkan semasa pengumpulan data
Konsisten	Mengukur tahap maklumat yang bercanggah	Rekod pelanggan yang sama mempunyai nombor kad pengenalan yang berbeza	Menetapkan aturan pemadanan data yang jelas untuk mengelakkan konflik maklumat

Kriteria Kualiti Data	Keterangan	Contoh Situasi Tidak Memenuhi Kriteria	Amalan Baik
Keunikan	Mengukur tahap data yang tidak berulang	Terdapat dua rekod yang sama dalam pangkalan data pelanggan	Melakukan penilaian berkala dan penyingkiran rekod berganda
Kesahihan	Mengukur tahap data yang sepadan dengan rujukan	Nombor kad pengenalan pelanggan adalah sah dan sepadan dengan rekod rasmi	Membandingkan data dengan sumber rujukan yang sah untuk pengesahan
Kepatuhan	Mengukur tahap data yang disimpan dalam format bukan standard	Penyimpanan tarikh dalam format yang berbeza daripada standard industri	Mematuhi standard data yang ditetapkan untuk format penyimpanan
Ketepatan	Mengukur tahap persetujuan dengan sumber maklumat yang betul	Data harga produk yang serupa dengan harga dalam laporan pembekal	Memeriksa dan membandingkan data dengan sumber yang dipercayai
Integriti	Mengukur tahap data yang rosak	Kerosakan data pada hasil cetakan akibat gangguan sistem	Melakukan penyemakan berkala integriti data dan membuat sandaran data
Kekinian	Mengukur tahap ketersediaan dan kekinian data untuk digunakan dalam tempoh yang ditetapkan	Maklumat stok produk yang dikemas kini setiap jam dan dapat diakses oleh pengguna pada bila-bila masa	Menjadualkan dan memantau pengemaskinian data secara berkala untuk menjaga ketepatan masa

3.3.4.7 Panduan Formula Pengukuran Tahap Kualiti Data

Rajah 3.7 memberi panduan kepada agensi untuk mengukur tahap kualiti data bagi empat (4) kriteria yang sering digunakan.



Rajah 3.7 : Formula Pengukuran Tahap Kualiti Data

3.3.4.8 DRSA-P3-B6: Borang Skor Kualiti Data

Borang DRSA-P3-B6 bertujuan untuk merekod skor kualiti data yang diperoleh bagi setiap atribut yang dinilai mengikut dimensi kualiti data yang telah ditentukan.

DRSA-P3-B6: Borang Skor Kualiti Data

No.	Jadual	Atribut	Kriteria Kualiti Data	Skor	Nilai Ambang		
					Merah	Kuning	Hijau
1	Kesalahan	NoKP	Unik	100	<50	>=50 dan <99	>=99 dan <=100

1

Dimensi kualiti data yang diperlukan untuk atribut (Rujuk Jadual 3.1)

2

Skor = 100% - Nilai Sasaran (Rujuk DRSA-P3-B5)

3

Nilai ambang yang diset untuk menunjukkan status kualiti atribut.
Hijau – menunjukkan kualiti data yang **BAIK**, boleh dipercayai, sesuai untuk analisis dan membuat keputusan.
Kuning – menunjukkan kualiti data **SEDERHANA** dan mungkin memerlukan pengesahan lanjut sebelum digunakan.
Merah – menunjukkan kualiti data yang **TIDAK BAIK** dan **TIDAK SESUAI** digunakan untuk analisis.

3.3.4.9 DRSA-P3-B7: Borang Penyediaan Data Bagi Medan Ujian

Data *sandbox* ialah satu persekitaran khusus yang digunakan untuk meneroka data tanpa menjejaskan persekitaran dan sumber data produksi. Ia berfungsi sebagai ruang yang selamat dan terkawal untuk mencuba, meneroka dan menganalisis data.

Setiap data *sandbox* yang dibangunkan akan direkod dalam borang ini. Borang ini boleh dijadikan sebagai rujukan kepada Peringkat 4 (Pemodelan dan Penilaian) dan Peringkat 5 (Pembangunan Produk Data).

DRSA-P3-B7: Borang Penyediaan Data Bagi Medan Ujian

No.	ID Sandbox	Nama Sandbox	Penerangan	Maklumat Data				
				Tarikh Diterbitkan	Versi	Disahkan Oleh	Pemilik Data	Data Steward
1	SB-JPJ-01	Kesalahan Trafik	Sandbox untuk staging data kesalahan trafik	15 Jun 2022	1.0	En. Shahrul	Bahagian Penguatkuasaan	En. Shahrul

3.3.4.10 DRSA-P3-B8: Borang Laporan Analisis Penerokaan Data

Borang ini digunakan untuk penyediaan laporan Analisis Penerokaan Data bagi mendapatkan gambaran keseluruhan mengenai corak data, kualiti data dan mengenal pasti kelainan dan kekurangan yang wujud dalam data.

DRSA-P3-B8: Borang Laporan Analisis Penerokaan Data

1.0 PENGENALAN <i>Gambaran keseluruhan laporan EDA dan matlamat analisis. Penerangan tentang set data yang sedang dianalisis dan kaitannya dengan projek atau masalah yang dihadapi.</i>
2.0 OBJEKTIF <i>Penerangan tentang objektif laporan EDA yang dilakukan.</i>
3.0 PENERANGAN DATA <i>Penerangan ringkas tentang set data, termasuk:</i> • Sumber – dari mana dan cara data dikumpulkan • Saiz • Format • Penerangan tentang pemboleh ubah atau atribut yang terdapat dalam set data <i>Maklumat tentang latar belakang yang berkaitan dengan data. Jika ia diperoleh daripada organisasi, tinjauan atau projek penyelidikan tertentu, berikan butiran tentang kaedah pengumpulan data dan sebarang kemungkinan bias.</i>

4.0 PEMBERSIHAN DATA

Penerangan tentang langkah-langkah yang diambil untuk membersihkan dan pra-pemrosesan data. Ini mungkin melibatkan pengendalian nilai yang hilang (missing data), menangani data terpencil, menyeragamkan format dan mengeluarkan duplikasi data. Perlu diterangkan juga sebarang andaian atau keputusan yang dibuat semasa proses pembersihan.

5.0 STATISTIK DESKRIPTIF DAN VISUALISASI DATA

Penerangan tentang statistik deskriptif untuk pemboleh ubah utama, termasuk statistik menggunakan ukuran seperti min, median, mod, julat, varian dan sisihan piawai (standard deviation). Gunakan jadual, carta dan graf untuk mewakili secara visual taburan data dan kecenderungan pusat (central tendency). Gunakan plot taburan, histogram, plot kotak, carta bar atau peta haba untuk menganalisis hubungan, corak dan taburan dalam data. Terangkan pandangan yang diperoleh daripada visualisasi dan implikasinya.

6.0 TRANSFORMASI DATA

Penerangan tentang teknik yang digunakan untuk mencipta pemboleh ubah baharu atau mengubah yang sedia ada. Contoh normalisasi, binning dan lain-lain.

7.0 ANALISIS KORELASI

Penerangan tentang analisis korelasi yang digunakan untuk mengkaji hubungan antara pemboleh ubah. Antara kaedah yang boleh digunakan adalah pekali korelasi seperti korelasi Pearson atau Spearman. Kenal pasti korelasi yang signifikan dan bincangkan implikasinya.

8.0 DAPATAN DAN INSIGHTS AWAL

Ringkasan penemuan utama dan insights yang diperoleh daripada EDA. Tumpukan kepada sebarang corak, aliran atau hubungan ketara yang ditemui semasa analisis. Bincangkan implikasi penemuan ini dan kaitannya dengan projek atau masalah yang sedang ditangani.