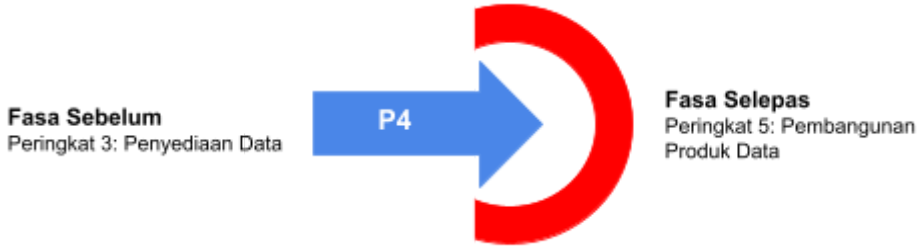
3.4 Pemodelan dan Penilaian (P4)

3.4.1 Pengenalan

Peringkat 4 melibatkan pembangunan dan penilaian prestasi model yang dicadangkan bagi menghasilkan keputusan yang boleh menjawab soalan dalam Spesifikasi Keperluan Bisnes di Peringkat 2. Rajah 3.8 menunjukkan kedudukan Peringkat 4 dalam metodologi DRSA.



Rajah 3.8: Kedudukan Peringkat 4 Dalam Metodologi DRSA

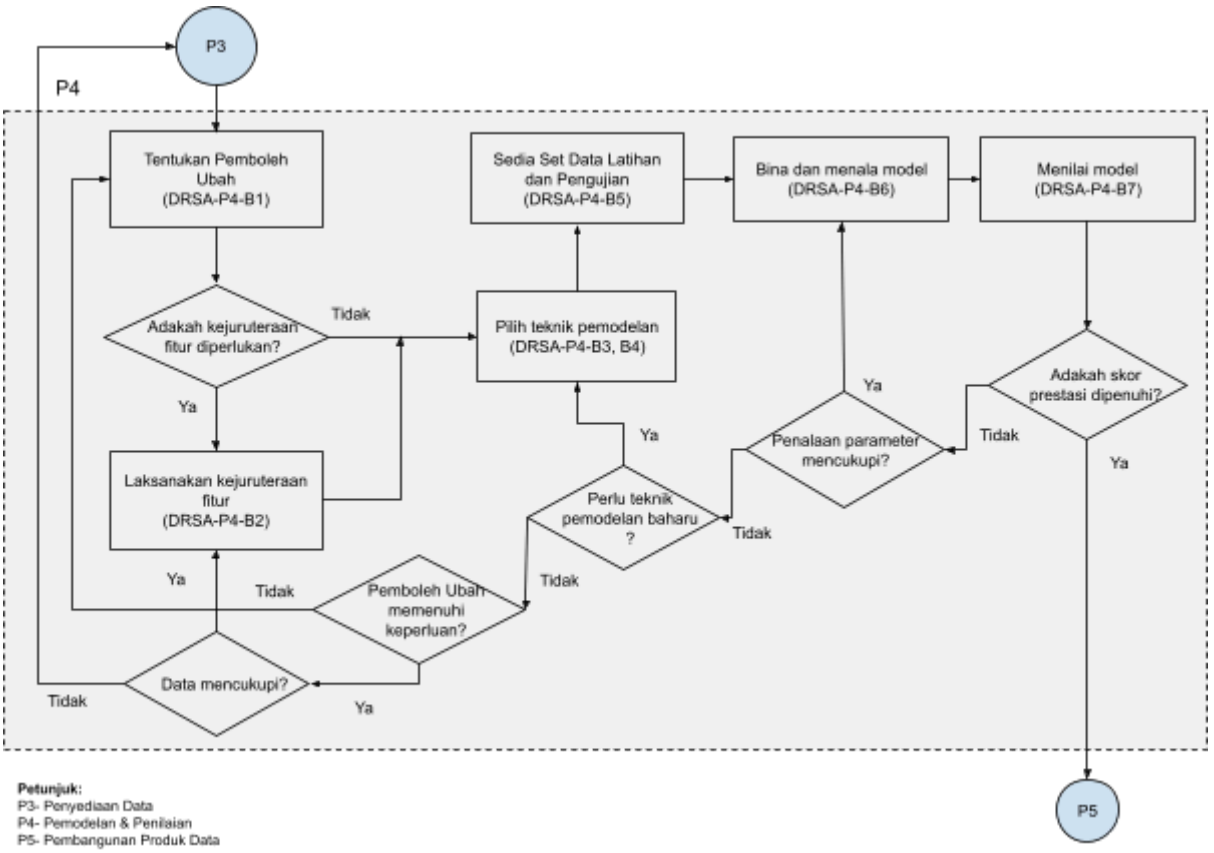
Terdapat pelbagai teknik dan algoritma yang boleh digunakan bagi menghasilkan model yang sesuai berdasarkan tahap analitis yang hendak dilaksanakan. Pelaksanaan peringkat ini melibatkan kejuruteraan fitur, iaitu pengubahan dan manipulasi data asal untuk mencipta ciri-ciri atau atribut baru yang lebih relevan atau informatif bagi model data. Contoh kejuruteraan fitur adalah normalisasi menggunakan Z-Score atau *binning*. Seterusnya model tersebut perlu diuji untuk mengenal pasti prestasi yang terbaik.

3.4.2 Hasil Serahan

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

3.4.3 Carta Alir

Aktiviti Pemodelan dan Penilaian boleh dirujuk pada Rajah 3.9.



Rajah 3.9: Carta Alir Pemodelan dan Penilaian

3.4.4 Borang dan Panduan

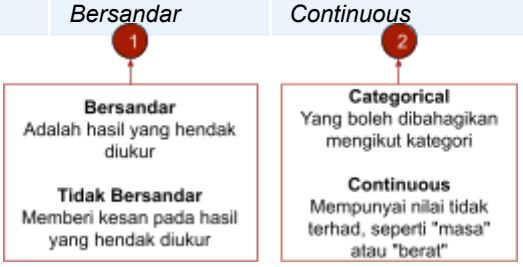
3.4.4.1 DRSA-P4-B1: Borang Penentuan Pemboleh Ubah

Borang DRSA-P4-B1 bertujuan untuk mengenal pasti pemboleh ubah/atribut yang terlibat di dalam analitis, jenis pemboleh ubah sama ada bersandar atau tidak bersandar dan kategori pemboleh ubah seperti *categorical, continuous* dan lain-lain jenis data.

Contoh: Membuat ramalan bilangan jenis kesalahan trafik mengikut peristiwa seperti cuti umum, musim perayaan, cuti sekolah dan sebagainya. Atribut Bilangan Jenis Kesalahan adalah pemboleh ubah bersandar manakala atribut Peristiwa adalah pemboleh ubah tidak bersandar.

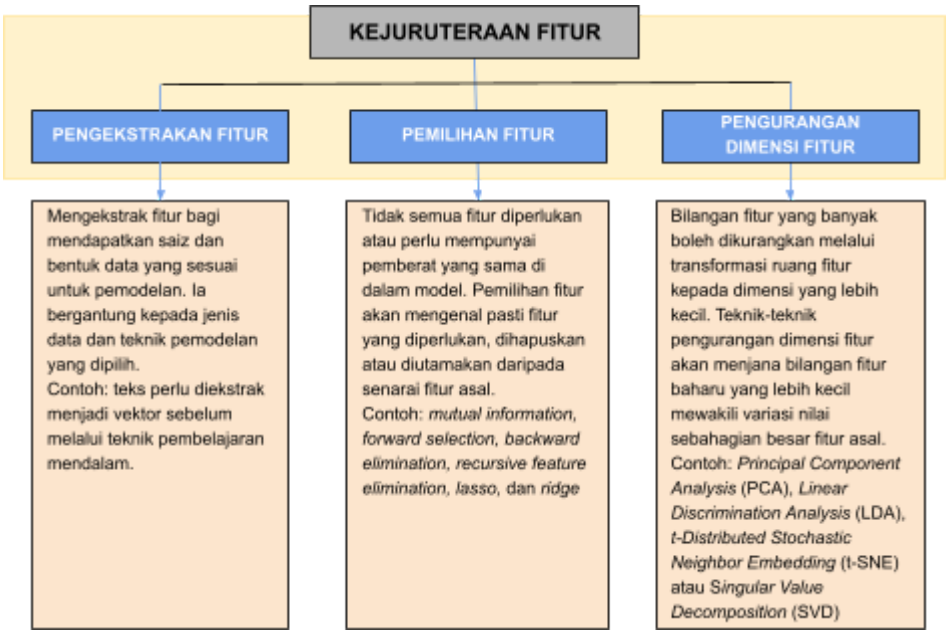
RSA-P4-B1: Borang Penentuan Pemboleh Ubah

Nama Atribut	Jenis Data	Jenis Pemboleh Ubah	Kategori Pemboleh Ubah
Tarikh	*Date*	*Tidak bersandar*	*Continuous*
Peristiwa	*Integer*	*Tidak bersandar*	*Categorical*
Daerah	*Integer*	*Tidak bersandar*	*Categorical*
Jenis Kesalahan	*Integer*	*Tidak bersandar*	*Categorical*
Bilangan Kesalahan	*Integer*	*Bersandar*	*Continuous*



3.4.4.2 DRSA-P4-B2: Borang Kejuruteraan Fitur

Tidak semua fitur boleh digunakan secara langsung di dalam model. Adakalanya fitur tersebut perlu diubah untuk mendapatkan model yang lebih tepat dan mempunyai prestasi tinggi. Terdapat beberapa jenis kejuruteraan fitur iaitu pengekstrakan fitur, pemilihan fitur atau/dan pengurangan dimensi fitur seperti penerangan di Rajah 3.10.



Rajah 3.10: Kejuruteraan Fitur

Borang DRSA-P4-B2 adalah contoh pengisian maklumat kejuruteraan fitur bagi atribut Jenis Kesalahan yang asalnya dalam bentuk teks kepada vektor. Borang ini perlu disesuaikan dengan jenis kejuruteraan fitur yang dilaksanakan berdasarkan pada keadaan data dan keperluan teknik pemodelan.

DRSA-P4-B2: Borang Kejuruteraan Fitur

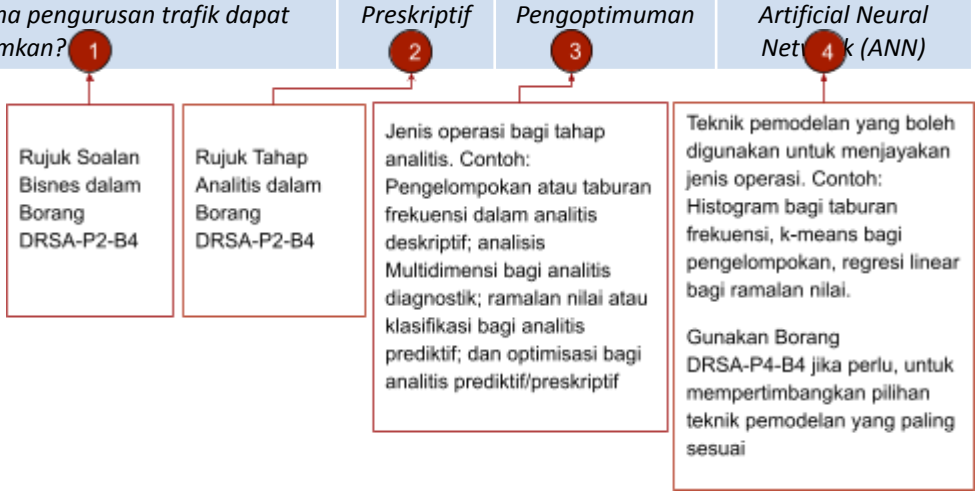
Jenis Kejuruteraan Fitur	Teknik	Fitur Terlibat
Pengekstrakan Fitur	Word2Vec	Jenis Kesalahan
Pemilihan Fitur	RFE	Semua Fitur
Pengurangan Dimensi Fitur	PCA	Semua Fitur

3.4.4.3 DRSA-P4-B3: Borang Analitis dan Operasi Pemodelan

Borang DRSA-P4-B3 bertujuan untuk menentukan tahap analitis, jenis operasi dan teknik pemodelan berdasarkan soalan bisnes daripada borang DRSA-P2-B4.

DRSA-P4-B3: Borang Analitis dan Operasi Pemodelan

Soalan Bisnes	Tahap Analitis	Jenis Operasi	Teknik Pemodelan
Apakah trend kekerapan kemalangan mengikut bulan?	*Deskriptif*	*Analisis Multidimensi*	*Histogram*
Apakah punca kekerapan kemalangan berlaku pada bulan-bulan tertentu?	*Diagnostik*	*Analisis Multidimensi*	*Time Series Analysis*
Apakah jangkaan jumlah jenis kesalahan pada masa hadapan mengikut lokasi berdasarkan peristiwa?	*Prediktif*	*Klasifikasi*	*Logistic Regression*
Bagaimana pengurusan trafik dapat dioptimumkan?	*Preskriptif*	*Pengoptimuman*	*Artificial Neural Network (ANN)*



3.4.4.4 DRSA-P4-B4: Borang Pemilihan Teknik Pemodelan

Borang DRSA-P4-B4 bertujuan untuk memilih teknik pemodelan yang terbaik bagi analitis. Pemilihan teknik pemodelan bagi peringkat ini mengambil kira faktor seperti kebolehskalaan, ketersediaan kepakaran, kos dan ketersediaan peralatan.

DRSA-P4-B4: Borang Pemilihan Teknik Pemodelan

Teknik Pemodelan Data	Batasan / Kekurangan yang diketahui	Pematuhan kepada Andaian Asas	Keboleh skalaan	Prestasi Terdahulu	Ketersediaan Tools/ Libraries	Ketersediaan Peralatan Visualisasi	Mudah Diguna Berdasarkan Pengalaman Pasukan	Ketersediaan Bantuan/ Latihan	Ketersediaan Syarikat / Perunding	Kos untuk Membangunkan atau Memperoleh	Keputusan (Y/T)
Logistic Regression	*Tidak berfungsi dengan baik jika hubungan antara input (Peristiwa) dan output (Jenis KesalahanTrafik) tidak linear.*	*Ya, terdapat perkaitan linear antara Peristiwa dengan Jenis Kesalahan*	*Bergantung kepada jumlah ciri dan data yang digunakan*	*Kajian literatur menunjukkan prestasi yang baik bagi data yang serupa*	*Terdapat pelbagai pakej terbuka seperti scikit-learn dalam bahasa Python yang menyediakan fungsi Logistic Regression*	*Terdapat pelbagai perisian visualisasi seperti Tableau, Power BI*	*Mudah digunakan*	*Terdapat pelbagai sumber bantuan yang boleh dirujuk seperti dalam talian dan forum komuniti*	*Terdapat syarikat yang mahir dalam pembangunan dan penggunaan Logistic Regression Dalam pelbagai aplikasi*	*Menggunakan kos yang rendah*	*Y*

3.4.4.5 DRSA-P4-B5: Borang Keperluan Set Data Latihan dan Pengujian

Borang DRSA-P4-B5 bertujuan untuk menyediakan set data latihan dan set data ujian yang diperlukan bagi melatih dan menguji model pembelajaran mesin.

DRSA-P4-B5: Borang Keperluan Set Data Latihan dan Pengujian

Keperluan Data Latihan dan Data Ujian	Sumber Data	Keperluan Pelabelan (penentuan)		Tindakan lain yang Diperlukan	Keperluan Sumber
		(Y/T)	Penerangan		
Data kesalahan trafik dan saman (80% data latihan, 20 % data ujian bergilir-gilir mengikut validasi silang k-fold)	*JPJ*	*T*	*Data telah sedia mempunyai bilangan kesalahan mengikut jenis kesalahan trafik. Maka, pemboleh ubah bersandar adalah bilangan kesalahan trafik dan selebihnya adalah pemboleh ubah tidak bersandar.*	1. *Medan Tarikh perlu dipecahkan kepada DD-MM-YYYY.* 2. *Nilai bagi atribut Peristiwa perlu ditukarkan kepada numerik.* 3. *Nilai bagi atribut Jenis Kesalahan perlu ditukarkan kepada numerik.*	*Tiada*

1. Data latihan dan data ujian untuk melatih dan menguji model. Perlu dinyatakan pembahagian data bagi data latihan dan data ujian
2. Tindakan yang diperlukan untuk melabelkan data bagi tujuan latihan dan ujian
3. Tindakan lain yang diperlukan untuk menyediakan data bagi keperluan dalam model. Contoh: Semua data perlu ditukarkan ke dalam bentuk nombor sebelum boleh digunakan dalam Rangkaian Neural
4. Keperluan sumber, manusia, bengkel, kewangan, dan API bagi pelabelan data, dan lain-lain.

3.4.4.6 DRSA-P4-B6: Borang Pembinaan dan Penalaan Model

Borang DRSA-P4-B6 bertujuan untuk menetapkan teknik, parameter dan proses yang akan digunakan untuk pembinaan model analitis. Penentuan parameter bagi pembinaan model adalah bergantung kepada jenis model yang dipilih.

Model dilatih menggunakan data latihan untuk mempelajari pola dan hubungan dalam data tersebut. Seterusnya, model yang terhasil daripada latihan tersebut diuji dengan menggunakan data ujian. Pengujian ini bertujuan untuk mengukur seberapa tepat model dapat melakukan ramalan daripada data ujian. Hasil pengujian memberikan pemahaman tentang kemampuan model dan memungkinkan penalaan lebih lanjut jika diperlukan untuk meningkatkan prestasi ramalan. Jika skor prestasi model tidak memuaskan, proses penalaan parameter model akan dibuat dan latihan akan diulang ataupun pemilihan model baharu akan dilakukan.

DRSA-P4-B6: Borang Pembinaan dan Penalaan Model

Tahap Analitis	Operasi	Teknik Pemodelan data	Langkah-langkah Penetapan	Penalaan Parameter Model
Prediktif	*Pengoptimuman*	*ANN*	*Langkah 1: Tentukan senibina model termasuk rekabentuk nod input dan output.* *Langkah 2: Tentukan lapisan padat dengan fungsi pengaktifan.* *Langkah 3: Susun model dengan fungsi pengoptimum dan kehilangan.* *Langkah 4: Model dilatih dengan set data latihan.*	1. *Jumlah lapisan.* 2. *Jumlah nod dalam setiap lapisan.* 3. *Fungsi pengaktifan yang sesuai untuk setiap nod.* 4. *Parameter bagi kadar pembelajaran dan kriteria pemberhentian atau jumlah epoch yang perlu sebelum pembelajaran ditamatkan.*

3.4.4.7 DRSA-P4-B7: Borang Penilaian Model

Borang DRSA-P4-B7 bertujuan untuk merekodkan hasil penilaian terhadap model yang dipilih berdasarkan kaedah penilaian yang digunakan. Penilaian model ialah proses menilai prestasi dan keberkesanan model pembelajaran mesin dalam menyelesaikan masalah. Ia melibatkan pengukuran ketepatan ramalan model, memahami kekuatan dan kelemahannya, dan menentukan sama ada ia memenuhi kriteria yang dikehendaki untuk tugas atau masalah

tertentu. Proses ini memerlukan penyediaan data latihan dan data ujian dengan kadar pembahagian tertentu.

Sekiranya prestasi model tidak memuaskan, beberapa penambahbaikan boleh dibuat sama ada senarai pemboleh ubah tidak bersandar disemak semula; model analitis lain perlu digunakan; data latihan perlu ditambah dan diperbaiki; atau senibina model dan penalaan parameternya perlu disemak semula dan ditambah baik.

DRSA-P4-B7: Borang Penilaian Prestasi Model

Teknik	Kaedah penilaian prestasi	Skor penilaian prestasi
Naïve Bayes *Decision Tree* *Random Forest* *ANN* *SVM*	*Average Accuracy*	*80%*
Linear Regression *Decision Tree*	*Pekali korelasi antara nilai benar dan ramalan*	*0.9*