MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN UNTUK MEMBUAT RAMALAN TAHAP PENULARAN WABAK DENGGI

NURUL NABIHAH BINTI MOHD AMINUDIN

SUHAILA BINTI ZAINUDIN

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

ABSTRAK

Menggunakan Teknik Pembelajaran Mesin Untuk Membuat Ramalan Tahap Penularan Wabak Denggi menggunakan pendekatan pembelajaran mesin bagi mengatasi kelemahan kaedah statistik tradisional yang kurang menyepadukan faktor persekitaran dan demografi secara holistik. Denggi merupakan salah satu penyakit bawaan nyamuk yang paling cepat merebak di dunia, dan Malaysia mencatatkan peningkatan kes yang signifikan saban tahun, misalnya pada tahun 2020 Pertubuhan Kesihatan Sedunia (WHO) menyatakan Malaysia merupakan salah satu negara yang mempunyai kadar demam denggi tertinggi di Rantau Pasifik Barat, dengan anggaran sebanyak 330,000 kes. Justeru, terdapat keperluan mendesak untuk membangunkan sistem ramalan yang lebih tepat dan berupaya memberikan kesedaran awal. Projek ini dijalankan bagi mencapai tiga objektif utama: membangunkan sistem bersepadu yang boleh memberi kesedaran awal kepada pengguna, membina model ramalan yang tepat dan responsif, serta memastikan model tersebut dapat menyesuaikan diri dengan data yang dinamik mengikut masa, lokasi dan perubahan persekitaran. Metodologi CRISP-DM digunakan untuk memandu keseluruhan proses pembangunan, manakala pelbagai model pembelajaran mesin seperti Hutan Rawak, Pokok Keputusan, Regresi Logistik, SVM, dan KNN telah dibandingkan. Model K-Nearest Neighbors (KNN) yang melalui proses Kejuruteraan Fitur, SMOTE dan Penalaan Hiperparameter menunjukkan prestasi terbaik dengan ketepatan 70.27% dan AUC 0.7042. Model ini kemudiannya diintegrasikan ke dalam sistem ramalan interaktif menggunakan platform Streamlit, yang membolehkan pengguna memasukkan data seperti suhu, kelembapan, minggu epidemik dan lokasi, seterusnya menerima output ramalan secara serta-merta dalam bentuk label risiko seperti "Zone 1", "Zone 2" atau "Zone 3" berserta kebarangkalian visual. Sistem ini berupaya membantu pegawai kesihatan awam dan orang awam dalam mengenal pasti kawasan berisiko serta meningkatkan kesedaran dan langkah pencegahan awal terhadap penularan wabak denggi.

Kata Kunci: Pembelajaran Mesin, Wabak Denggi, KNN, SMOTE, CRISP-DM

PENGENALAN

Demam denggi merupakan penyakit yang terjadi apabila seseorang dijangkiti virus yang merebak melalui gigitan nyamuk aedes. Terdapat empat jenis virus yang mampu mengakibatkan demam denggi iaitu virus DEN 1, DEN 2, DEN 3, dan DEN 4 (Hassan et al. 2012). Di seluruh dunia, terdapat lebih 3.9 bilion individu berisiko dijangkiti demam denggi. Penyakit ini juga bukanlah asing dan merupakan endemik di 128 buah negara di seluruh Asia Selatan, Asia Tenggara, Afrika, Amerika, Barat Pasifik, dan Kawasan Mediterranean Timur (Salim et al. 2021). Peningkatan kes denggi di Seremban mencetuskan kebimbangan kepada kesihatan awam di Malaysia. Pada tahun 2019, Kementerian Kesihatan Malaysia (KKM) melaporkan lebih 130,000 kes denggi di seluruh negara, terutamanya melibatkan bandar besar seperti Seremban. Di Malaysia, kes denggi mula merebak pada tahun 1902 dan mula menjadi salah satu risiko kepada kesihatan awam sekitar tahun 1970-an dimana wabak denggi yang pertama telah merebak pada tahun 1973. Pada 31 Disember 2022, sejumlah 4,110,465 kes denggi manakala 4099 kematian telah dicatatkan di peringkat global (Muhammad Bilal Khan et al., 2023). Menurut Pertubuhan Kesihatan Sedunia (WHO), Malaysia merupakan salah satu negara yang mempunyai kadar demam denggi tertinggi di Rantau Pasifik Barat, dengan anggaran sebanyak 330,000 kes dilaporkan pada tahun 2020 (Majeed et al. 2023). Pihak kerajaan telah melaksanakan beberapa langkah bagi mengawal dan mencegah wabak demam denggi termasuklah kawalan vektor (contoh: membuang tempat pembiakan nyamuk), kempen pendidikan awam dan vaksinasi. Namun, di sebalik usaha-usaha ini, demam denggi tetap menjadi kebimbangan kesihatan awam utama di Malaysia.

Secara amnya, projek "Menggunakan Teknik Pembelajaran Mesin Untuk Membuat Ramalan Tahap Penularan Wabak Denggi" memanfaatkan pembelajaran mesin bagi tujuan meramal wabak denggi demi mengurangkan risiko orang awam dijangkiti virus demam denggi. Kaedah tradisional dalam meramal wabak demam denggi sering bergantung pada model statistik, yang berkemungkinan tidak mampu menangkap interaksi kompleks antara faktor persekitaran, demografi, dan iklim. Oleh itu, terdapat keperluan yang semakin meningkat untuk pendekatan berasaskan data *(data-driven)* yang lebih maju bagi meramal wabak denggi dengan lebih tepat dan cekap. Projek ini akan menganalisis data kes sejarah, corak cuaca, dan faktor persekitaran. Ia akan mengumpul dan memproses data untuk mengenal pasti corak, membolehkan ramalan tepat bagi kawasan berisiko tinggi di Seremban. Sistem ini akan menyediakan kemas kini masa nyata melalui papan pemuka interaktif, membantu pegawai kesihatan awam menvisualisasikan risiko wabak. Sistem ini bertujuan untuk meningkatkan kesiapsiagaan wabak denggi dan meminimumkan risiko kesihatan awam.

KAJIAN LITERATUR

Isu penularan demam denggi telah menarik perhatian ramai penyelidik dalam pelbagai bidang, termasuk kesihatan awam, sains data dan kecerdasan buatan. Kajian literatur menunjukkan bahawa pendekatan awal dalam meramal kes denggi banyak bergantung kepada kaedah statistik tradisional seperti regresi linear, regresi logistik, dan model autoregresif (ARIMA). Kaedah ini biasanya digunakan untuk menganalisis hubungan antara pembolehubah cuaca seperti suhu, kelembapan, dan hujan dengan jumlah kes yang dicatatkan. Walau bagaimanapun, pendekatan statistik ini menghadapi keterbatasan dalam mengenal pasti corak bukan linear dan interaksi kompleks antara faktor persekitaran dan epidemiologi yang berubah-ubah mengikut masa dan lokasi.

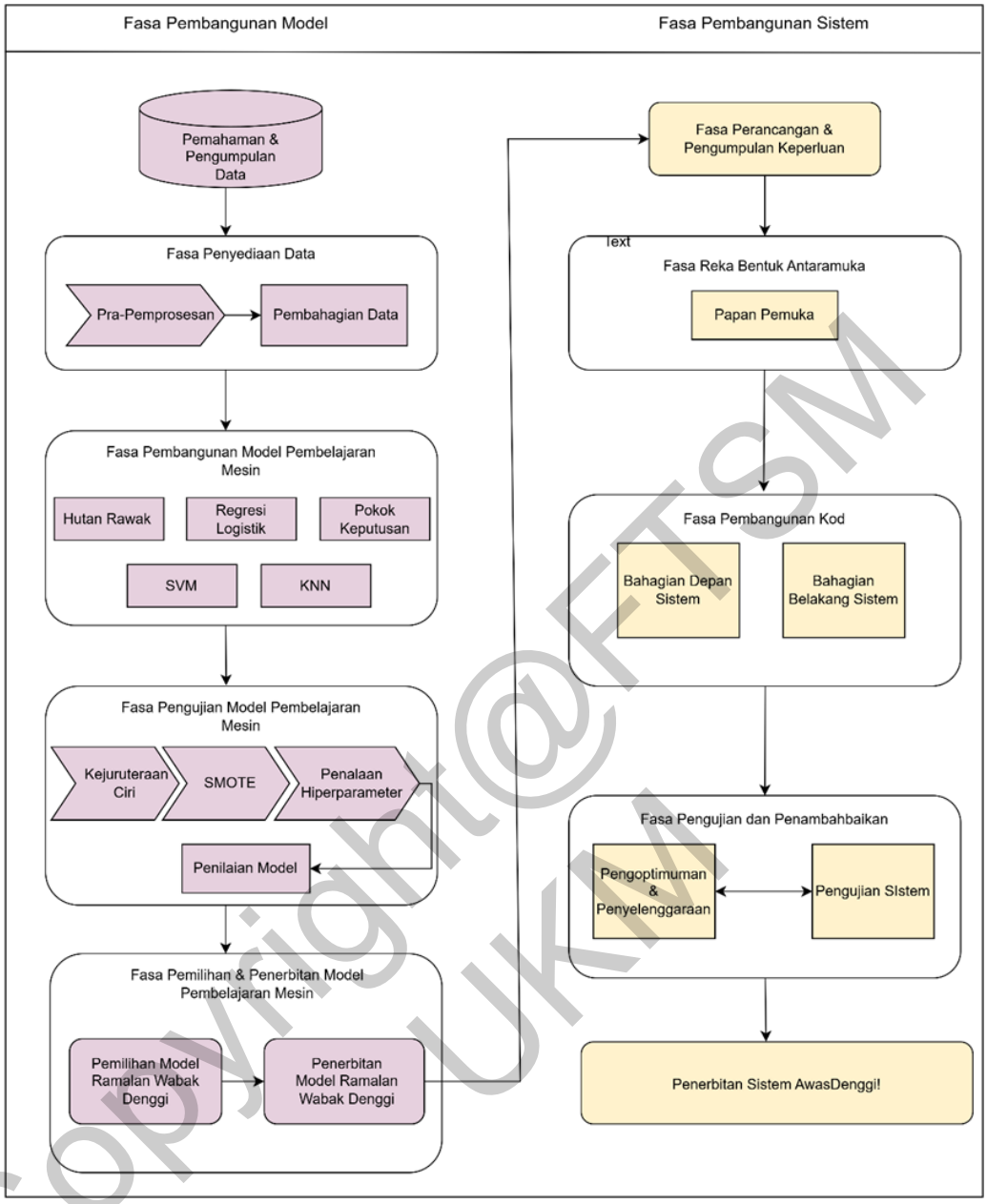
Bagi mengatasi kekangan ini, kaedah berasaskan pembelajaran mesin mula digunakan secara lebih meluas. Model seperti Pokok Keputusan, Hutan Rawak, *Support Vector Machine* (SVM), dan *K-Nearest Neighbors* (KNN) telah digunakan dalam pelbagai kajian terdahulu kerana keupayaannya dalam mengendalikan data berskala besar dan pelbagai jenis ciri. Sebagai contoh, kajian oleh Nazri et al. (2019) membandingkan prestasi beberapa model pembelajaran mesin dalam meramal kes denggi dan mendapati bahawa model Hutan Rawak menunjukkan prestasi yang konsisten dan stabil. Manakala, kajian oleh Ros et al. (2021) membuktikan bahawa SVM mampu memberikan ketepatan tinggi dalam senario yang melibatkan data kompleks dan tidak linear.

Tambahan pula, kajian literatur turut menekankan kepentingan proses Kejuruteraan Fitur *(Feature Engineering)* dan penyesuaian terhadap masalah ketidakseimbangan data. Kajian oleh Anwar et al. (2020) menunjukkan bahawa penggunaan teknik SMOTE (Synthetic Minority Over-sampling Technique) dapat meningkatkan ketepatan model dengan menghasilkan sampel baharu bagi kelas minoriti. Proses ini membantu mengurangkan berat sebelah model terhadap kelas majoriti, terutamanya dalam kes seperti penularan wabak, di mana kawasan risiko tinggi selalunya lebih sedikit berbanding kawasan risiko rendah. Kajian tersebut turut menekankan bahawa penalaan hiperparameter *(hyperparameter tuning)* adalah aspek penting dalam mempertingkatkan prestasi model pembelajaran mesin.

Akhir sekali, beberapa kajian telah mula menggabungkan model ramalan ke dalam sistem amaran awal berasaskan web atau mudah alih untuk kegunaan praktikal. Pendekatan ini tidak hanya memberi fokus kepada ketepatan ramalan, tetapi juga kepada kemudahan akses oleh pengguna dan pegawai kesihatan. Ini termasuk sistem visualisasi ramalan risiko wabak dalam bentuk zon, peta haba, atau amaran tahap risiko. Berdasarkan kajian-kajian ini, jelas bahawa pembelajaran mesin bukan sahaja sesuai digunakan untuk analisis ramalan denggi, tetapi juga mempunyai potensi tinggi untuk diintegrasikan ke dalam sistem bersepadu yang responsif dan mampu memberikan maklumat kesihatan awam secara proaktif kepada masyarakat.

METODOLOGI KAJIAN

Kajian ini merangkumi analisis keperluan, merangka reka bentuk model konseptual, pembangunan model, pengujian kebolehgunaan dan hasil. Metodologi menerangkan kaedah bagi mengatasi masalah yang dikenal pasti serta menerangkan proses kajian yang dilakukan. Metodologi yang digunakan dalam kajian ini ialah CRISP-DM *(Cross Industry Standard Process for Data Mining)*, iaitu satu pendekatan berstruktur yang terdiri daripada enam fasa utama: Pemahaman Perniagaan, Pemahaman Data, Penyediaan Data, Pemodelan, Penilaian, dan Penerapan. Pendekatan ini dipilih kerana kesesuaiannya dalam pembangunan sistem berasaskan data yang melibatkan proses analisis dan pembelajaran mesin secara sistematik. Carta alir bagi fasa-dasa pembangunan model dan sistem AwasDenggi! yang menggunakan CRISP-DM sebagai langkah untuk membangunkan sistem ini boleh dilihat dalam Rajah 1.



Rajah 1

Rajah 1 menunjukkan carta alir bagi pem pembangunan model ramalan wabak denggi dan juga sistem AwasDenggi! yang akan diintegrasikan bersama model ML yang telah dilatih. Fasa Pembangunan Model akan mengaplikasikan bahasa pengaturcaraan Python dan Visual Studio Code sebagai platform untuk membuat pengaturcaraan.

Fasa Pemahaman dan Pengumpulan Data

Langkah pertama dalam membangunkan model pembelajaran mesin adalah memahami dan mengumpul data untuk tujuan meramal wabak denggi menggunakan data sedia ada. Data yang digunakan untuk melatih model ML adalah data kes denggi di Seremban sekitar tahun 2003 – 2009 yang merangkumi data cuaca seperti suhu purata, kelembapan, taburan hujan,

maklumat masa seperti minggu atau bulan kejadian dan juga demografi pesakit. Data ini diperoleh daripada kajian terdahulu yang pernah dijalankan dan merupakan data berstruktur yang disimpan dalam format 'CSV'. Pemahaman terhadap data memainkan peranan penting bagi mengenal pasti jenis atribut yang akan digunakan, sama ada bersifat numerik, kategori atau masa. Selain itu, analisis penerokaan data (exploratory data analysis) turut dijalankan bagi mengenal pasti taburan data, nilai yang tidak lengkap serta pola tertentu dalam kes denggi mengikut masa atau kawasan. Pemahaman ini membolehkan ciri-ciri yang berpotensi menyumbang kepada berlakunya wabak denggi dikenal pasti, seterusnya memandu keputusan dalam pemilihan ciri untuk pembangunan model yang lebih berkesan. Rajah 2 di bawah menunjukkan sebahagian daripada keseluruhan data mentah yang digunakan dalam projek ini sebelum dipraproses untuk proses pemodelan. Data ini mempunyai 21 atribut secara keseluruhan.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	YEAR	WEEK	ACC_WEE	DF	DHF	CASE	TEMP_MA	TEMP_MIN	TEMP_AVE	HUMID	RAINFALL	BULAN	KES	UMUR	JANTINA	BANGSA	PEKERJAA	ALAMAT1I	MAJDAE	DAERAH	WABAK
2	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	L	INDIA	Pelajar	AMPANGA	MPS	SN	TKW
3	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	P	MELAYU	Suri Ruma	RASAH	MPS	SN	TKW
4	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Kanak-kan	L	MELAYU	Kanak-kan	AMPANGA	MPS	SN	DKW
5	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	L	MELAYU	Buruh Am	AMPANGA	MPS	SN	TKW
6	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	L	INDIA	Pelajar	RASAH	MPS	SN	MWB
7	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	L	INDIA	Buruh Am	RANTAU	MPS	SN	DKW
8	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Kanak-kan	P	MELAYU	Kanak-kan	AMPANGA	MPS	SN	TKW
9	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	L	MELAYU	Bukan Eks	RASAH	MPS	SN	DKW
10	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	P	CINA	Bukan Eks	RASAH	MPS	SN	DKW
11	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	L	MELAYU	Pelajar	AMPANGA	MPS	SN	DKW
12	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	L	INDIA	Buruh Am	SETUL	MPS	SN	TKW
13	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	L	MELAYU	Pelajar	AMPANGA	MPS	SN	DKW
14	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	L	MELAYU	Buruh Am	AMPANGA	MPS	SN	DKW
15	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	P	CINA	Sendiri	AMPANGA	MPS	SN	TKW
16	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	P	MELAYU	Suri Ruma	AMPANGA	MPS	SN	TKW
17	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	P	MELAYU	Eksekutif	AMPANGA	MPS	SN	TKW
18	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	L	CINA	Buruh Am	RASAH	MPS	SN	TKW
19	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	P	MELAYU	Pelajar	AMPANGA	MPS	SN	TKW
20	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Dewasa	P	MELAYU	Suri Ruma	AMPANGA	MPS	SN	TKW
21	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Belia	P	CINA	Pelajar	RASAH	MPS	SN	TKW
22	2003	1	1	100	9	109	31.5125	22.45	25.724	77.2521	50.9286	JAN	0	Warga Em	P	MELAYU	Warga Em	AMPANGA	MPS	SN	DKW

Rajah 2

Fasa Penyediaan Data

Bagi fasa penyediaan data, ianya merangkumi beberapa langkah penting yang perlu dilaksanakan sebelum proses latihan model dilakukan. Data yang digunakan terdiri daripada pelbagai ciri yang berkaitan dengan kes denggi di daerah Seremban, termasuk data cuaca, data masa dan lokasi kejadian. Data ini memerlukan pemprosesan awal bagi memastikan data adalah berish, konsisten dan sesuai digunakan dalam model pembelajaran mesin.

Langkah pertama dalam penyediaan data ialah menangani nilai hilang, khususnya dalam ciri suhu seperti 'TEMP_MAX', 'TEMP_MIN' dan 'TEMP_AVE'. Nilai yang hilang dalam ciri-ciri ini digantikan dengan purata keseluruhan bagi setiap ciri supaya tidak berlaku kehilangan maklumat yang penting semasa latihan model. Data kategori pula hendahlah ditukar dalam bentuk numerik supaya boleh dibaca oleh model ML. Penukaran ini dilakukan menggunakan pengekodan label (Label Encoder) terutamanya pada atribut seperi 'ALAMAT1D' dan 'WABAK' iaitu kelas sasaran bagi projek ini. Bagi memastikan konsistensi, semua nilai dalam 'ALAMAT1D' ditukar kepada huruf besar sebelum pengekodan dijalankan. Nilai unik dalam kelas sasaran adalah seperti berikut:

- 0: DKW – Berhampiran Kawasan Wabak
- 1: MWB – Mungkin Wabak Berlaku
- 2: TKW – Tiada Wabak Dikesan

Seterusnya, proses pemetaan turut dijalankan ke atas atribut 'BULAN' bagi memastikan kod bagi bulan adalah selari dengan nama bulan tersebut (e.g, JAN → 1). Data juga disusun dan ditapis untuk memastikan hanya ciri-ciri yang relevan digunakan. Melalui kaedah pemilihan seperti SelectKBest dengan ujian ANOVA (Analysis of Variance) F-Test, ciri-ciri yang paling berkait rapat dengan perubahan tahap wabak dikenal pasti dan dipilih untuk melatih model. Skor ini menggambarkan bahawa fitur-fitur ini berkemungkinan besar mempunyai hubungan yang kuat dengan tahap wabak denggi dan berpotensi dalam meningkatkan prestasi model sekiranya digunakan dalam latihan. Jadual 1 memaparkan skor bagi kesemua fitur yang terdapat dalam set data, tidak termasuk kelas sasaran.

Jadual 1

Fitur	Skor
CASE	544.1237
DF	540.2407
ACC_WEEK	436.7668
YEAR	430.7467
DHF	280.3624
WEEK	86.0553
BULAN	84.6834
TEMP_MIN	36.0049
TEMP_AVE	34.1891
TEMP_MAX	18.2495
ALAMAT1D	16.3588
BANGSA	13.2520
JANTINA	12.2176
HUMID	9.8652
MAJDAE	3.0786
DAERAH	1.5231
PEKERJAAN	1.3947
RAINFALL	0.9734
UMUR	0.6443
KES	NaN

Berdasarkan skor kepentingan fitur yang diperoleh berserta rujukan daripada beberapa kajian lepas, hanya 9 ciri atau fitur yang dipilih untuk melatih model ML. Data klinikal pesakit tidak digunakan untuk melatih model walaupun skor daripada ANOVA F-Test adalah tinggi kerana projek ini bukanlah bertujuan untuk membuat diagnosis individu, akan tetapi projek ini memerikan fokus kepada ramalan tahap wabak berdasarkan faktor persekitaran dan masa. Hal ini menjadikan data persekitaran adalah lebih relevan bagi tujuan pemantauan dan pencegahan pada peringkat kawasan, dalam masa yang sama meningkatkan prestasi ramalan dan ketepatan ramalan model. Oleh itu, fitur yang kurang relevan akan disisihkan daripada set data menggunakan drop column bagi memastikan hanya ciri yang paling signifikan digunakan dalam latihan model sekaligus meningkatkan kecekapan dan prestasi model. Rajah 3 menunjukkan 5 data terawal dengan 10 atribut dan 1 kelas sasaran sebelum pengekodan label dijalankan, menjadikan jumlah kolum sebanyak 11 kesemuanya.

YEAR	ACC_WEEK	CASE	TEMP_MAX	TEMP_MIN	TEMP_AVE	HUMID	BULAN	ALAMAT1D	WABAK	WABAK_ENCODED
2003	1	109	31.5125	22.45	25.723958	77.252083	1	0	TKW	2
2003	1	109	31.5125	22.45	25.723958	77.252083	1	8	TKW	2
2003	1	109	31.5125	22.45	25.723958	77.252083	1	0	DKW	0
2003	1	109	31.5125	22.45	25.723958	77.252083	1	0	TKW	2
2003	1	109	31.5125	22.45	25.723958	77.252083	1	8	MWB	1

Rajah 3

Data yang telah melalui proses prapemprosesan ini dibahagikan kepada dua set iaitu set latihan dan set ujian menggunakan teknik *Stratified Train-Test Split* bagi memastikan agihan kelas (label sasaran) yang seimbang dalam setiap set. Penggunaan teknik *stratified* adalah penting kerana set data ini melibatkan masalah klasifikasi berbilang kelas yang tidak seimbang. Dalam projek ini, kelas 2 (TKW - Tiada Wabak Dikesan) menunjukkan bilangan yang jauh lebih tinggi berbanding kelas 0 (DKW – Berhampiran Kawasan Wabak) dan kelas 1 (DKW – Mungkin Wabak Berlaku). Hal ini menyebabkan model ML cenderung untuk lebih mengenal pasti dan memberi keutamaan kepada kelas majoriti dan mengabaikan kelas minoriti yang akan menyebabkan prestasi model dalam mengesan kelas minoriti merosot. Teknik *Stratified Train-Test Split* memastikan kedua-dua subset mengekalkan perkadaran kelas yang sama seperti dalam keseluruhan set data dan membolehkan model mempelajari model darapada semua kelas dengan seimbang dan adil. Agihan ini mengggunakan nisbah 80:20 di mana 80% data dimasukkan ke dalam set latihan yang digunakan oleh model ML untuk mempelajari corak, hubugan antara fitur dan membuat generalisasi terhadap data. Manakala baki 20% daripada keseluruhan data pula akan dimasukkan ke dalam set ujian yang tidak pernah dilihat oleh model semasa latihan. Set ini digunakan untuk menilai sejauh mana model mampu membuat rmalan dengan tepat terhdapa data baru yang tidak dikenali, sekaligus mengukur prestasi sebenar model dari segi ketepatan, keupayaan klasifikasi dan kebolehan generalisasi. Rajah 4 menunjukkan bilangan data bagi setiap kelas selepas dibahagikan kepada set latihan dan set ujian.

```
✅ X_train shape: (4864, 9)
✅ y_train distribution:
 WABAK_ENCODED
2    3023
0    1540
1     301
Name: count, dtype: int64
✅ X_test shape: (1217, 9)
✅ y_test distribution:
 WABAK_ENCODED
2    757
0    385
1     75
```

Rajah 4

Fasa Pembangunan Model

Fasa pembangunan model ML dalam meramal wabak denggi melibatkan proses latihan dan pengujian ke atas lima model ML terbaik yang telah dikenal pasti berdasarkan prestasi dalam kajian-kajian lepas berserta kesesuaian terhadap jenis data tabular. Setiap model ini dilatih menggunakan data yang telah dipraproses dan dipilih berdasarkan skor kepentingan fitur. Lima model ML tersebut adalah Hutan Rawak, Regresi Logisitik, Pokok Keputusan, SVM (Support Vector Machine) dan KNN (K-Nearest Neighbor). Bagi memastikan penilaian yang adil dan menyeluruh, beberapa metrik penialaian prestasi telah digunakan iaitu ketepatan bagi

kedua-dua set latihan dan ujian, keluasan di bawah lengkungan (AUC) yang menilai keupayaan model membezakan antara kelas, serta skor F1 makro (macro F1-score) yang mengambil kira prestasi bagi setiap kelas secara seimbang.

a. Hutan Rawak

Algoritma Hutan Rawak membina banyak pokok keputusan di mana setiap pokok dilatih dengan subset data dan ciri yang berbeza. Hasil daripada setiap pokok digabungkan melalui undian majoriti untuk menghasilkan keputusan akhir. Parameter class-weight='balanced' digunakan untuk menghasilkan ketidakseimbangan kelas pada data 'WABAK' supaya model fokus kepada semua kelas dengan seimbang. Rajah 5 merupakan pseudokod bagi algoritma Hutan Rawak sebagai rujukan.

```
ALGORITMA Hutan_Rawak

1. Inisialisasi model RF:
   rf_model ← RandomForestClassifier(dengan random_state = 42,
      class_weight = 'balanced')

2. Lakukan pensilangan silang 10-lipatan:
   cv_scores ← cross_val_score(rf_model, X_train, y_train, cv = 10,
      scoring = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set):", min_cv, "±", std_cv

3. Latih model dengan set latihan penuh:
   latih(rf_model, X_train, y_train)

4. Buat ramalan ke atas set ujian:
   y_pred ← ramal(rf_model, X_test)
   y_proba ← ramal_kebarangkalian(rf_model, X_test)

5. Kira ketepatan ujian:
   test_accuracy ← skor(rf_model, X_test, y_test)
   Papar "Test Accuracy:", test_accuracy

TAMAT ALGORITMA
```

Rajah 5

b. Regresi Logistik

Regresi Logistik digunakan sebagai model klasifikasi linear yang seusuai untuk menangani masalah pelabelan berbilang kelas seperti tahap wabak denggi. Model ini menggunakan pendekatan penskalaan ciri melalui *StandardScaler* dan dilatih dengan kaedah *Pipeline* bagi memastikan aliran praproses yang konsisten. Rajah 6 merupakan pseudokod untuk algoritma Regresi Logistik.

```
ALGORITMA Regresi_Logistik

2. Bina saluran pemprosesan (pipeline):
   logreg_pipeline ← Pipeline(
       'scaler' = StandardScaler(),
       'logreg' = LogisticRegression(max_iter = 1000, class_weight =
      'balanced', random_state = 42)
   )

3. Jalankan pensilangan silang 5-lipatan:
   cv_scores ← cross_val_score(logreg_pipeline, X_train, y_train, cv =
      5, skor = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set):", min_cv, "±", std_cv

4. Latih model menggunakan set latihan penuh:
   latih(logreg_pipeline, X_train, y_train)

5. Buat ramalan ke atas set ujian:
   y_pred ← ramal(logreg_pipeline, X_test)
   y_proba ← ramal_kebarangkalian(logreg_pipeline, X_test)

6. Kira ketepatan model:
   test_accuracy ← kira_ketepatan(y_test, y_pred)
   Papar "Test Accuracy:", test_accuracy

TAMAT ALGORITMA
```

Rajah 6

c. Pokok Keputusan

Algoritma Pokok Keputusan berfungsi dengan cara membina struktutr seperti pokok yang membuat keputusan berdasarkan syarat tertentu pada setiap nod. Kelebihannya ialah ia mudah difahami dan tidak memerlukan penskalaan ciri. Dalam projek ini, model ini dilatih secara langsung ke atas data tanpa *Pipeline*. Rajah 7 menunjukkan pseudokod bagi algoritma Pokok Keputusan.

```
ALGORITMA Pokok_Keputusan

2. Inisialisasi model Pokok Keputusan:
   dt_model ← DecisionTreeClassifier(random_state = 42, class_weight =
      'balanced')

3. Jalankan pensilangan silang 5-lipatan ke atas set latihan:
   cv_scores ← cross_val_score(dt_model, X_train, y_train, cv = 5, skor
    = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set):", min_cv, "±", std_cv

4. Latih model menggunakan set latihan penuh:
   latih(dt_model, X_train, y_train)

5. Buat ramalan ke atas set ujian:
   y_pred ← ramal(dt_model, X_test)
   y_proba ← ramal_kebarangkalian(dt_model, X_test)

TAMAT ALGORITMA
```

Rajah 7

d. SVM *(Support Vector Machine)*

Model SVM menggunakan pendekatan mencari sempadan terbaik *(hyperplane)* untuk memisahkan kelas secara optimum dalam ruang ciri. Dalam projek ini, model SVM menggunakan kernel rbf yang sesuai untuk data bukaan linear, serta diproses melalui *Pipeline* yang mengandung penskalaan sebelum latihan. Sebahagian daripada keseluruhan pseudokod bagi algoritma ini boleh dilihat dalam Rajah 8 di bawah.

```
ALGORITMA SVM

2. Bina saluran pemprosesan (pipeline) yang mengandungi:
   - Penyesaran ciri (StandardScaler)
   - Model SVM dengan kernel = 'rbf', class_weight = 'balanced',
     probability = True
   svm_pipeline ← Pipeline(scaler + SVM)

3. Jalankan pensilangan silang 5-lipatan:
   cv_scores ← cross_val_score(svm_pipeline, X_train, y_train, cv = 5,
     skor = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set):", min_cv, "±", std_cv

4. Latih model menggunakan set latihan penuh:
   latih(svm_pipeline, X_train, y_train)

TAMAT ALGORITMA
```

Rajah 8

e. KNN *(K-Nearest Neighbor)*

Model KNN menjalankan fungsi dengan mencari sejumlah 'k' jiran terdekat daripada data latihan untuk menentukan kelas bagi data baharu berdasekan majoriti kelas tersebut. Nilai k=5 dan penskalaan ciri dilakukan terlebih dahulu untuk memastikan semua fitur berada pada skala yang sama. Model ini tidak mempunyai parameter pembelajaran secara eksplisit, akan tetapi ia membuat keputusan berdasarkan jarak dalam ruang ciri semasa proses ramalan. Pseudokod Dalam Rajah 9 merupakan sebahagian kod bagi model KNN.

```
ALGORITMA KNN

1. Bina saluran pemprosesan (pipeline) yang mengandungi:
   - Penyesaran ciri (StandardScaler)
   - Model KNN dengan n_neighbors = 5
   knn_pipeline ← Pipeline(scaler + KNN)

2. Jalankan pensilangan silang 5-lipatan:
   cv_scores ← cross_val_score(knn_pipeline, X_train, y_train, cv = 5,
     skor = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set):", min_cv, "±", std_cv

3. Latih model menggunakan set latihan penuh:
   latih(knn_pipeline, X_train, y_train)

TAMAT ALGORITMA
```

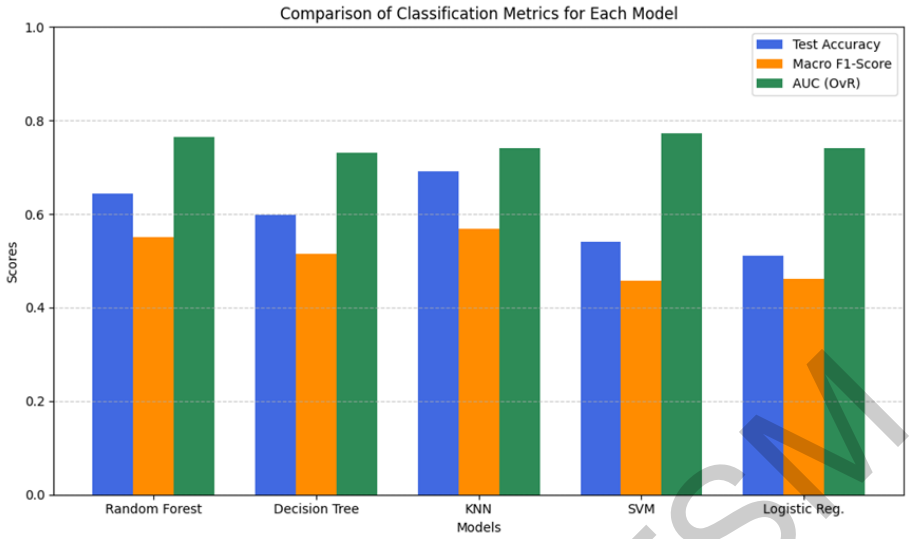
Rajah 9

Fasa Pengujian Model Pembelajaran Mesin

Berdasarkan kesemua lima model yang telah dilatih, keputusan yang diperolehi telah dicatat dalam Jadual 2 supaya perbandingan keputusan antara kesemua model dapat dilihat dengan lebih jelas. Rajah 10 pula menunjukkan gambaran carta bar bagi keputusan bagi ketepatan ujian, skor F1 makro dan keluasan bawah lengkungan yang diperoleh ke atas kesemua lima model.

Jadual 2

Model	Ketetpatan Ujian	AUC	Ketetpatan Makro	Ingatan Makro	Skor F1 Makro
Hutan Rawak	0.6434	0.7652	0.5350	0.5970	0.5498
Regresi Logistik	0.5103	0.7409	0.4766	0.5871	0.4612
Pokok Keputusan	0.5990	0.7310	0.5089	0.5818	0.5140
SVM	0.5415	0.7731	0.4946	0.6047	0.4851
KNN	0.6910	0.7405	0.5985	0.5505	0.5688



Rajah 10

Berdasarkan Jadual 2 yang menunjukkan prestasi lima model ML yang diuji, model KNN telah menunjukkan prestasi terbaik secara keseluruhan dalam ujian ramalan wabak denggi. KNN mencatatkan ketepatan tertinggi iaitu sebanyak 69.10% serta skor F1 Makro sebanyak 0.5688 dimana KNN menunjukkan keseimbangan prestasi terhadap ketiga-tiga kelas sasaran. Walaupun Hutan Rawak mencatatkan nilai AUC tertinggi iaitu sebanyak 0.7652, ketepatan ujiannya masih rendah jika dibandingkan dengan KNN iaitu 64.34%. Perkara ini menunjukkan bahawa walaupun Hutan Rawak berkebolehan membezakan kelas dengan baik, ramalan yang dibuat tidak setepat KNN dalam set data ujian.

Sebaliknya, model Regresi Logistik dan SVM mencatatkan prestasi yang lebih rendah dari segi ketepatan dengan nilai 51.03% dan 54.15% masing-masing walaupun AUC mereka adalah kompetitif iaitu sekitar 0.7409 dan 0.7731. Hal ini menunjukkan bahawa walaupun model-model tersebut agak baik dalam mengklasifikasikan kelas berdasarkan kebarangkalian, prestasi keseluruhan dalam membuat keputusan muktamad adalah kurang memuaskan. Model Pokok Keputusan pula berada di tahap sederhana dengan ketepatan 59.90% namun dari segi skor F1 makro dan keseimbangan ketiga-tiga metrik, KNN masih mendahului. Secara keseluruhan, KNN muncul sebagai model paling seimbang dan tepat, menjadikannya pilihan utama untuk dibangunkan ke peringkat sistem ramalan interaktif dalam projek ini. Akan tetapi, ketepatan KNN iaitu 69.10% masih boleh ditambah baik. Berdasarkan keputusan yang diperoleh ini, tiga model terbaik dipilih untuk ditambah baik prestasinya iaitu Hutan Rawak, Pokok Keputusan dan KNN. Beberapa langkah telah diambil untuk menambah baik prestasi melalui teknik Kejuruteraan Fitur *(Feature Engneering)*, SMOTE *(Synthetic Minority Oversampling Technique)* dan penelaan hiperparameter.

i. Kejuruteraan Fitur *(Feature Engineering - FE)*

Teknik kejuruteraan fitur telah digunakan dalam projek ini untuk menambah maklumat berkualiti kepada model dengan mencipta fitur baharu daripada fitur sedia ada. Dalam kod tersebut, tiga ciri baharu telah diwujudkan iaitu:

- TEMP_RANGE: Beza antara suhu maksimum dan minimum
- HUMID_TEMP_INT: Interaksi antara kelembapan dan suhu purata
- CASE_PER_WEEK: Kadar kes mingguan

Ciri-ciri ini direka berdasarkan pemahaman domain bahawa perubahan suhu, gabungan suhu dan kelembapan, serta beban kes mingguan mungkin memberi kesan terhadap kebarangkalian berlakunya wabak denggi. Rajah 11 menunjukkan bagaimana kejuruteraan fitur diaplikasikan dalam projek ini.

```
# Create new engineered features
data['TEMP_RANGE'] = data['TEMP_MAX'] - data['TEMP_MIN']
data['HUMID_TEMP_INT'] = data['HUMID'] * data['TEMP_AVE']
data['CASE_PER_WEEK'] = data['CASE'] / (data['ACC_WEEK'] + 1)  # +1 to avoid divide-by-zero
```

Rajah 11

Kejuruteraan fitur diaplikasikan terlebih dahulu sebelum SMOTE kerana SMOTE mensintesis data baharu berdasarkan nilai-nilai asal setiap ciri. Sekiranya SMOTE dijalankan terlebih dahulu sebelum kejuruteraan fitur, nilai-nilai baharu yang dijana mungkin tidak menggambarkan hubungan sebenar antara ciri-ciri yang direka dengan kelas sasaran, malah boleh boleh menghasilkan data sintetik yang tidak logik atau tidak konsisten. Jadual 3 menunjukkan keputusan apabila kejuruteraan fitur dijalankan ke atas tiga model terbaik dengan ketepatan yang tinggi iaitu Hutan Rawak, Pokok Keputusan dan KNN.

Jadual 3

Model	Ketepatan Ujian	AUC	Ketetpatan Makro	Ingatan Makro	Skor F1 Makro
Hutan Rawak + FE	0.6417	0.7574	0.5338	0.5957	0.5489
Pokok Keputusan + FE	0.5974	0.7299	0.5084	0.5814	0.5127
KNN + FE	0.6927	0.7473	0.5992	0.5649	0.5789

Berdasarkan keputusan yang ditunjukkan dalah Jadual 3, dapat dilihat bahawa penerapan kejuruteraan fitur memberi kesan yang berbeza-beza ke atas prestasi model. KNN mencatat peningkatan dalam ketepatan ujian daripada 69.10% kepada 69.27% dan skor F1 makro daripada 0.5688 kepada 0.5789, menjadikannya model paling konsisten dan stabil. Walaupun peningkatan bagi Hutan Rawak dan Pokok Keputusan agak kecil atau hampir kekal, namun perubahan nilai seperti AUC dan ketepatan makro menunjukkan bahawa FE membantu model dalam mengenal pasti pola yang lebih bermakna.

ii. SMOTE *(Synthetic Minority Oversampling Technique)*

Memandangkan kelas sasaran (‘WABAK’) adalah tidak seimbang antara 3 kategori (TKW, DKW dan MWB), SMOTE digunakan untuk menanganinya. Kelas majoriti seperti MWB mempunyai bilangan sample yang jauh lebih banyak berbanding kelas minoriti seperti DKW dan TKW. Ketidakseimbangan ini akan menyebabkan model cenderung mengabaikan corak daripada kelas minoriti. Dengan penggunaan SMOTE, sampel sintetik bagi kelas minoriti dijana untuk menyeimbangkan agihan kelas dalam set latihan, sekaligus membantu model membina pemahaman yang lebih adil terhadap ketiga-tiga kategori. Keupayaan model dapat ditingkatkan dalam mengenal pasti kemungkinan wabak berlaku walaupun kes yang dilaporkan adalah rendah. Rajah 12 menunjukkan pseudokod bagaimana SMOTE diaplikasikan ke atas model Hutan Rawak.

```
ALGORITMA Random_Forest_Dengan_SMOTE

1. Seimbangkan data latihan menggunakan SMOTE:
   smote ← SMOTE(random_state = 42)
   (X_train_sm, y_train_sm) ← smote.fit_resample(X_train, y_train)
   Papar "After SMOTE - y_train class distribution:",
      Counter(y_train_sm)

2. Inisialisasi model Random Forest tanpa class_weight:
   rf_model ← RandomForestClassifier(random_state = 42)

3. Jalankan pensilangan silang 10-lipatan ke atas data SMOTE:
   cv_scores ← cross_val_score(rf_model, X_train_sm, y_train_sm, cv =
      10, skor = 'accuracy')
   min_cv ← purata(cv_scores)
   std_cv ← sisihan_piawai(cv_scores)
   Papar "CV Accuracy (Train Set w/ SMOTE):", min_cv, "±", std_cv

4. Latih model menggunakan set latihan SMOTE:
   latih(rf_model, X_train_sm, y_train_sm)

5. Buat ramalan ke atas set ujian asal:
   y_pred ← ramal(rf_model, X_test)
   y_proba ← ramal_kebarangkalian(rf_model, X_test)

6. Kira ketepatan model ke atas set ujian:
   test_accuracy ← skor(rf_model, X_test, y_test)
   Papar "Test Accuracy:", test_accuracy

TAMAT ALGORITMA
```

Rajah 12

Pseudokod ini menggambarkan prosedur pelaksanaan model Hutan Rawak yang digabungkan dengan teknik SMOTE bagi menangani isu ketidakseimbangan kelas dalam set data latihan. Langkah pertama melibatkan penggunaan SMOTE untuk menghasilkan semula sampel data minoriti secara sintetik, sekali gus memastikan pengagihan kelas dalam *y_train* menjadi seimbang. Seterusnya, model Hutan Rawak diinisialisasi tanpa pemberat kelas *(class_weight)* memandangkan data telah diseimbangkan. Pensilangan silang *(cross-validation)* 10-lipatan digunakan ke atas set data baharu tersebut bagi menilai kestabilan dan ketepatan model secara awal. Model kemudiannya dilatih menggunakan set latihan SMOTE, dan proses ramalan dijalankan ke atas set ujian asal yang tidak mengalami sebarang perubahan. Ketepatan model dinilai berdasarkan keputusan ramalan terhadap set ujian, dan prestasi akhir dipaparkan untuk tujuan analisis.

Jadual 4 pula menunjukkan keputusan yang diperoleh untuk tiga model terbaik setelah SMOTE diaplikasikan ke atas ketiga-tiga model tersebut. Tiga model terbaik ini dipilih berdasarkan prestasi dalam mengeluarkan ketepatan tinggi dalam membuat ramalan.Pokok Keputusan dan SVM tidak dipilih berikutan prestasinya yang kurang memberangsangkan dalam menjana ramalan dengan ketepatan yang rendah dibandingkan Hutan Rawak, Pokok Keputusan dan KNN.

Jadual 4

Model	Ketepatan Ujian	AUC	Ketetpatan Makro	Ingatan Makro	Skor F1 Makro
Hutan Rawak + FE + SMOTE	0.6639	0.7663	0.5532	0.5922	0.5673
Pokok Keputusan + FE + SMOTE	0.6401	0.7369	0.5403	0.5877	0.5535
KNN + FE + SMOTE	0.6409	0.7342	0.5111	0.5355	0.5199

Keputusan dalam Jadual 4 menunjukkan bahawa aplikasi teknik SMOTE memberikan kesan yang bercampur kepada prestasi model. Untuk Hutan Rawak dan Pokok Keputusan, penggunaan SMOTE selepas kejuruteraan fitur dilihat membantu meningkatkan ketepatan ujian, AUC, serta skor makro F1, menunjukkan bahawa pengimbangan semula kelas melalui SMOTE membolehkan model mengenal pasti corak daripada kelas minoriti dengan lebih baik. Sebagai contoh, Hutan Rawak meningkat daripada F1 Makro 0.5489 kepada 0.5673. Walau bagaimanapun, bagi model KNN, penerapan SMOTE menyebabkan sedikit penurunan dalam semua metrik utama termasuk ketepatan ujian dan F1 Makro, mencadangkan bahawa KNN mungkin sensitif terhadap data sintetik dan tidak memberi tindak balas yang baik terhadap penambahan data dari kelas minoriti. Oleh itu, kesesuaian SMOTE bergantung kepada jenis algoritma yang digunakan, di mana ia lebih memberi manfaat kepada model seperti Hutan Rawak dan Pokok Keputusan berbanding KNN.

iii. Penelaan Hiperparameter

Daripada kaedah-kaedah yang telah diaplikasikan kepada model-model ini, dua model menunjukkan prestasi terbaik iaitu model Hutan Rawak + FE + SMOTE dengan ketepatan sebanyak 66.39% dan model terbaik iaitu KNN dengan ketepatan 69.10%. Langkah terakhir yang diambil untuk meningkatkan prestasi kedua-dua model tersebut adalah melalui teknik penelaan hiperparameter. Tujuan teknik ini adalah untuk mencari kombinasi parameter paling optimum agar model memberikan prestasi terbaik. Kaedah *Grid Search Cross Validation* digunakan untuk menguji pelbagai kombinasi parameter bagi dua model terbaik iaitu Hutan Rawak dan KNN. Parameter seperti bilangan pokok (*n_estimators*) dan jiran terdekat (*n_neighbors*) ditentukan berdasarkan nilai yang memberikan prestasi tertinggi. Proses ini membantu meningkatkan ketepatan dan kebolehpercayaan model dalam meramal tahap wabak denggi. Contoh aplikasi penelaan hiperparameter terhadap model KNN dapat dilihat dalam Rajah 13.

```
ALGORITMA: Penalaan Hiperparameter Model KNN Menggunakan GridSearchCV

1. Tetapkan julat hiperparameter untuk model KNN:
   - knn__n_neighbors ← [3, 5, 7, 9]
   - knn__weights ← ['uniform', 'distance']
   - knn__metric ← ['euclidean', 'manhattan']

2. Laksanakan pencarian grid (GridSearchCV) dengan parameter berikut:
   - Model asas ← knn_pipeline
   - Parameter grid ← param_grid_knn
   - Silang-sah ← 5-lipat (cv = 5)
   - Proses serentak ← Ya (n_jobs = -1)
   - Paparan proses ← verbose = 2

3. Latih model KNN terbaik hasil daripada GridSearchCV menggunakan set
      latihan:
   - grid_knn.fit(X_train, y_train)

4. Jalankan ramalan terhadap set ujian:
   - y_pred_knn ← grid_knn.predict(X_test)
   - y_proba_knn ← grid_knn.predict_proba(X_test)

TAMAT ALGORITMA
```

Rajah 13

Pseudokod di atas menerangkan proses penalaan hiperparameter bagi model KNN menggunakan teknik GridSearchCV. Langkah pertama menetapkan kombinasi nilai yang berbeza bagi tiga hiperparameter utama: bilangan jiran terdekat *(n_neighbors),* kaedah pemberat *(weights)*, dan metrik jarak *(metric)*. GridSearchCV kemudian digunakan untuk mencuba semua kombinasi tersebut secara sistematik melalui kaedah silang-sah 5-lipat, bagi mencari konfigurasi yang memberikan prestasi terbaik berdasarkan set latihan. Fungsi fit() melatih model menggunakan data latihan, manakala predict() dan predict_proba() digunakan untuk membuat ramalan kelas dan kebarangkalian bagi data ujian. Tujuan utama penggunaan

fungsi-fungsi ini adalah untuk memastikan model KNN yang dibina adalah dioptimumkan melalui penalaan parameter, sekaligus meningkatkan ketepatan ramalan terhadap data baharu.

Jadual 5 di bawah adalah keputusan akhir bagi dua model ML dengan prestasi terbaik setelah penambahbaikan dilakukan secara berperingkat; kejuruteraan fitur (FE), SMOTE dan Penelaan Hiperparameter.

Jadual 5

Model	Ketepatan Ujian	AUC	Ketetpatan Makro	Ingatan Makro	Skor F1 Makro
Hutan Rawak + FE + SMOTE + HT	0.6820	0.7880	0.58	0.60	0.59
KNN + FE + SMOTE + HT	0.7027	0.7042	0.65	0.57	0.60

Keputusan tersebut menunjukkan bahawa model KNN + FE + SMOTE + HT mempunyai prestasi terbaik dengan ketepatan ujian tertinggi iaitu 70.27%, diikuti oleh Hutan Rawak + FE + SMOTE + HT sebanyak 68.20%. Dari segi ketepatan makro dan skor F1 makro, model KNN juga mencatat nilai lebih tinggi iaitu masing-masing 0.65 dan 0.60, berbanding Hutan Rawak yang mencatat 0.58 dan 0.59. Walau bagaimanapun, dari aspek AUC, Hutan Rawak sedikit mendahului dengan nilai 0.7880 berbanding 0.7042 bagi KNN. Ini menunjukkan Hutan Rawak mempunyai kebolehan pemisahan kelas yang baik secara keseluruhan, tetapi KNN memberikan prestasi lebih seimbang dan stabil dari sudut ketepatan dan skor F1, menjadikannya model paling sesuai dan efektif untuk digunakan dalam sistem ramalan wabak denggi ini.

Fasa Pembangunan Sistem AwasDenggi!

Sistem AwasDenggi! dibina sebagai aplikasi berasaskan web menggunakan kerangka kerja Streamlit. Komponen sistem in boleh dibahagikan kepada dua bahagian utama iaitu *front-end* (antaramuka pengguna) dan juga *back-end* (logik ramalan dan model ML).

Dari segi pembangunan antaramuka pengguna *(front-end),* sistem ini dibangunkan menggunakan platform Streamlit, iaitu kerangka kerja sumber terbuka berasaskan Python yang memudahkan pembinaan aplikasi web interaktif untuk analisis data dan pembelajaran mesin. Antaramuka sistem direka bentuk agar mesra pengguna dan minimalis, membolehkan pengguna memasukkan input seperti minggu epidemik (ACC_WEEK), suhu maksimum, suhu minimum, suhu purata, kelembapan, bulan dan lokasi (alamat). Setelah input dimasukkan, sistem akan menjana tiga ciri tambahan secara automatik menggunakan kaedah kejuruteraan fitur. Hasil ramalan dipaparkan dalam bentuk label risiko seperti “Zone 0”, “Zone 1” atau “Zone 2” bersama visual amaran berwarna (kuning, merah atau hijau) yang membantu pengguna memahami tahap risiko secara intuitif.

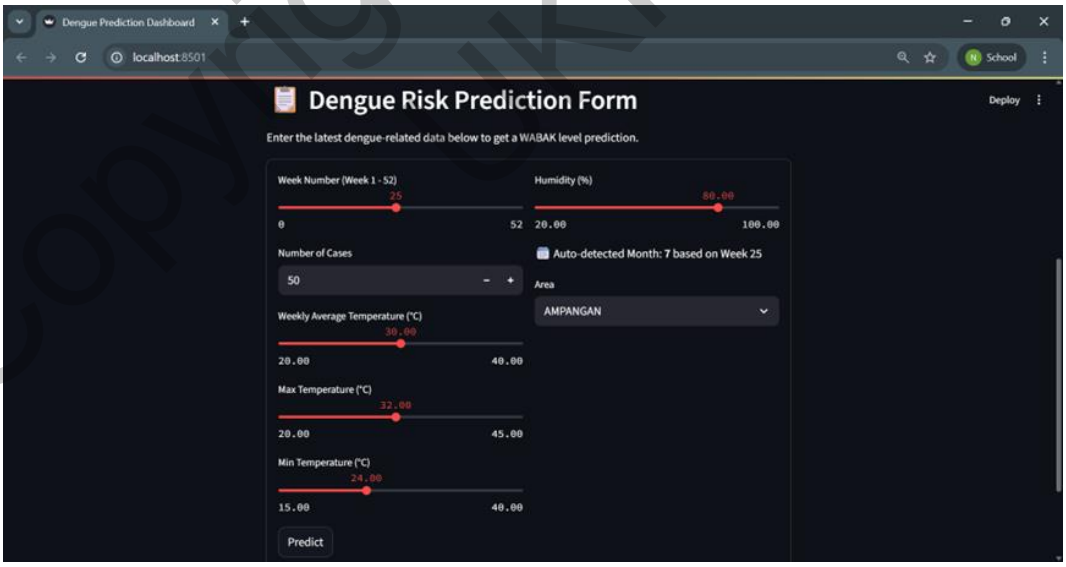
Bagi pembangunan sisi pelayan *(back-end),* sistem ini menggunakan model KNNyang telah dilatih sepenuhnya dan disimpan dalam bentuk fail .pkl menggunakan pustaka joblib. Apabila pengguna menghantar input, sistem akan menjalankan proses pra-pemprosesan seperti penskalaan dan pengekodan secara automatik sebelum data dihantar ke model untuk membuat ramalan. Selain itu, semua proses seperti pemetaan label ramalan, pemaparan kebarangkalian dan penjanaan visual keputusan dijalankan secara dinamik di peringkat pelayan.

Fungsi predict_proba() turut digunakan untuk memberikan kebarangkalian bagi setiap kelas yang diramalkan, yang kemudiannya dipaparkan kepada pengguna dalam bentuk bar progres. Keseluruhan sistem ini dibina dan diuji dalam persekitaran Anaconda menggunakan Python, dan boleh disebarkan ke mana-mana pelayan dengan konfigurasi yang sesuai untuk aplikasi Streamlit.

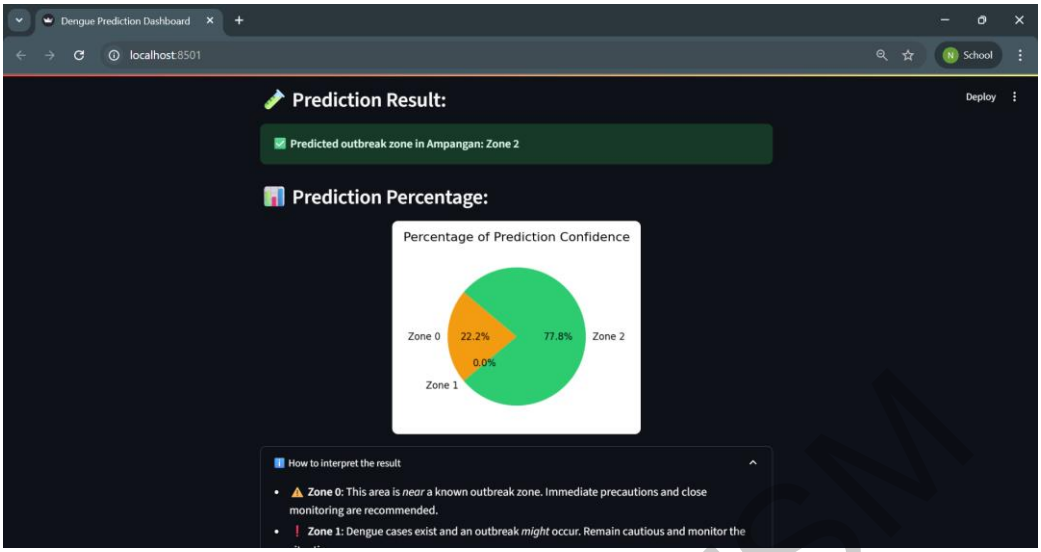
HASIL KEPUTUSAN DAN PERBINCANGAN

Selepas pembangunan sistem ramalan interaktif berdasarkan model pembelajaran mesin, satu proses pengujian dijalankan untuk menilai kebolehgunaan dan ketepatan sistem dalam senario sebenar. Dua jenis pengujian utama telah dilaksanakan, iaitu Pengujian Berfungsi dan Pengujian Tidak Berfungsi.

Pengujian Berfungsi menggunakan pendekatan *Black Box Testing*, di mana tumpuan diberikan kepada output sistem berdasarkan input yang dimasukkan oleh pengguna, tanpa mengambil kira bagaimana sistem berfungsi secara dalaman. Ujian ini melibatkan semua fungsi utama sistem termasuk pemprosesan input, kejuruteraan fitur automatik, pemanggilan model terlatih, dan pemaparan keputusan. Hasil ujian menunjukkan bahawa sistem dapat menjalankan fungsi-fungsi ini dengan baik di mana input yang dimasukkan menghasilkan output ramalan yang sepadan, manakala elemen visual seperti label zon dan mesej amaran dipaparkan dengan betul mengikut logik sistem. Rajah 14 menunjukkan contoh antaramuka sistem untuk pengguna memasukkan input. Setelah data dihantar, sistem AwasDenggi! akan menjana ramalan tahap wabak dan memaparkan keputusan dalam bentuk label zon bersama kebarangkalian serta isyarat visual berwarna seperti yang ditunjukkan dalam Rajah 15.



Rajah 14



Rajah 15

Pengujian Tidak Berfungsi pula memberi tumpuan kepada kestabilan sistem dalam menangani pelbagai input dan kebolehgunaan antara muka pengguna. Ujian ini menunjukkan bahawa sistem dapat beroperasi secara konsisten dengan pelbagai jenis input tanpa menghasilkan ralat, dan antaramuka yang dibangunkan menggunakan Streamlit adalah mudah difahami dan digunakan oleh pengguna bukan teknikal. Walaupun tidak diuji dari sudut kecekapan prestasi seperti masa tindak balas secara kuantitatif, sistem beroperasi dengan lancar sepanjang proses pengujian tanpa gangguan teknikal.

KESIMPULAN

Secara keseluruhan, projek pembangunan sistem AwasDenggi! telah berjaya mencapai matlamatnya untuk menjana ramalan tahap penularan wabak denggi di Seremban, Negeri Sembilan menggunakan teknik pembelajaran mesin. Melalui penggunaan metodologi CRISP-DM, data epidemiologi dan persekitaran dianalisis serta diproses dengan teliti bagi membina model yang mampu mengklasifikasikan tahap penularan wabak ke dalam tiga kategori berisiko. Model pembelajaran mesin yang digunakaniaitu algoritma KNN telah menunjukkan prestasi yang memuaskan dan berjaya diintegrasikan ke dalam sistem web yang mesra pengguna. Walaupun terdapat beberapa cabaran sepanjang proses pembangunan, semua kekangan tersebut telah dapat diatasi dengan pendekatan teknikal yang sesuai. Sistem ini berpotensi untuk terus diperluaskan dan ditambah baik bagi membantu usaha kawalan dan pencegahan denggi secara lebih proaktif pada masa akan datang.

Kelemahan Sistem

Terdapat beberapa kelemahan yang dikenal pasti sepanjang pembangunan sistem ini. Pertama, sistem ini masih bergantung kepada input manual daripada pengguna, yang boleh menyebabkan ketidaktepatan jika data yang dimasukkan tidak sah atau tidak terkini. Selain itu, sistem ini hanya dilatih menggunakan data dari satu lokasi iaitu daerah Seremban, Negeri Sembilan, yang mungkin mengehadkan kebolehgunaan model di lokasi lain yang mempunyai corak epidemiologi dan persekitaran yang berbeza. Tambahan pula, sistem ini tidak mengambil kira data masa nyata seperti bacaan cuaca semasa atau pergerakan populasi yang boleh mempengaruhi penularan wabak.

Kekuatan Sistem

Sistem ini menunjukkan potensi besar sebagai alat sokongan keputusan bagi pihak berkuasa kesihatan mahupun orang awam. Ia mampu memberi amaran awal kepada pengguna berdasarkan analisis data sedia ada, memaparkan output dalam bentuk visual yang mudah difahami serta menampilkan antara muka yang mesra pengguna. Penggunaan pembelajaran mesin secara holistik juga memberi kelebihan dalam memahami pola data yang kompleks berbanding kaedah statistik tradisional. Kejayaan membangunkan dan menguji sistem ini membuka ruang kepada penambahbaikan di masa hadapan seperti integrasi data masa nyata dan peluasan skop ramalan ke daerah atau negeri lain di Malaysia.

PENGHARGAAN

Alhamdulillah syukur ke hadrat Ilahi, pertama sekali saya ingin memanjatkan kesyukuran kepada Allah SWT kerana dengan berkat-Nya saya dapat menyiapkan projek ini.

Saya juga ingin mengucapkan setinggi-tinggi penghargaan kepada penyelia saya iaitu Prof. Madya Dr. Suhaila Zainudin kerana telah banyak memberikan tunjuk ajar serta banyak membantu dalam menyiapkan projek ini. Beliau juga tidak jemu dalam memberikan tunjuk ajar sepanjang menyiapkan projek tahun akhir ini yang akhirnya banyak membuka minda saya terutamanya berkaitan pembelajaran mesin.

Penghargaan khas ingin saya tujukan kepada kedua ibu bapa, adik-beradik serta rakan-rakan atas dorongan serta yang sentiasa diberikan sepanjang semester ini dan juga tidak lekang dalam bertukar pendapat sepanjang menyiapkan projek ini. Tanpa sokongan mereka, saya mungkin tidak mampu menyiapkan projek ini. Terima kasih semua.

RUJUKAN

Khan, M. B., Khan, A., Fatima, M., & others. (2023). Dengue overview: An updated systemic review. *Journal of Infection and Public Health, 16*(9), 1241–1248. https://doi.org/10.1016/j.jiph.2023.07.012

Majeed, M. A., Mohd, Z., Zed Zulkafli, & Aimrun Wayayok. (2023). A deep learning approach for dengue fever prediction in Malaysia using LSTM with spatial attention. International Journal of Environmental Research and Public Health, 20(5), 4130. https://doi.org/10.3390/ijerph20054130

Salim, N. A. M., Wah, Y. B., Reeves, C., Smith, M., Yaacob, W. F. W., Mudin, R. N., Dapari, R., Sapri, N. N. F. F., & Haque, U. (2021). Prediction of dengue outbreak in Selangor, Malaysia using machine learning techniques. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-020-79193-2

Nurul Nabihah Mohd Aminudin (A195948)
Assoc. Prof. Dr. Suhaila Zainudin
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia