ANALISIS PREDIKTIF KESAN COVID-19 JANGKA PANJANG KE ATAS PESAKIT POSITIF COVID-19 MENGGUNAKAN ALGORITMA PENGELOMPOKAN

YEE XIN JIE

PROF. DR. AZURALIZA ABU BAKAR

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

ABSTRAK

Kesan COVID-19 Jangka Panjang dikenal pasti sebagai satu siri simptom dan komplikasi yang berterusan berikutan fasa akut COVID-19. COVID-19 jangka panjang telah muncul sebagai kebimbangan kesihatan yang serius. Wujudnya pelbagai simptom berkaitan dengan COVID-19 jangka panjang yang boleh menjejaskan hampir seluruh sistem badan manusia, termasuk keletihan, sesak nafas, sakit dada, kabus otak, insomnia serta pelbagai isu fizikal dan psikologi yang lain. Simptom-simptom ini boleh berlangsung secara berterusan selama beberapa minggu, bulan, atau lebih lama lagi. Hal ini akan mengurangkan kualiti hidup dan menjejaskan kehidupan bagi mereka yang terjejas. Selain itu, kejadian ini boleh membawa implikasi sosial dan ekonomi yang mendalam, kerana individu yang menghidap COVID-19 jangka panjang sering bergelut untuk kembali bekerja atau melakukan aktiviti harian. Salah satu cabaran paling ketara dalam menangani COVID-19 jangka panjang adalah kekurangan pemahaman yang jelas tentang punca dan faktor risikonya. Oleh itu, objektif kajian ini bertujuan untuk mengenal pasti faktor yang mempengaruhi COVID-19 jangka panjang. Dalam kajian ini, algoritma pengelompokan digunakan untuk mengenal pasti corak dalam kalangan pesakit yang berkemungkinan menghidapi COVID-19 jangka panjang daripada kohort individu yang telah diuji positif untuk COVID-19. Set data klinikal telah digunakan dalam analisis kami diperoleh daripada laman web *Dataverse Havard*, yang terdiri daripada kira-kira 1,130 rekod dengan 186 atribut pesakit COVID-19 jangka panjang yang berpotensi di seluruh Eropah. Keberkesanan model pengelompokan ini dinilai menggunakan metrik evaluasi seperti *Davies Bouldin Index, Average within Centroid Distance, Silhouette Score* dan *Sum of Squares*. Pencapaian markah yang lebih baik dalam metrik evaluasi bukan sahaja mengesahkan potensi model tetapi juga membuktikan pengaplikasian dalam domain penjagaan kesihatan. Keupayaan untuk meramalkan pesakit COVID-19 jangka panjang dengan tepat mempunyai implikasi yang ketara, termasuk pencegahan awal untuk menambah baik penjagaan pesakit secara menyeluruh dan pengagihan sumber perubatan. Projek ini adalah sebahagian dari projek perundingan UKM-NIOSH "Kajian Simptom & Faktor Risiko Long Covid Dalam Kalangan Pekerja Sektor Pembuatan Di Malaysia" UKMP-S230424 oleh Prof. Dr. Azuraliza Abu Bakar.

Kata kunci: COVID-19 jangka panjang, Algoritma pengelompokan

PENGENALAN

Sindrom COVID-19 jangka panjang dikenali sebagai satu siri simptom dan komplikasi berterusan selepas fasa akut COVID-19. Ia telah menjadi kebimbangan kesihatan yang serius, dianggarkan menjejaskan sekurang-kurangnya 65 juta individu di seluruh dunia. Kejadian ini dianggarkan melibatkan 10-30% kes yang tidak memerlukan rawatan hospital, 50-70% kes yang memerlukan rawatan hospital, dan 10-12% kes yang telah divaksinasi (Davis et al. 2023). COVID-19 jangka panjang menjejaskan semua peringkat usia, dengan perkadaran tertinggi dalam kalangan mereka berusia 36 hingga 50 tahun. Kebanyakan kes ditemui pada bekas pesakit COVID-19 yang serius, individu yang tidak menerima vaksin, mereka dengan sindrom keradangan multisistem (MIS), dan mereka yang menghidap penyakit kronik seperti diabetes, asma, dan obesiti. Walau bagaimanapun, sesiapa sahaja yang pernah dijangkiti boleh menjadi mangsa COVID-19 jangka panjang (Health 2023). Menurut CDC, simptom yang paling kerap dilaporkan termasuk keletihan, sesak nafas, sakit dada, kabus otak, demam, insomnia, serta pelbagai isu fizikal dan psikologi lain (Anon 2023).

Salah satu cabaran utama dalam menangani COVID-19 jangka panjang adalah kekurangan pemahaman yang jelas tentang punca dan faktor risikonya. Sehingga kini, belum ada ujian makmal yang boleh mendiagnosis COVID-19 jangka panjang dengan tepat. Kes pesakit yang mempunyai simptom teruk walaupun ujian darah, sinar-X dada, dan EKG menunjukkan keputusan normal masih berlaku (Katella 2023). Diagnosis COVID-19 jangka panjang kini bergantung pada penilaian doktor terhadap simptom berterusan selepas jangkitan. Para saintis sedang berusaha untuk mengenal pasti penanda biokimia atau ujian makmal yang relevan (Belluck 2023).

Pembelajaran mesin, satu cabang kecerdasan buatan, membolehkan mesin untuk belajar daripada data dan mencari corak tanpa campur tangan manusia yang meluas (Flam 2023). Keupayaan pembelajaran mesin untuk memproses data besar dan kompleks serta keupayaan ramalan yang hebat menjadikannya medium ideal untuk kajian data kesihatan digital (Davenport et al. 2019). Terdapat kajian yang telah menggunakan pembelajaran mesin untuk meramal dan mengenal pasti faktor yang menyumbang kepada COVID-19 jangka panjang. Melalui pembelajaran mesin, pasukan penyelidik dari Amerika Syarikat yang disokong oleh National Institutes of Health (NIH) dapat mengenal pasti ciri-ciri individu yang mengalami dan mungkin mengalami COVID-19 jangka panjang (Pfaff et al. 2022). Mereka membina model XGBoost yang mengenal pasti pesakit berpotensi mengalami COVID-19 jangka panjang dengan menganalisis rekod kesihatan digital dalam pangkalan data National COVID Cohort Collaborative (N3C). Selain itu, pasukan penyelidik dari Scotland menggunakan analisis pengelompokan dan model regresi-logistik-multivariat untuk mengenal pasti faktor risiko epidemiologi berkaitan COVID-19 jangka panjang (Daines et al. 2022).

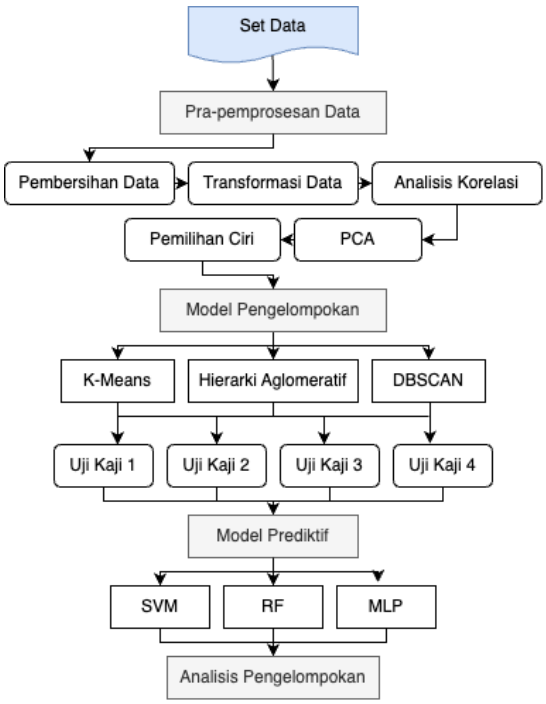
PERNYATAAN MASALAH

Penggunaan pembelajaran mesin dalam prognosis COVID-19 jangka panjang masih kurang diterokai, walaupun teknologi ini semakin meluas dalam perubatan. Kelemahan utama kajian sedia ada termasuk ketidakpastian data, kesukaran mengenal pasti faktor penyebab, dan kekurangan data kes COVID-19 jangka panjang. Penggunaan data besar melalui pembelajaran mesin adalah positif kerana ia dapat menganalisis gejala COVID-19 yang kompleks dan berpanjangan, memberi kesan besar pada kualiti hidup dan perkhidmatan kesihatan.

OBJEKTIF KAJIAN

Objektif utama bagi kajian ini adalah menggunakan algoritma pengelompokan untuk mengenal pasti faktor yang mempengaruhi COVID-19 jangka panjang. Selain itu, kajian ini membangunkan model prediktif bagi mengesah kualiti dan kestabilan kelompok-kelompok. Seterusnya menilai prestasi model pengelompokan dan prediktif.

METODOLOGI KAJIAN



Rajah 1: Carta aliran kajian

a. Pemahaman dan Penyediaan Data

Sumber data klinikal yang diterapkan dalam kajian ini dimuat turun daripada laman web pangkalan data Universiti Havard iaitu Havard Dataverse:Long Covid Dataset. Menurut kajian ini, set data yang digunakan terdiri daripada kira-kira 1,100 rekod dengan 186 atribut pesakit

COVID-19 jangka panjang yang berpotensi di seluruh Eropah. Dalam kajian ini, pemprosesan data rekod kesihatan digital pesakit positif COVID-19 akan menjalankan langkah-langkah penyediaan data seperti pembersihan data, transformasi data, analisis korelasi, analisis komponen utama (PCA) dan pemilihan ciri. Dalam langkah pembersihan data, data yang bertindih perlu ditangani untuk mencegah pertindihan maklumat dan memastikan kejelasan dalam tafsiran data. Selain itu, nilai ralat, nilai hilang, data yang tidak lengkap dan tidak konsisten harus dibuang bagi meningkatkan ketepatan data. Set data akan dibahagikan kepada *DataFrame* numerik dan *DataFrame* kategori. Selanjutnya, proses pembersihan data akan menjadi berbeza bergantung kepada jenis data, baik itu numerik mahupun kategori. Kehilangan nilai untuk data numerik akan diisi dengan nilai min manakala kehilangan nilai untuk data kategori akan diisi dengan nilai 'unknown'. Untuk transformasi data, *One-Hot Encoding* dilakukan untuk menukar data kategori kepada data matriks binari, di mana satu lajur baru akan dicipta untuk setiap nilai unik dalam lajur kategori yang asal (Seger 2018). Selain itu, *Label-Encoding* dijalankan untuk data kategori yang berkardinaliti. Akhir sekali, penskalaan *Min-Max* dijalankan atas data numerik supaya ciri-ciri tersebut mempunyai skala yang sama dan meningkatkan prestasi algoritma pembelajaran mesin. Di samping itu, matriks korelasi *upper triangle* digunakan untuk mengatur nilai korelasi kepada format yang mudah difahami dengan mengecualikan nilai berulangan . Ambang korelasi sebanyak 0.9 telah ditetapkan dan pasangan ciri-ciri yang mempunyai nilai korelasi melebihi nilai ini dianggap mempunyai korelasi yang kuat dan boleh digugurkan. Tambahan pula, PCA digunakan untuk mengenal pasti set ciri yang dikurangkan yang mewakili data asal dalam dimensi data yang lebih rendah dengan kehilangan maklumat yang minimum (Kherif et al. 2020). Dalam set data yang digunakan boleh memperlihatkan masih banyak ciri selepas transformasi data dan analisis korelasi. Oleh itu, *Extra Trees Classifier* digunakan untuk memilih 30 ciri yang paling penting dan relevan dengan pemboleh ubah sasaran.

b. Pembangunan Model

Tiga jenis algoritma pengelompokan digunakan untuk mengelompokkan pesakit COVID-19 jangka panjang ke dalam kumpulan yang serupa berdasarkan ciri-ciri mereka, seperti gejala dan faktor risiko iaitu pengelompokan K-Means, pengelompokan Hierarki Aglomeratif dan DBSCAN. Algoritma-algoritma tersebut dapat menyelesaikan masalah pengelasan dan dikenali sebagai pendekatan pembelajaran tidak terkawal yang paling luas digunakan untuk model pengelompokan. Seterusnya, label kelompok akan ditambah kepada setiap titik data berdasarkan hasil pengelompokan sebelum pembinaan model prediktif. Label ini berfungsi sebagai pemboleh ubah sasaran (*target variable*) untuk model pembelajaran terkawal. Terdapat beberapa model prediktif telah diaplikasikan untuk melatih model yang dapat meramalkan keanggotaan kelompok dengan tepat untuk titik data baru yang tidak terlihat, iaitu *Support Vector Machine* (SVM), *Random Forest* (RM) dan *Multilayer Perceptron* (MLP).

Bagi model pengelompokan, pengelompokan K-Means merupakan salah satu algoritma pengelompokan yang paling terkenal dan banyak digunakan kerana algoritma ini

dapat mengenal pasti struktur tersembunyi secara automatik (Zubair et al. 2021). Algoritma pengelompokan tersebut bertujuan untuk mengasingkan set data ke dalam kelompok K yang ditakrifkan sebelumnya yang berbeza dan tidak bertindih, di mana setiap titik data hanya tergolong dalam satu kumpulan. Seterusnya, pengelompokan hierarki aglomeratif ialah sejenis algoritma pengelompokan yang membina kelompok bersarang dengan menggabungkan titik data secara berturut-turut (Müllner 2011). Algoritma bermula dengan menganggap setiap titik data sebagai gugusan yang berasingan dan kemudian secara berulang menggabungkan kelompok terdekat berdasarkan metrik jarak. Hierarki kelompok boleh diwakili sebagai gambar rajah pokok yang dikenali sebagai *dendrogram*. Terdapat beberapa pilihan untuk mengabungkan titik data secara berturut-turut, termasuk pautan tunggal (jarak minimum), pautan lengkap (jarak maksimum), pautan purata (jarak purata) dan pautan Ward. Tambahan pula, DBSCAN adalah alat yang berguna untuk mengenal pasti struktur kompleks dalam data pengguna untuk pembahagian set data. DBSCAN sangat berkesan dalam mengenali kelompok dengan pelbagai bentuk dan ketumpatan, terutamanya dalam set data yang *noise* dan tidak seragam (Hicham et al. 2022). DBSCAN mengelompokkan titik data berdasarkan kedekatan dan ketumpatannya, mengenal pasti titik teras dengan bilangan minimum titik berhampiran. Algoritma ini boleh mengenal pasti kelompok padat dan jarang, memberikan pemahaman menyeluruh tentang trend tingkah laku set data. DBSCAN dapat mengenal pasti kelompok dengan bentuk dan saiz sebarang bentuk dan tahan terhadap *outliers*.

Untuk model prediktif, RF merupakan konsep pembelajaran ensemble yang menggabungkan pelbagai pengelas untuk menyelesaikan masalah yang kompleks bagi meningkatkan prestasi model. Algoritma ini merupakan pengelas yang mengandungi beberapa pokok keputusan pada subset data yang diberikan dan mengambil purata untuk meningkatkan ketepatan ramalan set data tersebut. RF tidak hanya bergantung kepada satu pokok keputusan tetapi mengambil keputusan ramalan daripada setiap pokok dan meramalkan keluaran terakhir berdasarkan undian majoriti ramalan. SVM pula, merupakan algoritma yang popular dalam menyelesaikan masalah klasifikasi dan regresi. Objektif utama mesin vektor sokongan adalah untuk mencari *hyperplane* dalam ruang dimensi *N* yang dapat mengelaskan titik data dengan nyata. SVM mewujudkan sempadan keputusan yang akan memisahkan ruang dimensi *N* kepada kelas yang beza untuk memastikan data atau titik baru dapat dikelaskan ke dalam kategori yang betul. Sempadan keputusan dikenali sebagai *hyperplane*. Titik data dekat pada *hyperplane* yang mempengaruhi lokasi *hyperplane* dirujuk sebagai vektor sokongan. Selain itu, MLP ialah varian model *Perceptron* asal yang dicadangkan oleh Rosenblatt pada tahun 1950-an (Rosenblatt 1958). Algoritma ini merupakan rangakaian neural *feedforward* yang mempunyai satu atau lebih lapisan tersembunyi antara lapisan input dan output, neuron-neuron diatur dalam lapisan-lapisan tersebut, isyarat input sentiasa diarahkan kepada satu araha dari lapisan ke lapisan, dan neuron-neuron dalam lapisan yang sama tidak saling berhubung.

c. Penilaian Model

Dalam kajian ini, empat metrik penilaian untuk model pengelompokan akan digunakan iaitu

Davies Bouldin Indeks (DBI), *Silhouette Score*, *Average within Centroid Distance* (AWCD) dan *Sum of Squared Error* (SSE). DBI adalah salah satu kaedah yang digunakan untuk mengukur kesahihan kelompok dalam sesuatu kaedah pengelompokan. Pengukuran dengan DBI adalah memaksimumkan jarak antara kelompok (*inter-cluster*) dan pada masa yang sama DBI cuba untuk meminimumkan jarak antara titik dalam satu kelompok (*intra-cluster*). Selain itu, *Silhouette Score* merupakan salah satu kaedah yang digunakan untuk menilai kesahihan pengelompokan. Menurut (Salihoun 2020), nilai purata yang diperoleh dari *Silhouette Score* menunjukkan dengan tepat bahawa betapa optimum bilangan kelompok yang dikelaskan. Nilai-nilai yang diperoleh selepas mempraktikkan *Silhouette Score* adalah antara -1 hingga +1. Semakin tinggi nilainya, semakin dekat objek itu dengan kelompoknya, dan sebaliknya, semakin kecil nilai yang diperoleh, semakin jauh objek itu dengan kelompoknya. AWCD pula mengira jarak purata setiap titik data ke centroid kelompok yang diberikan. Jarak purata yang lebih rendah menunjukkan kelompok yang lebih padat (Abdul Rahman et al. 2021). Seterusnya, SSE akan mengira jumlah jarak kuasa dua setiap titik data ke centroid kelompok yang diberikan (Van Beek 2005). Jika semua kes dalam satu kelompok adalah sama, maka SSE akan sama dengan 0. Metrik ini memberikan pandangan tentang penyebaran titik data dalam kelompok.

ANALISIS KEPUTUSAN

Reka bentuk eksperimen ini adalah untuk menganalisis dan mengenal pasti kelompok yang bermakna dalam set data *Long Covid* menggunakan teknik pembelajaran mesin tidak terkawal, dan seterusnya mengesahkan dan meramal keanggotaan kelompok menggunakan model pembelajaran mesin terkawal. Eksperimen ini disusun dalam empat uji kaji yang berbeza dengan mempraktikkan pelbagai algoritma pengelompokan dan menilai prestasi mereka menggunakan pelbagai metrik. Jadual 1 menunjukkan contoh bagi keputusan penilaian metrik model pengelompokan berdasarkan empat set data uji kaji yang berbeza. Setiap satu daripada empat set data uji kaji akan menggunakan tiga algoritma pengelompokan iaitu pengelompokan K-Means, Hierarki Aglomeratif dan DBSCAN.

Jadual 1: Contoh keputusan penilaian metrik model pengelompokan berdasarkan empat uji kaji set data

Uji kaji	**Set Data**	**Model-model Pengelompokan**	**Metrik penilaian**			
			SSE	*Silhouette Score*	DBI	AWCD
1	Processed Data		x	x	x	x
2	Processed Data + PCA		x	x	x	x
3	Selected Features Data		x	x	x	x
4	Selected Features Data + PCA		x	x	x	x

Bagi setiap kaedah pengelompokan dan konfigurasi, label kelompok ditambahkan kepada setiap titik data dalam set data uji kaji yang berkenaan. Label-label ini berfungsi sebagai pemboleh ubah sasaran (*target variable*) untuk model pembelajaran terkawal yang berikutnya. Empat-empat set eksperimen data dengan label kelompok yang telah ditambah

akan digunakan untuk melatih tiga model prediktif iaitu SVM, RF dan MLP. Jadual 2 menunjukkan contoh keputusan penilaian metrik model prediktif.

Jadual 2: Contoh keputusan penilaian metrik model prediktif berdasarkan empat uji kaji set data dengan label kelompok

Uji kaji	**Set Data**		**Metrik Penilaian**			
			Ketepatan	*Precision*	*Recall*	Skor F1
1	Processed Data + cluster labels	**Model-model Prediktif**	x	x	x	x
2	Processed Data + PCA + cluster labels		x	x	x	x
3	Selected Features Data + cluster labels		x	x	x	x
4	Selected Features Data + PCA + cluster labels		x	x	x	x

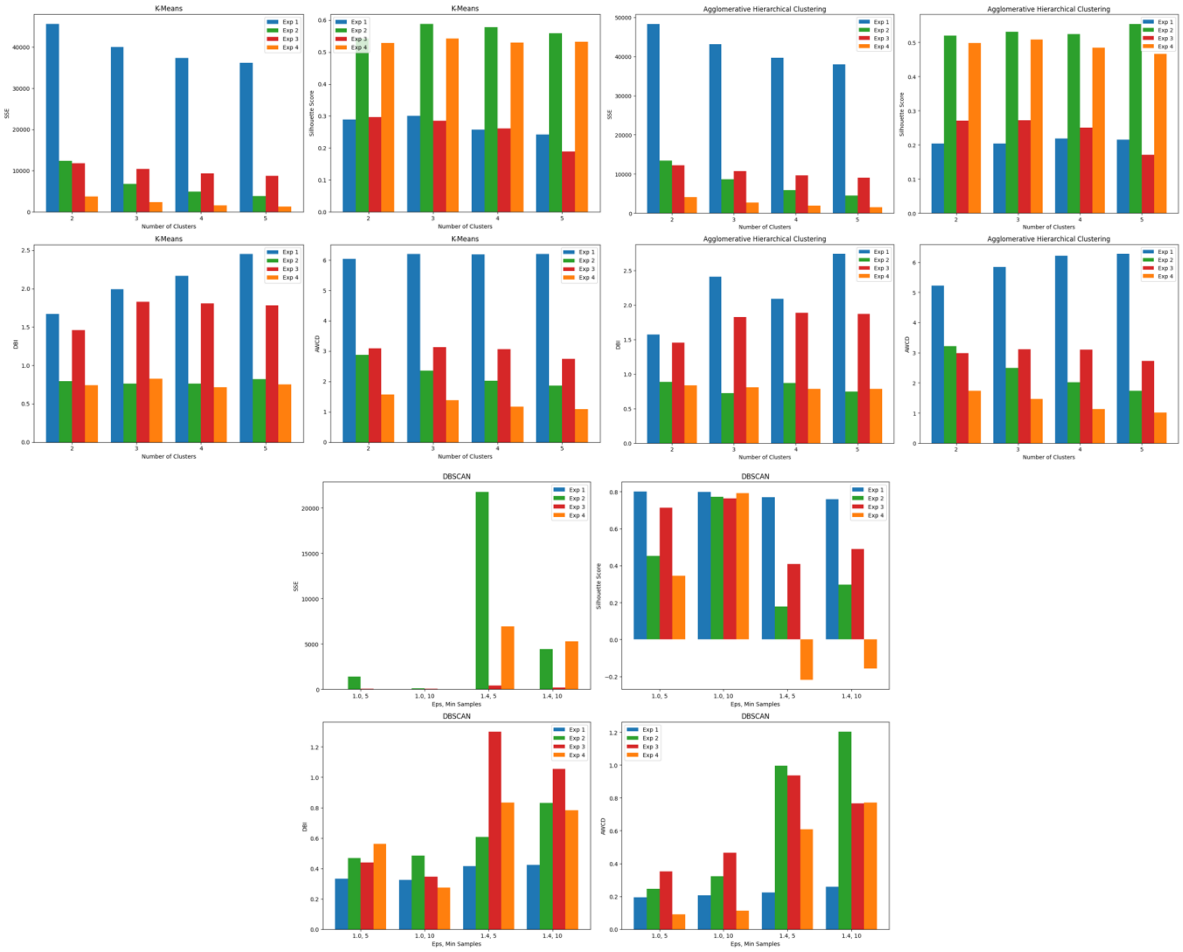
Uji kaji 1 menetapkan prestasi asas untuk pengelompokan menggunakan set data yang telah diproses sepenuhnya. Uji kaji ini menyediakan penanda aras untuk membandingkan kesan pemilihan ciri dan pengurangan dimensi. Bukan itu sahaja, uji kaji 1 menggunakan set data yang merangkumi keseluruhan atribut untuk memastikan tiada maklumat berharga yang diabaikan.

Uji kaji 2 menggunakan set data yang telah diproses sepenuhnya dan PCA untuk menilai kesan pengurangan dimensi terhadap prestasi pengelompokan. PCA membantu mengurangkan bilangan ciri sambil mengekalkan kebanyakan variasi dalam data. Kaedah ini sangat berguna apabila berurusan dengan set data *Long Covid* yang berdimensi tinggi (186 atribut). Selain itu, PCA dapat membantu menghapuskan *noise* dengan memberi tumpuan kepada komponen utama, yang membawa kepada pembentukan kelompok yang lebih jelas.

Uji kaji 3 pula menyelidiki kesan menggunakan subset atribut yang dipilih berdasarkan pemilihan ciri dengan *ExtraTreesClassifier*. Ciri yang dipilih berdasarkan metrik kepentingan dapat meningkatkan pengelompokan dengan memberi tumpuan kepada atribut yang paling bermakna. Tambahan pula, mengurangkan set ciri dapat menyederhanakan model dan mengurangkan kerumitan yang berpotensi membawa kepada keputusan yang lebih baik dan lebih boleh ditafsirkan.

Akhir sekali, uji kaij 4 bertujuan untuk menilai kesan gabungan pemilihan ciri dan pengurangan dimensi terhadap prestasi pengelompokan. Uji kaji ini menggabungkan pemilihan ciri dengan PCA memanfaatkan kekuatan kedua-dua teknik. Pemilihan ciri memastikan hanya ciri yang relevan dipertimbangkan, manakala PCA seterusnya mengurangkan dimensi. Di samping itu, uji kaji ini mengekalkan ciri yang paling bermakna dan mengurangkannya kepada komponen utama yang menangkap variasi maksimum. Gabungan ini dapat membawa kepada pengelompokan yang sangat efisien dan berkesan, yang berpotensi menemukan corak yang mungkin terlepas jika menggunakan semua ciri atau data yang telah diproses sahaja.

Rajah 3 telah menunjukkan graf perbandingan semua metrik penilaian antara setiap model pengelompokan yang diaplikasikan iaitu pengelompokan K-Means, Hierarki Aglomeratif dan DBSCAN.



Rajah 3: Graf perbandingan keputusan penilaian model pengelompokan

Menurut keputusan penilaian model-model prediktif SVM, RF dan MLP, nilai ketepatan, *precision*, *recall* dan skor F1 konsisten dengan julat 0.6824 hingga 0.8059. Julat metrik penilaian yang agak kecil menunjukkan bahawa teknik penyediaan data, model pengelompokan dan model prediktif adalah berkesan dan konsisten. Oleh itu demikian, boleh dikatakan bahawa model prediktif berjaya untuk mengesahkan kualiti dan kestabilan kelompok-kelompok serta meramal keanggotaan kelompok. Berdasarkan jadual 3, prestasi model prediktif yang terbaik dicapai oleh model SVM dengan nilai ketepatan 0.8059 yang menggunakan model pengelompokan K-Means dan nombor kelompok 3 dalam uji kaji 3 yang menggunakan *Selected Features Data*. Hal ini menunjukan SVM berkesan dalam mengendalikan data berdimensi tinggi dan mencari *hyperplane* optimum yang memaksimumkan margin antara kelas-kelas.

Jadual 3: Keputusan penilaian model SVM uji kaji 3

Model prediktif	**Model Pengelompokan**	**Nombor Kelompok/ eps, min samples**	**Metrik Penilaian**			
			Ketepatan	*Precision*	*Recall*	Skor F1
SVM	K-Means	2	0.7971	0.8000	0.8000	0.8000
		3	0.8059	0.8100	0.8100	0.8100
		4	0.7882	0.7900	0.7900	0.7900
		5	0.7941	0.8000	0.8000	0.8000
	Hierarki Aglomeratif	2	0.7912	0.7900	0.7900	0.7900
		3	0.8000	0.8000	0.8000	0.8000
		4	0.7882	0.7900	0.7900	0.7900
		5	0.7912	0.7900	0.7900	0.7900
	DBSCAN	1.0, 5	0.7941	0.8000	0.8000	0.7900
		1.0, 10	0.7941	0.8000	0.8000	0.7900
		1.4, 5	0.7824	0.7900	0.7800	0.7800
		1.4, 10	0.7971	0.8000	0.8000	0.8000

ANALISIS PENGELOMPOKAN

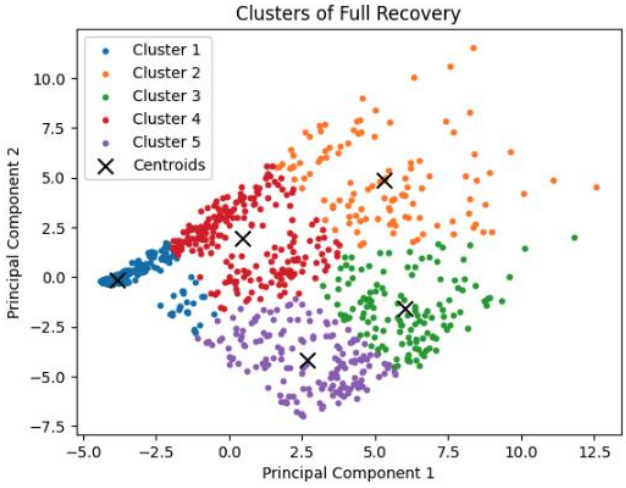
Berdasarkan keputusan penilaian model pengelompokan DBSCAN mencapai keputusan yang paling baik untuk keempat-empat uji kaji yang dijalankan. Jadual 4 menunjukkan keputusan penilaian yang paling baik menurut uji kaji 1 hingga uji kaji 4. Namun demikian, menurut rajah 8 dan rajah 9 yang memaparkan plot bersepah DBSCAN untuk uji kaji 2 dan 4, boleh diperhatikan bahawa terdapat banyak titik data telah dikelaskan sebagai *noise.* Kewujudan *noise* yang terlampau banyak dalam plot bersepah bukanlah sesuatu yang diingini kerana masalah ini menghalang penerokaan struktur sebenar dan hubungan dalam data, menjadikannya sukar untuk ditafsir dan dianalisis secara berkesan. Bukan itu sahaja, titik-titik *noise* boleh menyebabkan kekacauan visual, mengurangkan kejelasan sempadan kelompok dan menyebabkan pentafsiran yang salah terhadap taburan data. Hal ini telah menghalang keupayaan untuk mengenal pasti corak atau kelompok yang berbeza dan menyukarkan proses untuk membuat keputusan. Oleh itu demikian, model DBSCAN adalah tidak sesuai dalam kajian ini.

Jika membandingkan model pengelompokan K-Means dan Hierarki Aglomeratif sahaja, keputusan penilaian menunjukkan model pengelompokan K-Means mendapati nilai penilaian yang lebih baik. Berdasarkan keputusan penilaian jadual 4 serta visualisasi plot bersepah rajah 4 dan rajah 5, nombor kelompok yang paling baik dapat dijumpai apabila

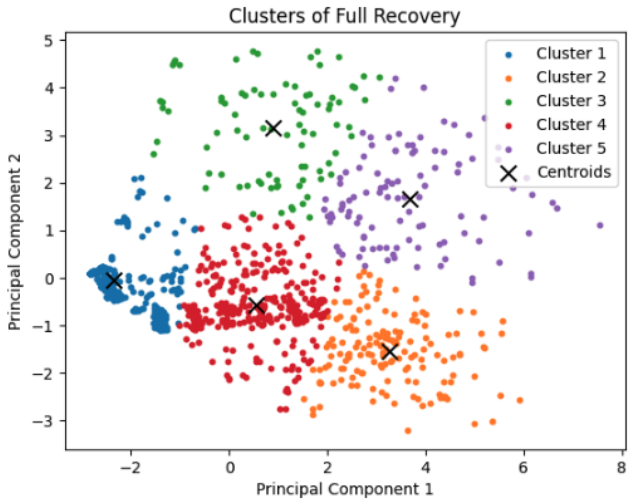
menggunakan model pengelompokan K-Means dan nombor kelompok 5 dalam uji kaji 4. Selain itu, kelompok dalam plot bersepah untuk model K-Means uji kaji 4 adalah jelas dan lebih mudah untuk membuat analisis pengelompokan. Hal ini demikian, model pengelompokan lebih sesuai dan nombor kelompok 5 uji kaji 4 model tersebut akan ditafsirkan dengan teliti supaya dapat memahami ciri-ciri apa yang membentuk kelompok-kelompok tersebut.

Jadual **Error! No text of specified style in document.**: Keputusan penilaian terbaik untuk setiap model pengelompokan dari uji kaji 1 ke uji kaji 4

Uji kaji	**Model Pengelompokan**	**Nombor kelompok/ eps, min samples**	**Metrik penilaian**			
			SSE	*Silhouette Score*	DBI	AWCD
1 (Processsed data)	K-Means	5	36193.2166	0.2426	2.4513	6.1993
	Hierarki Aglomeratif	5	38017.8108	0.2153	2.7387	6.2753
	DBSCAN	1.0, 10	7.2286	0.7980	0.3246	0.2061
2 (Processed Data + PCA)	K-Means	5	3819.3176	0.5586	0.8217	1.8698
	Hierarki Aglomeratif	5	4449.2473	0.5541	0.7460	1.7361
	DBSCAN	0.3, 10	83.6623	0.7717	0.4851	0.3219
3 (Selected Features Data)	K-Means	5	8723.5514	0.1890	1.7826	2.7411
	Hierarki Aglomeratif	5	9051.1178	0.1709	1.8674	2.7291
	DBSCAN	1.0, 5	62.7365	0.7131	0.4396	0.3524
4 (Selected Features Data + PCA)	K-Means	5	1279.1926	0.5323	0.7546	1.0847
	Hierarki Aglomeratif	5	1544.7770	0.4662	0.7848	1.0092
	DBSCAN	0.1, 10	26.0593	0.7918	0.2746	0.1131

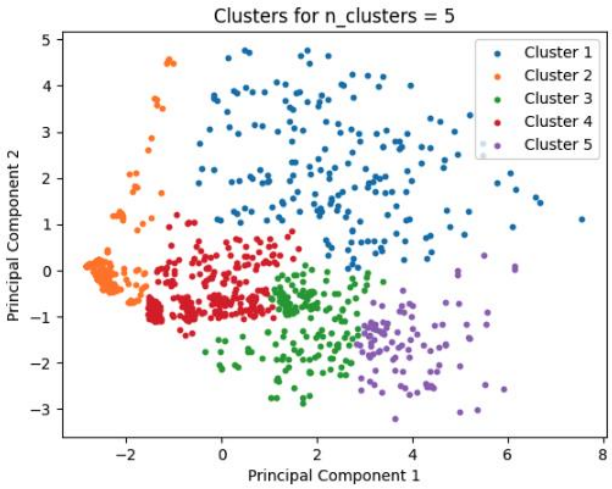


Rajah 4: Plot bersepah model pengelompokan K-Means uji kaji 2

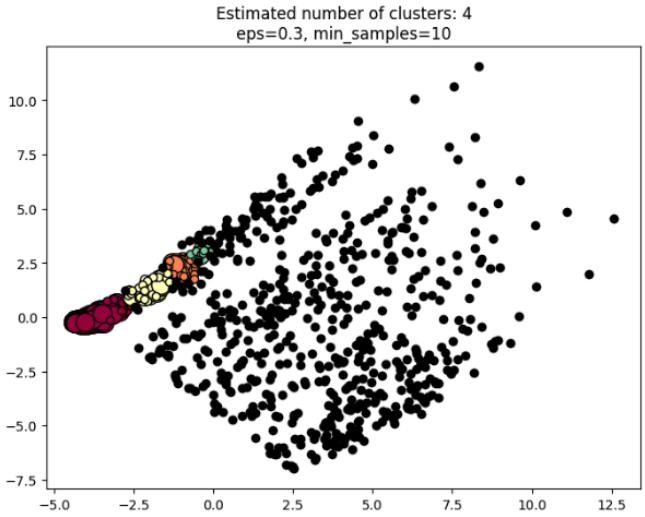


Rajah 5: Plot bersepah model pengelompokan K-Means uji kaji 4

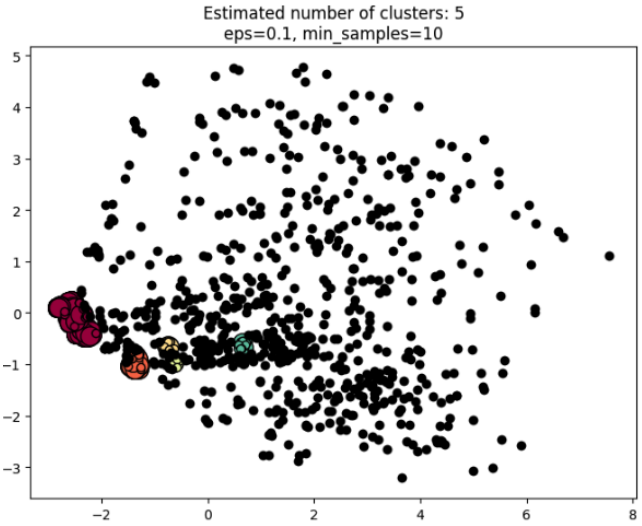
Rajah 6: Plot bersepah model pengelompokan Hierarki Aglomeratif uji kaji 2



Rajah 7: Plot bersepah model pengelompokan Hierarki Aglomeratif uji kaji 4



Rajah 8: Plot bersepah DBSCAN uji kaji 2



Rajah 9: Plot bersepah DBSCAN uji kaji 4

Analisis dan tafsiran hasil pengelompokan adalah salah satu aktiviti paling penting dalam pengelompokan. Setiap kelompok perlu diterokai dan dianalisis untuk memahami ciri-cirinya dan perbezaannya. Analisis untuk setiap kelompok akan dilaksanakan dengan mengkaji skor kepentingan ciri-ciri, saiz kelompok, boxplot ciri-ciri berdasarkan kelompok dan taburan ciri data dalam setiap kelompok. Dalam kajian ini, analisis dan tafsiran untuk setiap kelompok akan mengenal pasti faktor COVID-19 jangka panjang yang paling menonjol. Jadual 5 menunjukkan ciri-ciri yang paling berpengaruh dalam nombor kelompok terbaik k=5.

Jadual 5: Ciri-ciri berdasarkan skor kepentingan dalam k=5

No	Ciri	Skor Kepentingan
1	anxietydepression_AcuteCovid	0.141031
2	Last_fever	0.126516
3	Fatigue_now	0.120757
4	anxietydepression_now	0.106295

5	Skinrash_pastweek_unknown	0.080599
6	ageusia1_pastweek	0.079189
7	Headache_pastweek	0.077615
8	weaknessArmsLegs_now	0.073831
9	fever_Yes	0.054818
10	SOB1_pastweek	0.048347
11	fever_unknown	0.036283
12	income_0	0.030680
13	fever_No	0.024038

KESIMPULAN

Secara kesuluruhannya, pembangunan model pengelompokan berjaya mencapai matlamat perniagaan yang dinyatakan pada perancangan awal projek. Pemprosesan data pada peringkat awal terhadap set data *Long Covid* adalah untuk menjamin kualiti data dan memastikan set data tiada ralat yang akan menjejaskan keputusan model. Keputusan model pengelompokan untuk mengenal pasti faktor mempengaruhi COVID-19 jangka panjang dan model prediktif dihasilkan melalui perbandingan prestasi antara model-model dengan menggunakan metrik penilaian. Cadangan untuk menambahbaik kajian ini adalah mengumpul set data yang mempunyai lebih banyak variasi atribut dan rekod agar dapat meningkatkan kebolehan proses pembelajaran mesin. Sebagai contoh, pengumpulan data mengenai ciri-ciri klinikal ataupun jenis simptom dan faktor COVID-19 jangka panjang. Sistem ramalan menggunakan model pengelompokan K-Means boleh dibangunkan supaya pengguna dapat memasukkan data yang lain untuk meramalkan indikator atau simptom COVID-19 jangka panjang.

PENGHARGAAN

Pertama sekali, saya ingin menyampaikan jutaan terima kasih kepada penyelia saya, Prof. Dr. Azuraliza Binti Abu Bakar, atas nasihat, inspirasi, dan bimbingan yang tak ternilai. Ribuan terima kasih juga saya ucapkan kepada rakan-rakan seperjuangan atas bantuan dan kerjasama mereka dalam menyempurnakan projek ini. Penghargaan turut diberikan kepada Fakulti / Institusi / Pusat Pengajian / Jabatan atas kemudahan yang sempurna.

Ucapan terima kasih ini juga ditujukan kepada semua pihak yang terlibat secara langsung atau tidak langsung dalam menjayakan projek ini. Bantuan dan sokongan mereka amat saya hargai kerana tanpa mereka, tugasan ini mungkin tidak dapat dilaksanakan dengan baik.

RUJUKAN

Abdul Rahman, M., Sani, N. S., Hamdan, R., Ali Othman, Z. & Abu Bakar, A. 2021. A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. *PloS one* 16(8): e0255312.

Anon. 2023. Post COVID Conditions. *Centers for Disease Control and Prevention*

Belluck, P. 2023. Scientists offer a new explanation for long Covid. *The New York Times*

Daines, L., Mulholland, R. H., Vasileiou, E., Hammersley, V., Weatherill, D., Katikireddi, S. V., Kerr, S., Moore, E., Pesenti, E. & Quint, J. K. 2022. Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data. *BMJ open* 12(7): e059385.

Davenport, T. & Kalakota, R. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6(2): 94.

Davis, H. E., Mccorkell, L., Vogel, J. M. & Topol, E. J. 2023. Long COVID: major findings, mechanisms and recommendations. *Nature Reviews Microbiology* 21(3): 133-146.

Flam, D. S. 2023. Machine Learning in Healthcare: Guide to Applications & benefits. *ForeSee Medical*

Health, N. I. O. 2023. Long COVID Information and Resources *NIH COVID-19 Research*

Hicham, N. & Karim, S. 2022. Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. *International Journal of Advanced Computer Science and Applications* 13(10):

Katella, K. 2023. What happens when you still have long COVID symptoms? . *Yale Medicine*

Kherif, F. & Latypova, A. 2020. Principal component analysis. Dlm. (pnyt.). *Machine learning*, hlm. 209-225. Elsevier.

Müllner, D. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*

Pfaff, E. R., Girvin, A. T., Bennett, T. D., Bhatia, A., Brooks, I. M., Deer, R. R., Dekermanjian, J. P., Jolley, S. E., Kahn, M. G. & Kostka, K. 2022. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *The Lancet Digital Health* 4(7): e532-e541.

Rosenblatt, F. 1958. *The perceptron: a theory of statistical separability in cognitive systems (Project Para).* Cornell Aeronautical Laboratory.

Salihoun, M. 2020. State of art of data mining and learning analytics tools in higher education. *International Journal of Emerging Technologies in Learning (iJET)* 15(21): 58-76.

Seger, C. 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

Van Beek, P. 2005. *Principles and Practice of Constraint Programming-CP 2005.* Springer.

Zubair, M., Asif Iqbal, M., Shil, A., Haque, E., Moshiul Hoque, M. & Sarker, I. H. 2021. An efficient k-means clustering algorithm for analysing covid-19. *Hybrid Intelligent Systems: 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020*, hlm. 422-432.

Yee Xin Jie (A187328)

Prof. Dr. Azuraliza Abu Bakar

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia