PTA-FTSM-2023-A179637

RAMALAN KEPARAHAN PESAKIT COVID-19 DENGAN MENGGUNAKAN MODEL PEMBELAJARAN MESIN

Ooi Teng He[1*],
Zalinda Othman[2]

[1,2]*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia*

Commented [U1]: Untuk keperluan daftar ke dalam e-rep. No rujukan akan dimasukkan juga sebagai keperluan daftar ke dalam e-rep.

ABSTRAK

Penyakit berjangkit baharu yang dinamakan Coronavirus Sindrom Pernafasan Akut Teruk 2 (SARS-CoV-2) yang pada mulanya ditemui semasa wabak kes penyakit pernafasan di Wuhan, China, dikenali sebagai virus korona penyakit 2019 (COVID-19). Hal ini telah menyebabkan bumi kita mengalami pandemik akibat COVID-19. Malaysia juga merupakan antara negara yang teruk dilanda oleh COVID-19 dan terkesan teruk dari segi kesihatan, ekonomi, dan sosial. Oleh itu, kemajuan kecerdasan buatan adalah penting dalam mendiagnosis dan meramalkan penyebaran COVID-19 dan pendekatan pembelajaran mesin berpotensi meramal kesan penyebaran virus ini. Kajian ini bertujuan untuk membina model ramalan menggunakan beberapa algoritma pembelajaran mesin dan menilai prestasinya untuk mencari model terbaik bagi ramalan klasifikasi tahap keparahan pesakit COVID-19. Secara khususnya, terdapat lima jenis model ramalan telah digunakan dalam projek ini, iaitu regresi logistik, hutan rawak, mesin vektor sokongan, *Gaussian Naïve Bayes*, dan pohon keputusan. Kajian ini dilakukan dengan menggunakan 131788 rekod data pesakit COVID-19 di kawasan pemantauan Pejabat Kesihatan Daerah Hulu Langat, Selangor dan terdapat 13 atribut di dalam set data. Prestasi model akan dinilai dari segi ketepatan, *precision*, *recall*, skor F1 dan nilai AUC. Kajian ini mendapati bahawa model hutan rawak talah mencatatkan ketepatan, *precision*, *recall*, skor F1 dan nilai AUC yang paling tinggi iaitu sebanyak 0.91 dan model ini akan digunakan untuk melakukan ramalan keparahan pesakit COVID-19. Keputusan ramalan tahap keparahan pesakit COVID-19 ditunjukkan di aplikasi web dengan menggunakan *Streamlit*. Akhirnya, model yang

diterangkan boleh dipercayai dan boleh digunakan untuk meramal tahap keparahan pesakit COVID-19 dengan cara yang cepat dan berkesan dalam konteks semasa pandemik COVID-19.

Kata Kunci: COVID-19, Ramalan tahap keparahan, Kecerdasan buatan, Pembelajaran mesin, *Streamlit*

Pengenalan

Sejarah virus korona bermula pada tahun 1930-an apabila jangkitan virus bronkitis berjangkit (IBV) disebabkan oleh ayam peliharaan. Pada manusia, kesnya pertama kali dilaporkan pada tahun 1960-an (Kahn & McIntosh 2005). Beberapa virus korona diketahui yang menyebabkan jangkitan pernafasan daripada selsema hingga penyakit yang lebih teruk seperti Sindrom Pernafasan Timur Tengah (MERS) dan Sindrom Pernafasan Akut Teruk (SARS). Pada Disember 2019, penyakit pernafasan berjangkit baharu muncul di Wuhan, wilayah Hubei, China, dan dinamakan oleh Pertubuhan Kesihatan Sedunia (WHO) sebagai COVID-19, penyakit virus korona 2019 yang sangat berjangkit kepada sistem pernafasan manusia, sistem hepatik dan sistem gastrousus, dan gangguan neurologi (Verma & Prakash 2020). COVID-19 merupakan virus asid ribonukleik (RNA) dan ahli keluarga virus korona yang tinggal dalam mamalia dan burung (Su et al. 2016; Zhu et al. 2020). Virus ini boleh merebak antara manusia, ternakan, dan haiwan liar, seperti burung, kelawar, dan tikus. SARS-CoV-2 boleh menyebabkan penyakit pernafasan akut yang teruk yang mengakibatkan kematian dalam pelbagai kes. Simptom COVID-19 adalah batuk, demam, hidung tersumbat, sesak nafas, dan kadang-kadang cirit-birit (Kementerian Kesihatan Malaysia 2020).

WHO telah mengisytiharkan virus korona COVID-19 sebagai kecemasan kesihatan awam dengan potensi pandemik pada 30 Januari 2020. Hal ini demikian kerana virus ini menyebabkan sejumlah besar kes dijangkiti dan kadar kematian yang tinggi. Sehingga

November 2022, 230 negara dengan lebih 650 juta kes positif yang telah dilaporkan oleh WHO, dengan kadar kematian sebanyak 1% daripada jumlah keseluruhan kes tertutup disebabkan oleh penularan manusia ke manusia yang pantas ("COVID Live - Coronavirus Statistics by Country-Worldometer" 2022). Seperti negara lain, Malaysia juga merupakan antara negara yang paling teruk dilanda oleh COVID-19 dengan 4 juta kes disahkan COVID-19 menurut laporan analisis yang diterbitkan di laman web COVID-19 Kementerian Kesihatan Malaysia. Sejak permulaan pandemik, semua kes positif COVID-19 telah dimasukkan ke hospital tanpa mengira tahap keparahan penyakit itu. Tambahan pula, peningkatan drastik dalam kes di seluruh dunia telah menyebabkan kapasiti hospital mencapai 100%, meletakkan beban berat kepada kemudahan perubatan (Pourhomayoun & Shakibi 2021). Kadar kematian meningkat dengan ketara akibat penghantaran oksigen yang tidak mencukupi, terutamanya dalam pesakit COVID-19. Oleh itu, kebanyakan negara di seluruh dunia mengehadkan interaksi sosial melalui langkah berjaga-jaga dan mengambil penjagaan responsif seperti mencuci tangan, memakai topeng muka, penjarakan fizikal, dan mengelakkan perhimpunan besar-besaran. Strategi penguncian sementara dan tinggal di rumah telah dilaksanakan sebagai tindakan yang diperlukan untuk mengawal penularan penyakit (Pokhrel & Chhetri 2021). Walaupun pihak berkuasa telah menggunakan beberapa strategi untuk mencegah penularan COVID-19, tetapi penyakit ini masih menular di seluruh dunia dengan tahap risiko tinggi kerana orang ramai tidak mematuhi prosedur operasi standard (SOP) yang disyorkan kerajaan. Akibatnya, kekurangan katil ICU dan sumber perubatan lain telah menyukarkan pihak berkuasa untuk memperuntukkan sumber kepada pesakit kritikal. Oleh itu, penggunaan teknik yang membolehkan pengecaman pantas pesakit berisiko tinggi untuk bentuk teruk dan tidak teruk untuk kemasukan ke hospital keutamaan adalah kritikal (Jiang et al. 2020). Penggunaan teknik pembelajaran mesin (ML) yang merupakan salah satu subbidang kecerdasan buatan (AI) adalah penting dalam mengesan corak dan meramalkan tahap keparahan pesakit COVID-19.

Beberapa kajian penyelidikan telah dilakukan mengenai ramalan keparahan COVID-19 menggunakan model pembelajaran mesin. Salah satunya adalah kajian oleh Aljameel et al. yang menggunakan data klinikal dan demografi pesakit dari Hospital Universiti *King Fahad*, *Dammam*, Kerajaan Arab Saudi untuk meramalkan keparahan penyakit dalam pesakit COVID-19. Keputusan menunjukkan bahawa hutan rawak mengatasi pengelas lain dengan ketepatan 0.95 dan AUC 0.99 (Aljameel et al. 2021). Kajian lain oleh Gull et al. di Pakistan menggunakan algoritma pembelajaran mesin untuk meramalkan keparahan pesakit COVID-19 dengan tujuan mengatasi kekurangan sumber. Hasilnya menunjukkan mesin vektor sokongan sebagai model terbaik kerana model tersebut telah menunjukkan 60% ketepatan yang paling tinggi, dan telah membahagikan keparahan pesakit kepada tahap ringan, sederhana dan teruk (Gull et al. 2020). Selain itu, Buvana dan Muthumayil juga menggunakan pembelajaran mesin untuk memilih daripada koleksi ciri berdimensi tinggi yang merupakan ciri terpenting dan pada masa yang sama meningkatkan kecekapan pengelas dengan masa pengiraan yang berkurangan. Mereka mendapati bahawa pengelas pohon keputusan memberikan hasil yang mengagumkan dengan mempunyai 94% *precision*, 96% *recall*, 95% skor f1, dan ketepatan 96.2% (Buvana & Muthumayil 2021). Di samping itu, kajian oleh Rustam et al. juga menggunakan pembelajaran mesin dan menunjukkan bahawa *exponential smoothing* (ES) adalah model terbaik untuk meramalkan faktor ancaman dari COVID-19 (Rustam et al. 2020). Terakhir, Jahangirimehr et al. menerapkan model algoritma berdasarkan pembelajaran mesin menggunakan data klinikal dan paraklinikal pesakit untuk penilaian keparahan COVID-19. Hasilnya menunjukkan bahawa mesin vektor sokongan menghasilkan hasil terbaik dengan mempunyai precision 95.5%, recall 94%, skor F1 94.8%, ketepatan 95%, dan AUC sebanyak 94% (Jahangirimehr et al. 2022).

Dalam kajian ini, tujuan penyelidikan ini adalah untuk membangunkan model ramalan menggunakan algoritma pembelajaran mesin yang berbeza dan seterusnya memilih model yang terbaik untuk menggunakan dalam meramal keparahan pesakit COVID-19. Selain itu, kajian

ini juga akan membangunkan aplikasi web pada sistem ramalan dengan menggunakan rangka kerja aplikasi sumber terbuka, iaitu *Streamlit* untuk menganalisis ciri-ciri pesakit dalam mengenal pasti tahap jangkitan pesakit. Hal ini akan membantu pihak berkuasa bukan sahaja memutuskan laluan klinikal untuk pengiktirafan kes kritikal tetapi juga membantu mereka menguruskan sumber mereka mengikut keutamaan kesihatan pesakit. Hasilnya, sistem yang dibangunkan boleh membantu hospital dan kemudahan perubatan memutuskan siapa yang perlu mendapat perhatian terlebih dahulu, siapa yang mempunyai keutamaan yang lebih tinggi untuk dimasukkan ke hospital, dan menghapuskan kelewatan dalam menyediakan penjagaan yang diperlukan. Oleh itu, kajian penyelidikan ini amat penting dan ia akan membantu orang ramai dalam sektor kesihatan untuk bersiap sedia dan mengambil semua langkah berjaga-jaga yang diperlukan untuk meminimumkan penyebaran wabak. Model ini juga akan berfungsi sebagai sistem amaran awal untuk mengenal pasti pesakit berisiko tepat pada masanya.

Skop kajian ini akan memfokuskan kepada ramalan tahap keparahan pesakit COVID-19, dan data terhad kepada pesakit COVID-19 di kawasan pemantauan Pejabat Kesihatan Daerah Hulu Langat, Selangor. Data akan meliputi tempoh 1 Mei 2021 hingga 31 Disember 2021. Selain itu, algoritma pembelajaran mesin yang akan digunakan untuk membangunkan model ramalan adalah regresi logistik, mesin vektor sokongan, Gaussian Naïve Bayes, hutan rawak dan pohon keputusan. Hasil ramalan tahap keparahan pesakit COVID-19 akan ditunjukkan dalam papan pemuka dengan menggunakan aplikasi web *Streamlit*.

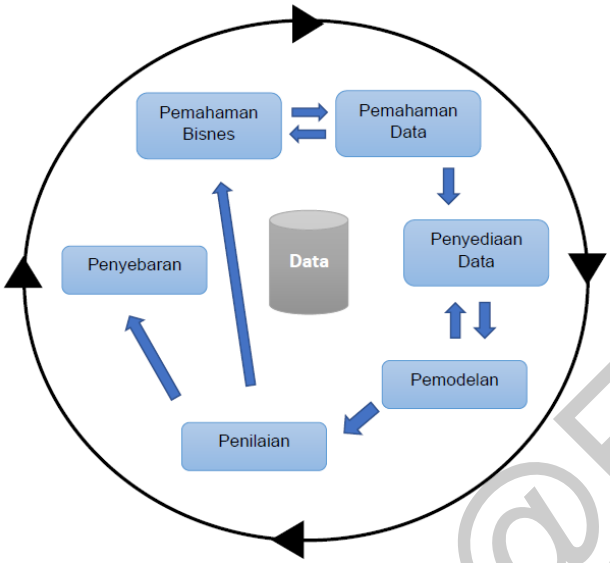
Beberapa kekangan berlaku sepanjang keseluruhan kajian ini. Pertama sekali, ingatan capaian rawak (RAM) komputer adalah salah satu kekangan dalam projek ini. Hal ini demikian kerana ingatan capaian rawak mestilah mencukupi supaya segala proses dapat berjalan dengan lancar dan menghalang proses daripada putus. Seterusnya, masa juga merupakan satu lagi kekangan untuk projek ini. Hal ini bukan disebabkan oleh tempoh yang diberikan untuk menyelesaikan projek, tetapi terutamanya disebabkan oleh saiz pengumpulan data yang besar,

model pembelajaran mesin perlu mengambil banyak masa untuk membuat latihan. Selain itu, kekurangan pengalaman dalam membangunkan papan pemuka pada sistem ramalan menggunakan aplikasi web *Streamlit* yang menyepadukan model pembelajaran mesin juga merupakan salah satu cabaran bagi saya.

Sebagai kesinambungan daripada itu, laporan ini disusun dalam tiga bahagian: metodologi kajian, keputusan dan perbincangan, dan kesimpulan. Metodologi kajian akan menjelaskan tentang kaedah dan pendekatan yang digunakan dalam menjalankan kajian. Ia juga menerangkan model proses pembangunan khusus yang digunakan serta jelaskan mengapa model proses berkenaan dipilih. Manakala keputusan dan perbincangan akan memaparkan hasil kajian dan maklumat yang diperoleh serta memberi makna dan kesimpulan kepada kajian yang telah dijalankan. Bahagian kesimpulan akan memberi gambaran terhadap hasil dan maklumat yang diperoleh dari kajian. Ia juga menerangkan apakah kekuatan serta kelemahannya untuk kerja-kerja akan datang.

Metodologi Kajian

Metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) telah digunakan dalam projek ini untuk membangunkan model ramalan yang dapat meramalkan keparahan pesakit COVID-19. Metodologi ini merupakan salah satu model proses perlombongan data yang menerangkan pendekatan yang biasa digunakan yang digunakan oleh pakar untuk menangani masalah. Metodologi ini mengandungi 6 fasa untuk membina projek secara teratur dan fasa-fasa ini boleh berkitar berdasarkan keperluan pembangun, iaitu fasa pemahaman bisnes, fasa pemahaman data, fasa penyediaan data, fasa pemodelan, fasa penilaian dan fasa penyebaran (Luna 2021). Rajah 1 menunjukkan urutan fasa metodologi CRISP-DM.



Rajah 1 Fasa-fasa Metodologi CRISP-DM

Fasa pemahaman perniagaan (*Business Understanding*) merupakan fasa pertama yang memfokuskan pada pemahaman objektif dan keperluan projek dari perspektif perniagaan. Matlamat utama dalam projek ini adalah untuk membangunkan model pembelajaran mesin yang boleh meramalkan keparahan kesihatan pesakit COVID-19 serta membangunkan sistem ramalan melalui aplikasi web *Streamlit* berdasarkan penemuan yang telah dikaji. Pengumpulan pelbagai kertas penyelidikan, penerbitan, jurnal dan buletin mengenai pandemik COVID-19 dan perlombongan data sangat diperlukan sebelum memulakan projek ini. Hal ini amat berguna dalam proses pembangunan model projek ini kerana pengetahuan yang berkaitan boleh digunakan sebagai panduan. Kriteria yang akan digunakan untuk menilai sama ada projek ini berjaya dari sudut perniagaan mesti ditakrifkan pada fasa ini.

Pemahaman Data (*Data Understanding*) merupakan fasa kedua yang bermula dengan pengumpulan data awal dan seterusnya memahami jenis dan bentuk data untuk mengenal pasti masalah yang terlibat dalam data yang diperolehi. Dalam kajian ini, set data diperoleh daripada data pesakit COVID-19 di kawasan pemantauan Pejabat Kesihatan Daerah Hulu Langat,

Selangor. Sebanyak 131788 rekod data pesakit COVID-19 dalam tempoh dari 1 Mei 2021 hingga 31 Disember 2021 telah diekstrak untuk projek penyelidikan ini. Sebanyak 13 atribut rekod pesakit COVID-19 akan digunakan dalam pembangunan model dalam projek ini, termasuklah umur, jantina, sakit tekak, demam, selesema, hilang deria rasa, hilang deria bau, batuk, sesak nafas, cirit birit, lain-lain dan komorbid akan digunakan sebagai pemboleh ubah tak bersandar manakala hasil diagnosis (ringan, sederhana dan serius) digunakan sebagai pemboleh ubah sasaran. Selain itu, taburan label kelas, kecenderungan taburan dan visualisasi data akan dihasilkan untuk lebih memahami data yang digunakan dalam kajian ini. Korelasi antara atribut data juga diperhatikan untuk memperoleh pengetahuan yang lebih baik tentang hubungan antara pemboleh ubah yang akan digunakan.

Penyediaan data (*Data Preparation*) ialah proses membersihkan dan mengubah data mentah sebelum pemprosesan dan analisis. Fasa ini merupakan langkah yang penting sebelum pemprosesan dan selalunya melibatkan pemformatan semula data, membuat pembetulan pada data dan menggabungkan set data untuk memperkayakan data supaya set data tersebut dapat sesuai untuk penyepaduan ke dalam model pembelajaran mesin. Dalam kajian ini, data yang diperoleh mesti melalui langkah pemprosesan data termasuklah pergabungan data, penerokaan data, pembersihan data, dan pendiskretan data. Peringkat awal dalam pemprosesan data adalah untuk menggabungkan data daripada set data yang berlainan. Rekod data pesakit COVID-19 dari tempoh 1 Mei 2021 hingga 31 Disember 2021 daripada set data pesakit COVID-19 akan diekstrak dan digabungkan ke set data yang baru. Pada masa yang sama, pemilihan atribut juga dilakukan dengan mengikut kajian kesusasteraan supaya dapat menghasilkan keputusan ramalan yang berkualiti. Seterusnya, penerokaan data seperti penyemakan data, dimensi data, jenis data, taburan label kelas, kecenderungan taburan, visualisasi data dan korelasi antara atribut telah dilaksanakan untuk mendapatkan pemahaman yang lebih baik tentang data yang tersedia sebelum melakukan analisis yang lebih lanjut atau membangun model ramalan. Proses

pembersihan data membersihkan data dengan mengisi nilai yang hilang, pelicinan data hingar (noisy data), mengenal pasti atau mengalih keluar elemen luaran, *one-hot encoding* untuk mengubah nilai kategori kepada nilai berangka dan menyelesaikan masalah data yang tidak konsisten. Pendiskretan data membantu mengubah format data agar sesuai dengan model pembelajaran mesin. Ketepatan model boleh dijamin dengan menggunakan set data yang telah melalui proses pra-pemprosesan.

Fasa pemodelan (*Modelling*) ialah proses di mana teknik pembelajaran mesin digunakan pada set data bersih dalam model ramalan untuk keparahan pesakit Covid-19. Dalam kajian ini, algoritma pembelajaran mesin yang digunakan untuk meramalkan tahap keparahan pesakit COVID-19 adalah regresi logistik, mesin vektor sokongan, *Gaussian Naïve Bayes*, hutan rawak dan pohon keputusan. Setiap algoritma akan dibangunkan menggunakan model latihan dan diuji dengan menggunakan data ujian. Semua algoritma akan dibangunkan menggunakan parameter lalai yang terdapat dalam dokumentasi *sklearn* ("scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation" n.d.). Set data akan diasingkan kepada 80 peratus untuk latihan dan 20 peratus untuk ujian sebelum membina model ramalan. Hal ini kerana set ujian digunakan untuk mengesahkan prestasi model ramalan manakala set latihan dikendalikan dalam pemodelan. Model dengan hasil pemodelan terbaik akan dipilih untuk model ramalan kajian ini.

Fasa penilaian (*Evaluation*) merupakan fasa yang melakukan penilaian terhadap model pembelajaran mesin dengan teliti dan menyemak langkah-langkah yang dilaksanakan untuk membina model sebelum meneruskan ke pelaksanaan terakhir model. Penilaian model ialah bahagian teras membina model pembelajaran mesin yang berkesan. Dalam kajian ini, prestasi setiap model pembelajaran dinilai dari segi ketepatan, *precision*, *recall*, skor F1 dan nilai AUC (*Area under curve*). Penilaian dan pengujian pada model ramalan akan dijalankan ke atas set data input pesakit COVID-19 dari tahun 2022. Pengujian *t-test* juga digunakan sebagai ujian

hipotesis untuk membandingkan prestasi model ramalan dan membuktikan model yang terpilih mempunyai prestasi yang tinggi dalam ramalan keparahan pesakit COVID-19.

Fasa pelaksanaan (*Deployment*) merupakan fasa terakhir proses pembangunan model ramalan yang membentangkan hasil kajian dengan cara yang berguna dan boleh difahami, dan projek harus mencapai matlamatnya. Dalam kajian ini, hasil ramalan keparahan pesakit COVID-19 akan dipaparkan melalui aplikasi web *Streamlit*. Alat visualisasi ini menyediakan papan pemuka pengguna yang jelas untuk membolehkan pengguna memahami hasil analisis dengan cepat dan mudah. Pemantauan dan penyelenggaraan mesti dilakukan secara berkala untuk menjamin bahawa tiada ralat berkembang dalam keputusan kajian dan maklumat yang tidak betul tidak disebarkan kepada orang ramai.

Keputusan dan Perbincangan

Penilaian metrik dijalankan terhadap algoritma pembelajaran mesin yang telah dibangunkan untuk menentukan algoritma yang paling sesuai untuk mengklasifikasikan keparahan pesakit COVID-19. Metrik penilaian yang digunakan untuk menilai prestasi model ramalan adalah ketepatan, *precision*, *recall*, skor F1 dan nilai AUC. Semakin tinggi skor penilaian bermakna prestasi model ramalan tersebut adalah lebih tinggi daripada yang lain. Bahagian ini mempamerkan keputusan prestasi model yang menggunakan set data pesakit COVID-19. Dalam set data ini mendapati bahawa masalah ketidakseimbangan kelas label yang akan menjejaskan ketepatan pemodelan. Oleh itu, pelbagai kaedah pensampelan semula seperti SMOTE (*oversampling*) dan NearMiss-3 (*undersampling*) telah digunakan untuk mengatasi masalah ini. Jadual 1, 2, 3, 4 dan 5 menunjukkan keputusan pemodelan berdasarkan pelbagai kaedah pensampelan mengikut model masing-masing.

Jadual 1 Keputusan klasifikasi pohon keputusan dalam pelbagai kaedah pensampelan semula

Metrik Penilaian		**ketepatan**	***precision***	***recall***	***f1-score***
Tidak mengguna sebarang kaedah pensampelan semula	*Train*	0.91	0.91	0.91	0.91
	Test	0.91	0.91	0.91	0.91
Pensampelan semula dengan SMOTE (*Oversampling*)	*Train*	0.87	0.87	0.87	0.87
	Test	0.87	0.87	0.87	0.87
Pensampelan semula dengan *Near Miss* (*Undersampling*)	*Train*	0.70	0.70	0.70	0.70
	Test	0.47	0.46	0.47	0.46

Jadual 2 Keputusan klasifikasi *Gaussian Naive Bayes* dalam pelbagai kaedah pensampelan semula

Metrik Penilaian		**ketepatan**	***precision***	***recall***	***f1-score***
Tidak mengguna sebarang kaedah pensampelan semula	*Train*	0.90	0.91	0.90	0.90
	Test	0.90	0.91	0.90	0.90
Pensampelan semula dengan SMOTE (*Oversampling*)	*Train*	0.79	0.79	0.79	0.79
	Test	0.80	0. 80	0.80	0.79
Pensampelan semula dengan *Near Miss* (*Undersampling*)	*Train*	0.54	0.54	0.54	0.53
	Test	0.50	0.50	0.50	0.48

Jadual 3 Keputusan klasifikasi regresi logistik dalam pelbagai kaedah pensampelan semula

Metrik Penilaian		**ketepatan**	***precision***	***recall***	***f1-score***
Tidak mengguna sebarang kaedah pensampelan semula	*Train*	0.91	0.91	0.91	0.91
	Test	0.91	0.91	0.91	0.91
Pensampelan semula dengan SMOTE (*Oversampling*)	*Train*	0.80	0.80	0.80	0.79
	Test	0.80	0.80	0.80	0.80
Pensampelan semula dengan *Near Miss* (*Undersampling*)	*Train*	0.56	0.54	0.56	0.53
	Test	0.50	0.48	0.50	0.47

Jadual 4 Keputusan klasifikasi hutan rawak dalam pelbagai kaedah pensampelan semula

Metrik Penilaian		**ketepatan**	***precision***	***recall***	***f1-score***
Tidak mengguna sebarang kaedah pensampelan semula	*Train*	0.91	0.91	0.91	0.91
	Test	0.91	0.91	0.91	0.91
Pensampelan semula dengan SMOTE (*Oversampling*)	*Train*	0.87	0.87	0.87	0.87
	Test	0.87	0.87	0.87	0.87
Pensampelan semula dengan *Near Miss* (*Undersampling*)	*Train*	0.70	0.70	0.70	0.69
	Test	0.46	0.44	0.46	0.44

Jadual 5 Keputusan klasifikasi mesin vektor sokongan dalam pelbagai kaedah pensampelan semula

Metrik Penilaian		**ketepatan**	***precision***	***recall***	***f1-score***
Tidak mengguna sebarang kaedah pensampelan semula	*Train*	0.91	0.91	0.91	0.91
	Test	0.91	0.91	0.91	0.91
Pensampelan semula dengan SMOTE (*Oversampling*)	*Train*	0.83	0.84	0.83	0.83
	Test	0.84	0.84	0.84	0.83
Pensampelan semula dengan *Near Miss* (*Undersampling*)	*Train*	0.58	0.59	0.58	0.53
	Test	0.52	0.48	0.52	0.47

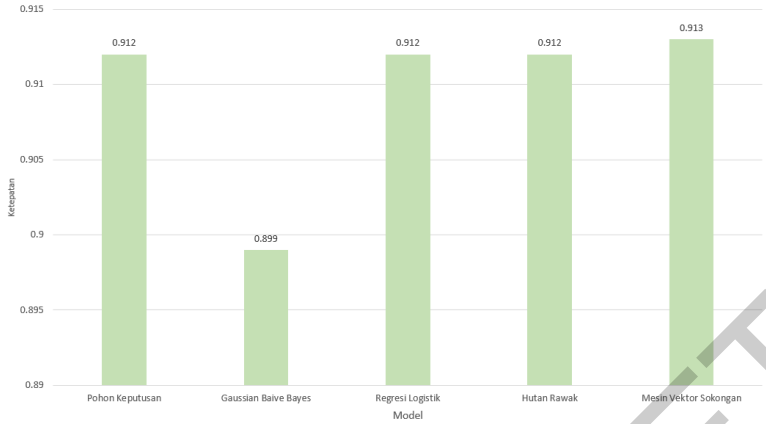
Selepas perbandingan penilaian pemodelan berdasarkan pelbagai kaedah pensampelan semula dalam semua model, kita dapati bahawa semua model yang tidak mengguna sebarang kaedah pensampelan semula telah memperoleh keputusan terbaik berbanding dengan model yang mengguna kaedah pensampelan semula. Oleh itu, perbandingan penilaian metrik antara model yang tidak mengguna sebarang kaedah pensampelan semula akan dilaksanakan. Penilaian metrik set data ujian akan dipilih untuk membuat perbandingan prestasi ramalan

antara model kerana set data latihan merupakan elemen yang penting untuk membina model ramalan, manakala set data ujian adalah untuk mengkaji dan mengesahkan prestasi model yang dibina. Penilaian algoritma klasifikasi akan dijalankan perbandingan untuk menentukan model ramalan keparahan pesakit COVID-19 yang dapat membuat ramalan dengan prestasi yang tinggi. Jadual 6 telah meringkaskan semua penilaian metrik model ramalan.

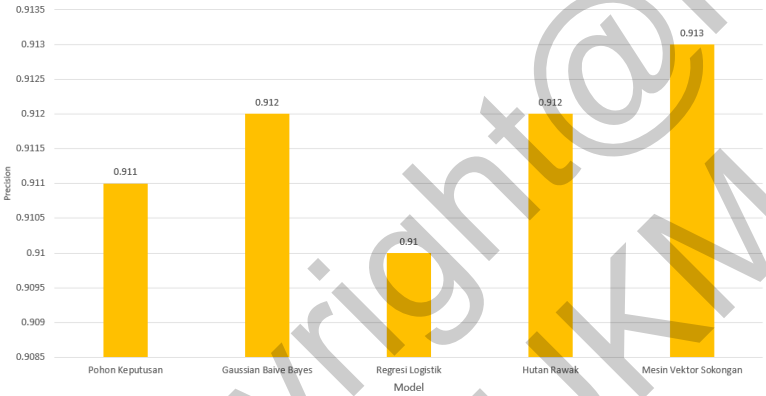
Jadual 6 Penilaian metrik model ramalan

Metrik Penilaian	**ketepatan**	***precision***	***recall***	**skor f1**	**nilai AUC**
Pohon Keputusan	0.912	0.911	0.912	0.910	0.928
Gaussian Naive Bayes	0.899	0.912	0.899	0.903	0.903
Regresi Logistik	0.912	0.910	0.912	0.910	0.914
Hutan Rawak	0.912	0.912	0.912	0.911	0.931
Mesin Vektor Sokongan	0.913	0.913	0.913	0.912	0.897

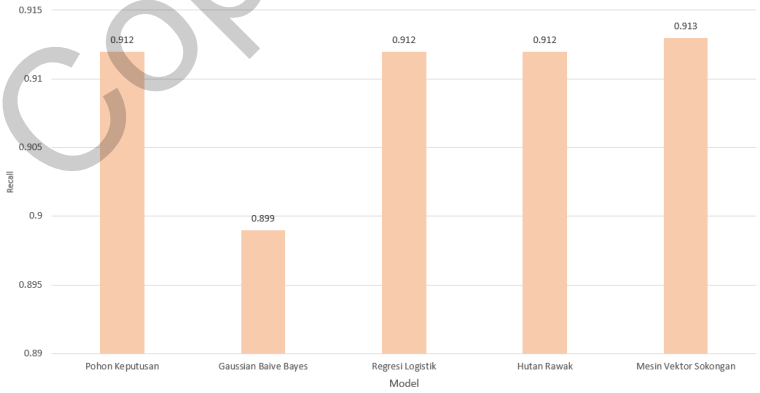
Daripada keputusan yang diperoleh, semua model telah mencapai semua metrik penilaian yang tinggi iaitu sebanyak 0.90 dan ke atas dalam membuat ramalan keparahan pesakit COVID-19. Metrik penilaian ketepatan, *recall* dan skor f1 bagi pohon keputusan, regresi logistik, hutan rawak dan mesin vektor sokongan adalah sama tinggi dengan nilai sebanyak 0.91. Dari segi *precision*, semua model telah mencapai *precision* yang sama tinggi iaitu nilai sebanyak 0.91 juga. Model hutan rawak telah memperoleh nilai AUC yang lebih tinggi sedikit, iaitu 0.931. Selepas perbandingan dijalankan di antara model ramalan, semua model mempunyai penilaian metrik yang lebih kurang sama. Oleh itu, penilaian dan pengujian pada model ramalan akan dilaksanakan untuk membuktikan bahawa mana model yang mempunyai metrik penilaian paling tinggi dalam ramalan keparahan pesakit COVID-19. Rajah 2, 3, 4, 5 dan 6 telah menunjukkan graf perbandingan prestasi model ramalan.



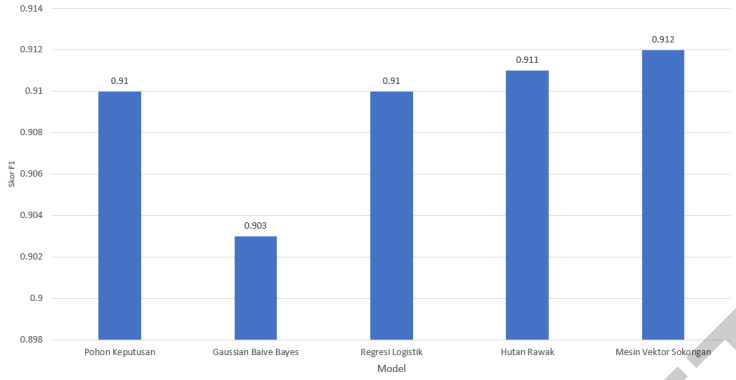
Rajah 2 Perbandingan ketepatan antara model



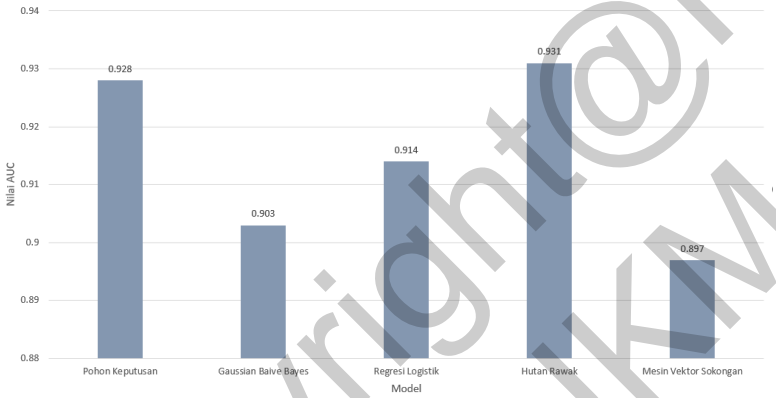
Rajah 3 Perbandingan *precision* antara model



Rajah 4 Perbandingan *recall* antara model



Rajah 5 Perbandingan skor F1 antara model



Rajah 6 Perbandingan nilai AUC antara model

Penilaian dan pengujian pada model ramalan akan dijalankan ke atas set data input pesakit COVID-19 dari tahun 2022. Hal ini kerana keputusan eksperimen yang muktamad memerlukan berbilang kali perjalanan proses pengujian terhadap set data input yang berbeza. Justeru, terdapat 10 bilangan set data pesakit COVID-19 yang baru akan digunakan dalam eksperimen untuk melakukan pengujian dan pengesahan untuk memilih model yang terbaik. Jadual 7 mempamerkan set data input yang akan diaplikasikan dalam eksperimen.

Jadual 7 Set data input eksperimen

kes ujian	kategori umur	jantina	sakit tekak	demam	selesema	hilang deria rasa	hilang deria bau	batuk	sesak nafas	cirit birit	lain-lain	komorbid	hasil ramalan yang dijangkakan
1	5	1	0	1	0	0	0	1	0	0	0	1	2
2	2	0	1	1	0	0	0	0	0	0	0	1	2
3	4	1	0	0	0	0	0	0	0	0	0	0	1
4	4	1	0	1	0	1	1	1	0	0	0	0	2
5	4	1	0	1	0	1	1	0	1	1	1	0	3
6	4	0	0	0	0	0	0	0	0	0	0	1	1
7	3	1	0	1	0	0	0	0	0	0	0	0	2
8	2	0	0	1	0	0	0	1	1	0	0	0	3
9	1	1	1	0	0	0	0	0	0	0	0	0	1
10	5	0	0	1	0	0	0	1	1	0	0	0	3

Selepas penilaian dan ujian dijalankan dalam semua model ramalan, kita dapati bahawa model pohon keputusan dan model hutan rawak adalah model yang dapat meramalkan hampir semua kes betul. Oleh itu, kedua-dua model adalah model yang paling sesuai digunakan untuk melakukan ramalan keparahan pesakit COVID-19. Walau bagaimanapun, ujian-t antara pohon keputusan dan hutan rawak akan dilaksanakan untuk memilih model terbaik.

Ujian t ialah ujian statistik yang digunakan untuk membandingkan min dua kumpulan. Ia sering digunakan dalam ujian hipotesis untuk menentukan sama ada proses atau rawatan sebenarnya mempunyai kesan ke atas populasi yang diminati, atau sama ada dua kumpulan berbeza antara satu sama lain. Dalam kajian projek ini, ujian t digunakan sebagai ujian hipotesis untuk membandingkan prestasi model ramalan dan membuktikan model yang terpilih mempunyai prestasi yang tinggi dalam ramalan keparahan pesakit COVID-19. Aras signifikan yang digunakan adalah *alpha* dengan nilai 0.05 (95% sela keyakinan). Jadual 8 menunjukkan pengujian *t-test* bagi model pohon keputusan dan hutan rawak.

Jadual 8 Pengujian *T-test* (Pohon keputusan dan hutan rawak)

T-test	**Perincian**
Hipotesis Nol (H0)	Prestasi pohon keputusan kurang tinggi daripada hutan rawak dan tiada perbezaan min yang ketara antara satu sama lain. (DT <= RF)
Hipotesis Alternatif (H1)	Prestasi pohon keputusan adalah lebih tinggi daripada hutan rawak dan terdapat perbezaan min yang ketara antara satu sama lain. (DT > RF)
Statistik Ujian T	-1.825742
Nilai-p	0.070964
Kesimpulan	Oleh kerana nilai-p(=0.070964) > alpha(=0.05), hipotesis H0 diterima. DT <= RF

Hipotesis sifar menyatakan prestasi pohon keputusan tidak tinggi daripada hutan rawak manakala hipotesis alternatif menyatakan prestasi pohon keputusan lebih tinggi daripada hutan rawak. Nilai T-statistik adalah -1.825742 dan nilai-p adalah 0.070964. Keputusan menunjukkan nilai-p adalah lebih tinggi daripada aras signifikan (0.05). Oleh itu, kita dapat membuktikan bahawa prestasi hutan rawak adalah lebih tinggi daripada pohon keputusan pada 95% sela keyakinan. Model hutan rawak akan digunakan untuk melakukan ramalan keparahan pesakit COVID-19.

Seterusnya, perbandingan dengan kajian lepas telah dilaksanakan. Jadual 9 menunjukkan prestasi model ramalan kajian ini bersama dengan hasil keputusan penilaian model ramalan yang telah dihasilkan dalam kajian lepas. Bahagian ini akan membandingkan prestasi model ramalan kajian ini dengan kajian Gull (Gull et al. 2020) kerana kedua-dua kajian mempunyai persamaan dalam metodologi kajian.

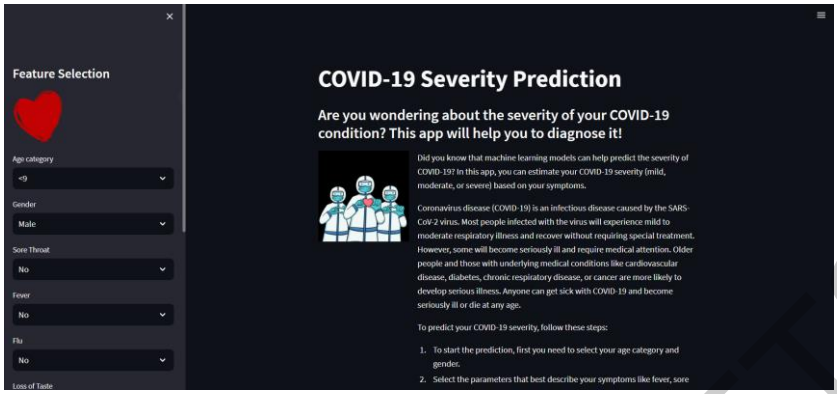
Jadual 9 Perbandingan prestasi model ramalan secara umum dengan kajian lepas

Metrik Penilaian	**Kajian ketepatan**	**Kajian Gull et al. (2020) ketepatan**
Pohon Keputusan	0.912	0.501
Gaussian Naive Bayes	0.899	0.508
Regresi Logistik	0.912	0.547
Hutan Rawak	0.913	0.511

Mesin Vektor Sokongan	0.914	0.603
Analisis diskriminan linear	X	0.566
K Nearest Neighbor	X	0.501

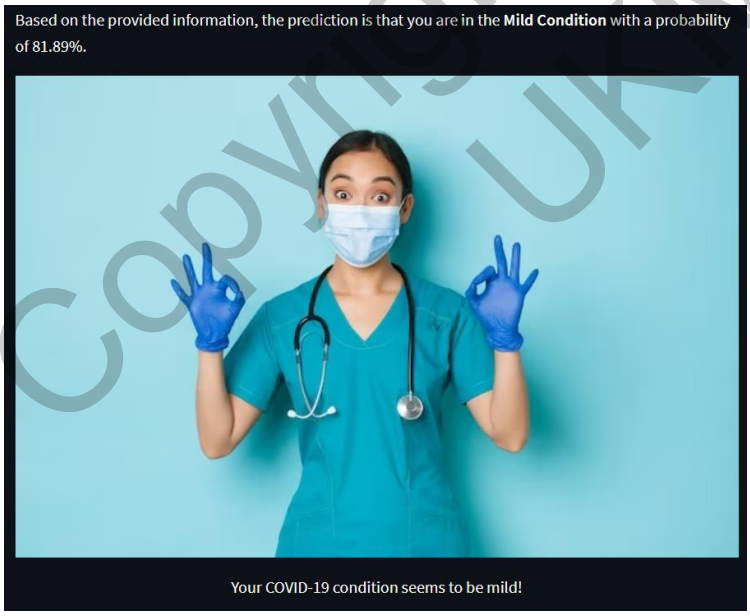
Berdasarkan jadual 9, kajian ini tidak mengaplikasikan model analisis diskriminan linear dan model *K-Nearest Neighbor*. Justeru, tiada keputusan mengenai model analisis diskriminan linear dan model *K-Nearest Neighbor* dipaparkan dalam jadual 9 dan digantikan dengan simbol 'X'. Kajian Gull et al. (2020) mendapati bahawa mesin vektor sokongan merupakan antara model ramalan yang mencapai prestasi yang paling tinggi dalam ramalan keparahan pesakit COVID-19, iaitu model tersebut telah mencapai ketepatan sebanyak 60.3%. Kajian ini pula mendapati bahawa model hutan rawak berprestasi tinggi dalam ramalan keparahan pesakit COVID-19 selepas penggunaan ujian hipotesis *T-test*. Secara keseluruhannya, prestasi semua model dalam kajian ini adalah tinggi berbanding dengan kajian Gull et al. (2020) kecuali model analisis diskriminan linear dan model *K-Nearest Neighbor*. Hal ini demikian kerana bilangan set data yang digunakan oleh kajian Gull et al. (2020) adalah kecil iaitu sebanyak jumlah 992 rekod pesakit. Kajian ini menggunakan jumlah set data yang besar iaitu mengandungi 131788 rekod pesakit. Hutan rawak dipilih sebagai model yang berprestasi tinggi kerana hutan rawak merupakan model *ensemble* yang menggabungkan pokok keputusan untuk menghasilkan model yang cekap dan mengendali masalah *overfitting* dengan baik berbanding dengan model ramalan yang lain.

Untuk papan pemuka kajian ini, *Streamlit* akan digunakan untuk mencipta aplikasi web yang boleh meramal keparahan pesakit COVID-19. Aplikasi web ini dapat menerima input pengguna dan kemudian memberikan keputusan tahap keparahan keadaan COVID-19. Rajah 7 telah memaparkan papan pemuka aplikasi web *Streamlit* yang mengandungi maklumat tentang COVID-19 dan juga sistem ramalan keparahan pesakit COVID-19 yang memerlukan pengguna memasukkan pelbagai ciri input yang berkaitan dengan gejala COVID-19.



Rajah 7 Sistem ramalan keparahan pesakit COVID-19

Selepas pengguna memilih pelbagai ciri input yang berkaitan dengan gejala COVID-19 menggunakan komponen *sidebar Streamlit selectbox*, keputusan ramalan keparahan keadaan pesakit COVID-19 akan dikeluarkan dengan kebarangkalian. Rajah 8 telah menunjukkan bahawa pesakit COVID-19 berada dalam keadaan ringan dengan kebarangkalian 81.89%.



Rajah 8 Keputusan ramalan keparahan pesakit COVID-19 berada dalam keadaan ringan

Rajah 9 telah menunjukkan bahawa pesakit COVID-19 berada dalam keadaan sederhana dengan kebarangkalian 97.37%.

Rajah 9 Keputusan ramalan keparahan pesakit COVID-19 berada dalam keadaan sederhana

Rajah 10 telah menunjukkan bahawa pesakit COVID-19 berada dalam berada dalam keadaan serius dengan kebarangkalian 91.56%.

Rajah 10 Keputusan ramalan keparahan pesakit COVID-19 berada dalam keadaan serius

Kesimpulan

Hasil kajian ini menunjukkan bahawa model hutan rawak mengatasi model ramalan lain seperti regresi logistik, mesin vektor sokongan, *Gaussian Nave Bayes*, dan pohon keputusan. Model hutan rawak diiktiraf sebagai model yang paling sesuai bagi ramalan keparahan pesakit COVID-19. Keputusan ramalan keparahan COVID-19 boleh diakses oleh pengguna pada aplikasi web *Streamlit*. Oleh itu, objektif yang ditetapkan dalam fasa pemahaman bisnes telah berjaya dicapai melalui kajian ini. Matlamat utama projek ini adalah untuk membangunkan model pembelajaran mesin untuk meramalkan tahap keparahan pesakit COVID-19, dan ini telah dicapai dengan tahap ketepatan yang memuaskan. Hasil kajian ini akan memberi impak dan implikasi yang signifikan dalam sains kesihatan dan industri berkaitan. Dengan

menggunakan teknik pembelajaran mesin, kami boleh menyumbang untuk mengatasi cabaran yang dihadapi dalam menguruskan wabak COVID-19. Model yang dibangunkan boleh membantu pihak berkuasa dalam memperuntukkan sumber dengan lebih cekap, meningkatkan peluang pemulihan untuk pesakit, dan mengurangkan beban pada sistem penjagaan kesihatan. Walaupun objektif kajian telah dicapai dan perancangan awal telah diikuti, terdapat beberapa kekangan yang dihadapi. Keberkesanan model ramalan sangat bergantung pada kualiti dan keseragaman data yang digunakan untuk melatih algoritma. Dalam projek ini, set data yang digunakan adalah tidak seimbang, iaitu data pesakit yang dalam kelas serius sangat kurang dan apabila membuat ramalan akan menyebabkan sukar untuk meramal pesakit dalam keadaan serius. Hal ini telah memberi kesan kepada kebolehpercayaan model. Selain itu, memahami dan mengenal pasti atribut penting yang berkaitan dengan keparahan pesakit COVID-19 adalah penting dalam membangunkan model yang berkesan. Kekangan pengetahuan dalam bidang perubatan atau epidemiologi mungkin mempengaruhi pemilihan atribut dan interpretasi hasil model. Untuk meningkatkan model ramalan pada masa hadapan. cadangan yang boleh dibuat untuk menambahbaik aplikasi web ini adalah memperoleh akses kepada lebih banyak data pesakit COVID-19 yang berkualiti tinggi dan mencukupi boleh membantu meningkatkan keupayaan model. Hal ini boleh dilakukan dengan bekerjasama dengan institusi perubatan atau pihak berkepentingan yang mempunyai data yang relevan. Di samping itu, menggunakan teknik pensampelan yang lebih sesuai bagi klasifikasi berbilang kelas boleh menyelesaikan masalah ketidakseimbangan data.

Secara keseluruhannya, kajian ini berjaya membangunkan model pembelajaran mesin yang berkesan untuk meramalkan tahap keparahan pesakit COVID-19. Penemuan ini mempunyai implikasi penting untuk mengurus sumber kesihatan dan menyediakan penjagaan tepat pada masanya kepada pesakit. Walaupun terdapat beberapa kelemahan dalam kajian ini, hasilnya tetap memberi sumbangan yang berharga dalam menghadapi cabaran pandemik

COVID-19. Penyelidikan masa depan boleh memperhalusi dan meningkatkan model ini untuk memberikan manfaat yang lebih besar dalam pengurusan wabak masa depan.

Penghargaan

Terlebih dahulu, saya ingin mengambil peluang ini untuk mengucapkan ribuan terima kasih kepada penyelia saya yang amat dihormati, Prof. Madya Dr. Zalinda Othman diatas segala tunjuk ajar dan bantuan yang diberikan kepada saya sepanjang projek ini dijalankan. Beliau telah meluangkan masa yang emas dengan penuh kesabaran untuk memberikan bimbingan dan nasihat yang baik kepada saya untuk melengkapkan projek ini dengan lancar. Jutaan terima kasih juga saya ucapkan kepada para pensyarah FTSM yang telah menaburkan ilmu pengetahuan kepada saya sepanjang pengajian saya di Universiti Kebangsaan Malaysia (UKM). Selain itu, saya juga ingin merakamkan setinggi-tinggi terima kasih kepada kedua-dua ibu bapa serta ahli keluarga saya. Mereka sentiasa memberikan galakan dan dorongan serta menemani saya sepanjang proses menyiapkan projek ini. Dengan ini, saya amat bersyukur kerana mendapat sokongan moral yang penuh daripada ibu bapa saya. Di samping itu, saya juga ingin merakamkan penghargaan saya kepada rakan saya. Mereka telah banyak menyumbangkan idea-idea serta nasihat kepada saya semasa perjalanan projek ini. Mereka juga sanggup berkongsi maklumat tentang projek ini dan menghulurkan bantuan kepada saya. Dengan sokongan moral mereka, saya dapat menyiapkan projek ini dengan lancar sekali. Akhir sekali, saya ingin mengucapkan jutaan terima kasih sekali lagi kepada semua pihak yang terlibat secara langsung atau tidak langsung yang telah memberikan bantuan kepada saya dalam proses menghasilkan projek tahun akhir ini. Sekian, terima kasih.

Rujukan

Aljameel, S.S., Khan, I.U., Aslam, N., Aljabri, M. & Alsulmi, E.S. 2021. Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients. *Scientific Programming* 2021.

Buvana, M. & Muthumayil, K. 2021. Prediction of covid-19 patient using supervised machine learning algorithm. *Sains Malaysiana* 50(8): 2479–2497.

COVID Live - Coronavirus Statistics by Country- Worldometer. 2022. .https://www.worldometers.info/coronavirus/ [9 Disember 2022].

Gull, H., Krishna, G., Aldossary, M.I. & Iqbal, S.Z. 2020. Severity Prediction of COVID-19 Patients Using Machine Learning Classification Algorithms: A Case Study of Small City in Pakistan with Minimal Health Facility. *2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020*: 1537–1541.

Jahangirimehr, A., Abdolahi Shahvali, E., Rezaeijo, S.M., Khalighi, A., Honarmandpour, A., Honarmandpour, F., Labibzadeh, M., Bahmanyari, N. & Heydarheydari, S. 2022. Machine learning approach for automated predicting of COVID-19 severity based on clinical and paraclinical characteristics: Serum levels of zinc, calcium, and vitamin D. *Clinical Nutrition ESPEN* 51: 404–411.

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., Zhang, T., Wu, Z., He, G. & Huang, Y. 2020. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials and Continua* 63(1): 537–551.

Kahn, J.S. & McIntosh, K. 2005. History and Recent Advances in Coronavirus Discovery. *Pediatric Infectious Disease Journal* 24(11): S223–S227.

Luna, Z. 2021. Understanding CRISP-DM and its importance in Data Science projects - Medium. https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-

importance-in-data-science-projects-91c8742c9f9b [17 Disember 2022].

Pokhrel, S. & Chhetri, R. 2021. A Literature Review on Impact of COVID-19 Pandemic on Teaching and Learning. *Higher Education for the Future* 8(1): 133–141.

Pourhomayoun, M. & Shakibi, M. 2021. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* 20(February).

Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. & Choi, G.S. 2020. COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access* 8: 101489–101499.

scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation. (n.d.). .https://scikit-learn.org/stable/index.html [18 Julai 2023].

Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y. & Gao, G.F. 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology* 24(6): 490–502.

Tanda-Tanda Orang Dewasa dan Kanak-Kanak Yang Dijangkiti COVID-19. 2020. .https://covid-19.moh.gov.my/garis-panduan/gp-umum-covid19/gp-tanda-orang-dewasa-dan-kanak-kanak-dijangkiti-covid-19 [9 Disember 2022].

Verma, A.K. & Prakash, S. 2020. Impact of COVID-19 on Environment and Society. *Journal of Global Biosciences* 9(5): 7352–7363.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F. & Tan, W. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine* 382(8): 727–733.

Ooi Teng He (A179637)
Prof. Madya Dr. Zalinda Othman
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia

Commented [U2]: No. Matriks Pelajar