MODEL PEMBELAJARAN MESIN UNTUK MERAMAL KESAN JANGKA PANJANG KE ATAS PESAKIT POSITIF COVID-19

CHEOK KAH YEEK

AZURALIZA ABU BAKAR

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

ABSTRAK

Kesan-kesan jangka panjang COVID-19 yang berterusan, sering dirujuk sebagai COVID berjangka panjang tetap menjadi cabaran besar walaupun puncak wabak telah berlalu. Selain daripada sesetengah individu yang berpulih, majoriti pesakit masih mengalami gejala yang berterusan seperti keletihan, demam, dan masalah pernafasan serta jantung yang boleh berlarutan selama minggu-minggu atau tahunan. Kondisi berterusan ini bukan sahaja membawa kesan buruk kepada kesejahteraan individu tetapi juga mengakibatkan kehilangan produktiviti dalam tenaga kerja. Oleh itu, terdapat keperluan yang semakin meningkat untuk model pembelajaran mesin yang boleh meramalkan kesan jangka-panjang ke atas pesakit positif Covid-19. Kajian ini dapat mengatasi ketidakpastian mengenai pesakit yang mungkin mengalami gejala COVID berjangka panjang dengan menganalisis rekod kesihatan pesakit. Rekod penjagaan kesihatan elektronik yang komprehensif digunakan oleh model pembelajaran mesin untuk mengenal pasti pesakit COVID-19 yang berisiko mengalami gejala berpanjangan. Dengan menyemak rekod kesihatan dari 1,132 pesakit COVID Eropah yang mempunyai 186 ciri, atribut yang berkaitan dengan COVID berjangka panjang diekstrak dan digunakan untuk melatih model pokok keputusan XGBoost. Model ini membantu mengenal pasti pesakit COVID berjangka panjang yang berpotensi dan dinilai menggunakan metrik seperti ketepatan, kejituan, ingatan, dan skor F1. Keberkesanan model ini berpotensi untuk membolehkan pengesanan awal, pencegahan proaktif, pengagihan sumber yang dioptimumkan, peningkatan kesedaran kesihatan awam, dan sokongan untuk penyelidikan dan pembangunan dasar. 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 kesan jangka panjang, pembelajaran mesin

PENGENALAN

Kesan jangka panjang COVID-19 yang berterusan (COVID-19 jangka panjang), juga dikenali sebagai Sindrom Post Covid telah menjadi cabaran besar kepada masyarakat yang pernah mengalami wabak COVID-19. COVID-19 jangka panjang merujuk kepada keadaan di mana bekas pesakit COVID-19 menunjukkan gejala-gejala bagi tempoh 12 minggu atau lebih yang tidak dapat dijelaskan dengan sebarang diagnosis alternatif. Antara simptom yang kerap dialami oleh bekas pesakit COVID-19 adalah kelesuan, kesukaran bernafas, sukar tidur malam, batuk dan kegelisahan. Simptom-simptom tersebut berpunca daripada komplikasi jangkitan COVID-19 terhadap fungsi pelbagai sistem dan organ dalam tubuh pesakit (KKM 2021). Sekurang-kurangnya 65 juta individu di seluruh dunia mengalami kesan jangkitan COVID-19 yang berpanjangan, berdasarkan anggaran kejadian yang konservatif sebanyak 10% daripada individu yang dijangkiti dan lebih daripada 651 juta kes COVID-19 yang berdokumentasikah di seluruh dunia (Balllering et al. 2022).

Salah satu ciri COVID-19 jangka panjang adalah ia mempengaruhi bekas pesakit COVID-19 dalam semua keparahan penyakit. Kajian-kajian telah mendapati bahawa COVID-19 jangka panjang mempengaruhi kes-kes ringan hingga sederhana dan dewasa muda yang tidak memerlukan sokongan pernafasan, rawatan hospital, atau rawatan rapi. Pesakit-pesakit yang tidak lagi positif terhadap COVID-19 dan telah dibenarkan keluar dari hospital, serta pesakit rawatan luar, juga boleh mengalami COVID-19 jangka panjang.

Sifat ketidakpastian COVID-19 jangka panjang telah menjadi cabaran untuk mengenal pasti dan meramalkan kesan-kesan yang akan berlaku (Sudre, C. et al. 2021). Terdapat beberapa kajian berusaha untuk mengenal pasti hubungan atau faktor-faktor yang menyebabkan COVID-19 jangka panjang. Dalam kajian Roman Kessler (Kessler,R et al. 2023), rekod kesihatan digital pesakit COVID-19 yang teliti telah digunakan untuk melatih model pembelajaran mesin. Beberapa ciri seperti usia, jantina dan rekod lengkap data diagnosis dan preskripsi telah dipilih untuk penilaian faktor-faktor COVID-19 jangka panjang dan simptomnya. Terdapat sesuatu kajian yang telah berjaya meramalkan COVID-19 jangka panjang dengan menggunakan faktor sosiodemografi dan keparahan gejala ketika jangkitan COVID-19 dengan reka bentuk kes-kawalan (Sudre et al. 2021). Di samping itu, sesuatu kajian dilakukan menggunakan model peningkatan gradien yang dilatih dengan maklumat pesakit

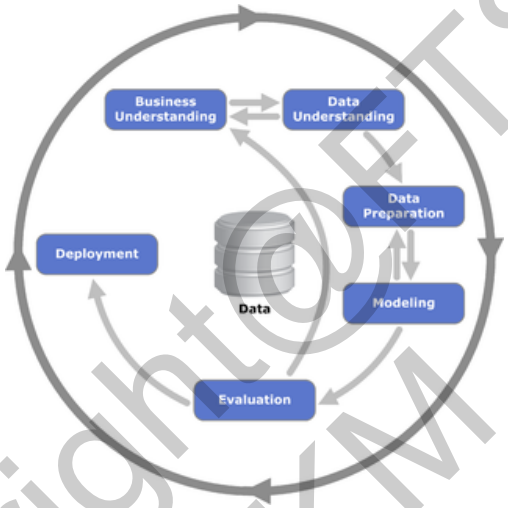
COVID-19 dirawat di klinik COVID-19 jangka panjang khusus (Pfaff, E.R et al. 2022). Dalam kerjanya, ciri-ciri penting data ditentukan oleh Shapley value. Model ini berjaya membuat ramalan yang tepat, mempunyai ketepatan 0.92 bagi semua pesakit.

Pembelajaran mesin dan pembelajaran mendalam ialah salah satu teknologi kecerdasan buatan (AI) yang memberikan keupayaan ramalan yang hebat dan mengatasi prestasi pemodelan statistik tradisional (Beam & Kohane, 2018; Miguel-Hurtado et al., 2016; Singal et al., 2013).Bagi mengatasi masalah ketidakpastian kesan jangka panjang COVID-19, terdapat beberapa kajian yang menggunakan pembelajaran mesin untuk meramalkan kesan jangka panjang COVID-19. Berdasarkan kerja Natalya Shakhovska, sebuah model pembelajaran mesin hibrid digunakan untuk ramalan kesan jangka panjang COVID-19 serta keparahannya(Natalya Shakhovska 2022). Model hibrid ini mempunyai tiga lapisan, iaitu pembinaan peraturan asosiatif, pemilihan pengelas lemah dan model Random Forest. Pendekatan ini dapat menggabungkan beberapa model untuk mendapat manfaatmya. Selain itu, berdasarkan kerja Emily R Pfaff, sebuah model pembelajaran mesin XGBoost telah digunakan untuk mengelaskan pesakit yang berpotensi mengalami kesan jangka COVID-19 (Emily R Pfaff et al. 2022). Di samping itu, teknik pembelajaran mesin regresi logistik juga digunakan dalam kajian (Kulenovic,A. Et al. 2022). Dalam kajian ini, sebuah model regresi logistik dibangunkan untuk setiap kesan jangka panjang COVID-19. Kajian ini mengenal pasti wujudnya hubungan antara pelbagai ciri-ciri kesan jangka panjang COVID-19.

Kesimpulannya, kajian ini dicadangkan kerana cabaran yang dihadapi oleh pengkaji untuk meramalkan kesan jangka-panjang ke atas pesakit positif Covid-19 dalam era pasca pandemic COVID-19 ini. Pendekatan secara pembelajaran mesin dicadangkan untuk meramal COVID-19 jangka panjang dalam kalangan pesakit COVID-19. Pendekatan ini dapat dilaksanakan kerana ia menunjukkan hubungan antara parameter dan ramalan COVID-19 jangka panjang berdasarkan kajian lain-lain. Hasil akhir kepada kajian ini dapat membantu pihak berkuasa untuk menilai risiko COVID-19 jangka panjang dan mengambil inisiatif yang sesuai untuk membantu individu yang mengalami COVID-19 jangka panjang kembali kepada kehidupan biasa.

METODOLOGI KAJIAN

Metodologi yang digunakan untuk kajian model pembelajaran mesin untuk meramal kesan jangka-panjang ke atas pesakit positif COVID-19 ialah Cross Industry Standard Process for Data Mining (CRISP-DM). Metodologi ini adalah model proses yang berperanan sebagai asas untuk proses sains data. Ia mempunyai enam fasa berurutan iaitu Pemahaman Perniagaan (Business Understanding), Pengetahuan Data (Data Understanding), Penyediaan Data (Data Preparation), Pemodelan (Modelling), Penilaian (Evaluation) dan Penggunaan (Deployment)



Fasa Pemahaman Data

Fasa pemahaman data bertujuan memastikan bahawa data sesuai untuk analisis yang diinginkan. Ini dicapai dengan meneroka data, mengenal pasti sebarang masalah, dan membuat penyesuaian yang diperlukan kepada data untuk memastikan kualiti dan kebolehpercayaannya (Han et al. 2012). Dalam fasa ini, beberapa maklumat tentang data yang dikumpul perlu dikenal pasti seperti tempat untuk mendapatkan data, alat yang digunakan untuk mendapatkan data dan sejauh mana data yang tersedia adalah penting (Luna 2021). Sumber data yang digunakan untuk kajian projek ini adalah data rekod kesihatan digital pesakit COVID-19 daripada laman web Harvard Dataverse:Long Covid Data Set. Selepas muat turun set data, pemeriksaan struktur dan kandungan set data dijalankan untuk mengenal pasti pemboleh ubah, jenis data dan hubungan yang mungkin (Hillier 2023). Pembersihan awal data juga dilakukan untuk menangani nilai yang hilang dengan menggantikan atau membuang mereka. Langkah ini adalah untuk membetulkan sebarang kesilapan atau inkosistensi dalam set data. Di samping itu, hubungan antara ciri-ciri akan diteroka dengan menggunakan analisis korelasi untuk

memahami hubungan dalam data. Selepas itu, ringkasan awal data tentang sebarang cabaran atau had dalam set data akan dibuatkan dengan merangkumi penemuan utama dari proses pemeriksaan dan pembersihan. Akhirnya, representasi visual akan dicipta untuk memahami corak dan trend dalam data melalui carta, graf dan plot. Dalam projek ini, fasa pemahaman data dalah sangat penting disebabkan set data yang digunakan mempunyai sebanyak 186 ciri-ciri dan bersifat rumit. Rajah di bawah menunjukkan set data mentah sebelum pemprosesan.

Fasa Penyediaan Data

Fasa Penyediaan Data adalah fasa yang memproseskan data mentah dan membersihkan data. Fasa ini dapat menghasilkan set data yang bersih dan sesuai untuk digunakan sebagai set data latihan dan ujian dalam pembangunan model pembelajaran mesin (Han et al. 2012). Dengan menggunakan Google Colab, pemprosesan data rekod kesihatan digital pesakit COVID-19. Langkah-langkah penyediaan data yang diambilkan adalah pembersihan data, transformasi data, visualisasi data, analisis korelasi dan pemilihan ciri.

Fasa Pemodelan

Pembangunan algoritma dalam projek ini adalah untuk meramalkan kesan jangka panjang COVID-19. Berdasarkan kajian dalam ramalan kesan jangka panjang COVID-19, terdapat beberapa model dan algoritma yang dibangunkan bagi mencapai objektif kajian. Antara model yang akan dibangunkan adalah Algoritma Penggalakan Gradien (XGBoost), Algoritma Regresi Logistik, Algoritma Random Forest dan Algoritma Decision Tree. Penalaan Hiperparameter dijalankan menggunakan GridSearch untuk menambahbaik persembahan dan keputusan model-model tersebut. GridSearch ialah teknik yang melibatkan pencarian secara menyeluruh melalui satu set nilai-nilai yang telah ditentukan untuk setiap hiperparameter. Tujuannya adalah untuk mengenal pasti kombinasi hiperparameter yang memberikan prestasi terbaik berdasarkan metrik penilaian tertentu. Melalui kaedah ini, setiap kombinasi yang mungkin akan diuji, dan hasilnya akan dianalisis untuk memilih parameter yang paling optimum.

Fasa Pemodelan

Fasa Penilaian adalah fasa di mana model pembelajaran mesin dinilai untuk memastikan ia mencapai objektif bisnes yang ditetapkan dalam fasa pemahaman bisnes. Prestasi model

diperiksa untuk memastikan keputusan yang dihasilkan adalah memuaskan. Jika keputusan yang diperoleh tidak memuaskan, proses pembinaan model akan diulang dan punca-punca yang menyebabkan keputusan buruk akan dikaji semula. Dalam kajian ini, metrik penilaian yang digunakan untuk menilai peramalan risiko kesihatan pesakit COVID-19 termasuklah AUROC, ketepatan (accuracy), precision, recall, dan skor F1.

Fasa Penyebaran

Fasa penyebaran (Deployment) adalah fasa terakhir proses pembangunan model ramalan kesan jangka panjang COVID-19. Keputusan ramalan yang dihasilkan kini siap untuk disebarkan dan digunakan oleh orang ramai. Hasil model ramalan bagi setiap negeri di Malaysia akan dipaparkan melalui alat visualisasi Tableau secara atas talian, membolehkan pihak berkepentingan memahami dan meneroka hasil analisis dengan cepat dan mudah. Pemantauan dan penyelenggaraan secara berkala perlu dijalankan untuk memastikan hasil ramalan adalah tepat dan mengelakkan penyebaran maklumat palsu kepada masyarakat.

KEPUTUSAN DAN PERBINCANGAN

Penilaian metrik dijalankan terhadap model pembelajaran mesin yang telah dibangunkan untuk menentukan model yang paling sesuai untuk ramalan kesan jangka panjang COVID-19.

Model dinilai dengan menggunakan fungsi roc_auc_score, accuracy_score, classification_report dan confusion_matrix dari sci-kit learn. Dengan menggunakan fungsi-fungsi ini AUROC, ketepatan, precision, skor F1 dan recall dapat dikirakan.

Jadual 1.1 Keputusan Model

	AUROC	**Ketepatan**	**Precision**	**Skor F1**	**Recall**
Algoritma XGBoost	0.8142	0.7137	0.7157	0.7133	0.7142
Algoritma Regresi Logistik	0.8816	0.7974	0.7987	0.7973	0.7978
Algoritma Random Forest	0.8486	0.7533	0.7533	0.7533	0.7533

Algoritma Decision Tree	0.6681	0.6696	0.6695	0.6695	0.6695

Jadual 1.1 menunjukkan perbandingan prestasi antara empat algoritma pembelajaran mesin: XGBoost, Regresi Logistik, Random Forest, dan Decision Tree, berdasarkan metrik AUROC, ketepatan, precision, skor F1, dan recall. Algoritma Regresi Logistik menunjukkan prestasi tertinggi dengan nilai AUROC 0.8816 dan ketepatan 0.7974, mengungguli algoritma lain dalam semua metrik yang dinilai, mencerminkan konsistensi dan keberkesanan yang tinggi. Sebaliknya, Algoritma Decision Tree memperlihatkan prestasi terendah dengan AUROC 0.6681 dan ketepatan 0.6696, menunjukkan kelemahan yang ketara berbanding algoritma lain. Algoritma Random Forest dan XGBoost menunjukkan prestasi sederhana dengan nilai AUROC masing-masing 0.8486 dan 0.8142, di mana Random Forest sedikit lebih baik daripada XGBoost. Perbandingan ini menekankan keunggulan Regresi Logistik dalam tugas pengelasan yang diuji, sementara Decision Tree menunjukkan kekurangan yang ketara. Algoritma Random Forest dan XGBoost, walaupun tidak setinggi Regresi Logistik, masih menunjukkan potensi yang baik dengan prestasi yang konsisten dalam pelbagai metrik.

Jadual 1.2 Keputusan Model dengan Penalaan Hiperparameter

	AUROC	**Ketepatan**	**Precision**	**Skor F1**	**Recall**
Algoritma XGBoost	0.8806	0.7885	0.7899	0.7884	0.7890
Algoritma Regresi Logistik	0.8759	0.7974	0.7976	0.7974	0.7976
Algoritma Random Forest	0.8531	0.7797	0.7797	0.7797	0.7797
Algoritma Decision Tree	0.8047	0.7445	0.7445	0.7445	0.7446

Jadual 1.2 di atas menunjukkan prestasi empat algoritma pembelajaran mesin setelah melalui penalaan hiperparameter. Algoritma XGBoost menunjukkan peningkatan dari sebelumnya dengan nilai AUROC 0.8806, sedangkan Regresi Logistik menunjukkan kestabilan dalam prestasi dengan nilai AUROC 0.8759, hampir serupa dengan sebelumnya. Random Forest menunjukkan peningkatan sedikit dalam beberapa metrik dengan AUROC 0.8531. Namun, Algoritma Decision Tree tetap menunjukkan prestasi yang lebih rendah dengan AUROC 0.8047, meskipun ada peningkatan dari sebelumnya.

Perbandingan antara dua set data menunjukkan bahawa penalaan hiperparameter memberikan manfaat yang berbeza kepada setiap algoritma. Regresi Logistik menunjukkan prestasi yang konsisten, sementara XGBoost menunjukkan peningkatan yang signifikan dalam AUROC setelah penalaan. Random Forest menunjukkan peningkatan sedikit, sementara Decision Tree menunjukkan perbaikan namun masih di bawah algoritma lain. Secara keseluruhan, penalaan hiperparameter memperbaiki prestasi secara umum, walaupun kesan peningkatan bervariasi di antara algoritma yang berbeza.

Tafsiran Model

Jadual 1.3 berikut menyenaraikan ciri-ciri signifikan yang digunakan dalam model-model ramalan kesan jangka panjang COVID-19. Ciri-ciri ini merangkumi faktor klinikal dan demografi yang diambil kira dalam pelbagai model seperti XGBoost, Regresi Logistik, Random Forest, dan Decision Tree. Setiap ciri yang disenaraikan menunjukkan kepentingannya dalam membantu model membuat ramalan yang tepat mengenai kesan jangka panjang COVID-19.

Jadual 1.3 Ciri-ciri Signifikan dalam Model Ramalan

Kategori	**Ciri**	**Penjelasan**
Ciri Klinikal	fever_No	Ketiadaan demam merupakan penunjuk kuat dalam ramalan kesan jangka panjang COVID-19.
	anxietydepression_AcuteCovid_maybe	Pengalaman kemurungan atau kebimbangan semasa jangkitan COVID-19 adalah signifikan dalam ramalan.
	Heart_attack2_maybe	Kemungkinan serangan jantung semasa atau selepas jangkitan COVID-19 merupakan faktor penting.
	Fatigue_now2_0.0	Keletihan semasa jangkitan COVID-19 menunjukkan kepentingan yang signifikan.
	fainting_now2_maybe	Pengsan semasa jangkitan COVID-19 merupakan ciri penting dalam ramalan.
	Type_vaccine_maybe	Jenis vaksin yang diterima memainkan peranan penting dalam model.
	Headache_covid_1.0	Sakit kepala semasa

		jangkitan COVID-19 turut merupakan ciri yang signifikan.
	Hospitalisation_maybe	Kemungkinan kemasukan ke hospital semasa atau selepas jangkitan COVID-19 diambil kira dalam ramalan.
Ciri Demografi	sex_female	Jantina wanita menunjukkan bahawa wanita mengalami risiko kesan jangka panjang COVID-19 yang lebih tinggi.
	agegroup_1.0 dan agegroup_2.0	kumpulan umur menunjukkan bahawa faktor umur mempengaruhi hasil jangka panjang COVID-19.
	Vaccination_status_No	Status vaksinasi turut mempengaruhi ramalan, di mana individu yang tidak divaksinasi mungkin mempunyai risiko yang berbeza berbanding mereka yang divaksinasi.
	Vaccine_doses_2.0	Dos vaksin yang diterima juga mempengaruhi keputusan model.

Jadual 1.5 Perbandingan Ciri-ciri antara Model Pembelajaran Mesin

XGBoost	**Regresi Logistik**	**Random Forest**	**Decision Tree**
fever_No	Fever_No	Fever_No	Anxietydepression_acuteCovid
fainting_now	Fever_not sure	Median_followupTime	Fever_No
anxietydepression_AcuteCovid	Fever_Yes	Heart_Attack	Type_vaccine
Heart_Attack	Last_Fever	Anxietydepression_now	Fever_Yes
Fatigue	Heart_attack	Anxietydepression_acute Covid	Headache_covid
Last_Fever	Vaccination_status	Fatigue	Heart_attack
Anxietydepression_now	Anxietydepresssion_acuteCovid	Fainting_now	Median_followupTime
Fever_Yes	Days_acuteCovid	Sex_female	Fatigue_now
Fatigue_Now	Fatigue	Fever_yes	Last_fever
Days_acuteCovid	Type_vaccine	weaknessArmsLegs_now	Anxietydepression_now

Model-model pembelajaran mesin seperti XGBoost, Regresi Logistik, Hutan Rawak, dan Pokok Keputusan menitikberatkan ciri-ciri berbeza dalam meramalkan hasil kesihatan

berkaitan Covid. XGBoost menekankan gejala seperti demam, pengsan, dan masalah kesihatan mental semasa fasa Covid akut, memanfaatkan keupayaannya untuk menangkap interaksi kompleks. Regresi Logistik memberi tumpuan kepada dinamik demam, status vaksinasi, dan gejala Covid segera, menyediakan pandangan langsung berdasarkan petunjuk klinikal. Hutan Rawak menonjolkan faktor demografi, masa pemantauan sederhana, dan gejala penting seperti serangan jantung dan kelemahan fizikal. Manakala, Pokok Keputusan memberi keutamaan kepada gejala akut, jenis vaksin, dan penanda kesihatan mental, menggunakan pendekatan hierarki untuk membuat keputusan yang jelas dalam penilaian kesihatan.

Kepelbagaian dalam pemilihan ciri-ciri ini membolehkan model-model ini menumpukan kepada aspek yang berbeza dalam penilaian kesihatan berkaitan COVID-19, membantu penyedia penjagaan kesihatan dan penyelidik untuk lebih memahami dan menguruskan risiko dan hasil berkaitan COVID-19.

KESIMPULAN

Secara keseluruhannya, projek ini membina model pembelajaran mesin untuk meramal kesan jangka panjang COVID-19 ke atas pesakit. Sumber data dari Havard Dataverse dijalankan pra pemprosesan dan diguna untuk pemodelan. Pelbagai model pembelajaran mesin, termasuk Algoritma Penggalakan Gradien, Algoritma Regresi Logistik, Algoritma Random Forest dan Algoritma Decision Tree. Penalaan Hiperparameter juga dijalankan untuk mencapai prestasi model yang optimum. Keputusan model dinilai dengan metrik-metrik seperti AUROC, Ketepatan, Precision, Recall dan Skor F1. Selepas itu, ciri-ciri yang menyumbang kepada pembangunan model juga dianalisis dan diinterpretasikan.

PENGHARGAAN

Pertama sekali, saya ingin mengucapkan terima kasih kepada Tuhan yang Maha Esa atas limpahan rahmat dan petunjuk-Nya sepanjang proses penyusunan tesis ini. Kepada penyelia Prof. Dr. Azuraliza Abu Bakar , terima kasih atas bimbingan, sokongan, dan nasihat yang berharga sepanjang perjalanan penyelidikan ini. Saya juga ingin menyampaikan penghargaan

kepada Fakulti / Institut / Pusat Pengajian / Jabatan atas kemudahan penyelidikan yang disediakan, yang telah menyokong kelancaran laporan usulan ini. Akhir sekali, saya juga mengucapkan ribuan terima kasih kepada semua pihak yang turut serta memberi sumbangan dan sokongan, yang telah menjadikan laporan usulan ini berjaya. Saya tidak ketinggalan untuk berterima kasih kepada rakan-rakan dan keluarga saya yang telah berkerjasama dengan saya menjalankan laporan usulan saya. Mereka turut memberi bantuan dan bimbingan.

RUJUKAN

Bruce K. Patterson, Jose Guevara-Coto, Ram Yogendra, Edgar B. Francisco, Emily Long, Amruta Pise, Hallison Rodrigues, Purvi Parikh, Javier Mora & Mora-Rodr´ıguez, R. A. 2021. Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning. *Frontiers in Immunology* 12(700782.

Chen, T. & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. University of Washington.

Chou, E. H., Wang, C. H., Hsieh, Y. L., Namazi, B., Wolfshohl, J., Bhakta, T., Tsai, C. L., Lien, W. C., Sankaranarayanan, G., Lee, C. C. & Lu, T. C. 2021. Clinical Features of Emergency Department Patients from Early COVID-19 Pandemic that Predict SARS-CoV-2 Infection: Machine-learning Approach. *West J Emerg Med* 22(2): 244-251.

Dasu, T. & Johnson, T. 2003. *Exploratory data mining and data cleaning.* John Wiley & Sons.

Eckerson, W. W. 2002. Data quality and the bottom line. *TDWI Report, The Data Warehouse Institute* 1-32.

George, D. & Mallery, P. 2018. Descriptive statistics. Dlm. (pnyt.). *IBM SPSS Statistics 25 Step by Step*, hlm. 126-134. Routledge.

Kessler, R., Philipp, J., Wilfer, J. & Kostev, K. 2023. Predictive Attributes for Developing Long COVID-A Study Using Machine Learning and Real-World Data from Primary Care Physicians in Germany. *J Clin Med* 12(10):

Kulenovic, A. & Lagumdzija-Kulenovic, A. 2022. Using Logistic Regression to Predict Long COVID Conditions in Chronic Patients. *Stud Health Technol Inform* 295(265-268.

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., Mcmurry, J. A., Moffitt, R., Walden, A., Chute, C. G., Haendel, M. A. & Consortium, N. C. 2022. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health* 4(7): e532-e541.

Reina Reina, A., Barrera, J. M., Valdivieso, B., Gas, M. E., Mate, A. & Trujillo, J. C. 2022. Machine learning model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical patients. *Sci Rep* 12(1): 5723.

Shafranovich, Y. 2005. Common format and MIME type for comma-separated values (CSV) files.

Shakhovska, N., Yakovyna, V. & Chopyak, V. 2022. A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system. *Math Biosci Eng* 19(6): 6102-6123.

Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J.-H. & Hsieh, J.-G. 2021. COVID-19 Prediction Applying Supervised Machine Learning

Algorithms with Comparative Analysis Using WEKA. *Algorithms*

Waskom, M. L. 2021. Seaborn: statistical data visualization. *Journal of Open Source Software* 6(60): 3021.

Wolcott, H. F. 1994. *Transforming qualitative data: Description, analysis, and interpretation.* Sage.

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Cheng, C., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H. & Yuan, Y. 2020. An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence* 2(5): 283-288.

Yu, L., Zhou, R., Chen, R. & Lai, K. K. 2022. Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade* 58(2): 472-482.

Zoabi, Y., Deri-Rozov, S. & Shomron, N. 2021. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med* 4(1): 3.

Cheok Kah Yeek (A189479)
Prof. Dr. Azuraliza Abu Bakar
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia