MODEL PENGELASAN BAGI MERAMAL KETAKMAMPUAN KOGNITIF RINGAN (MCI) DALAM KALANGAN WARGA EMAS

Eileen Tong Hui Guan[1] & Suhaila Zainudin[2]

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

Abstrak

Ketakmampuan kognitif merupakan kecacatan terhadap kemahiran kognitif sama ada dari segi peringatan, pemahaman atau kesedaran. Bagi mengkaji hubungan antara metrik kesihatan, keadaan sosial, dan trend ketakmampuan kognitif warga emas, set data Long-Term Research Grant Scheme project: Towards Usual Ageing (LRGS TUA) telah dibentuk. Ketakmampuan kognitif boleh dirawat sekiranya ia dikesan awal. Trend penuaan penduduk Malaysia menyebabkan nisbah populasi warga emas meningkat, tetapi ujian daya kognitif jarang dimasukkan dalam rutin saringan kesihatan kerana mengambil masa untuk dijalankan. Satu model pengelasan ketakmpmuan kognitif ringan (MCI) dicadangkan bagi memudahkan usaha mengesan warga emas yang mengalami ketakmampuan kognitif ringan berdasarkan metrik kesihatan dan keadaan sosial mereka. Objektif projek adalah untuk menjalankan pra-pemprosesan data terhadap data kajian longitudinal LRGS TUA dan menilai algoritma pemodelan terbaik bagi membina model pengelasan ketakmampuan kognitif ringan. Projek menjalankan pemilihan fitur secara konseptual daripada data LRGS TUA, menapis rekod set data, menggabungkan fitur, membetulkan data, mengimput data dan mengekod semula fitur bagi menyediakan set data untuk tujuan latihan model pengelasan. Model pengelasan yang dilatih ialah model linear regresi logistik, pepohon keputusan C4.5 mesin vektor sokongan, model XGBoost dan ensembel Hutan Rawak. Model pengelasan dinilai melaui skor kejituan, skor panggil semula (recall), ROC AUC, logistic loss dan min ralat kuasa dua (min squared error). Projek juga mencuba membuat pengklusteran terhadap set data. Konfigurasi model

dengan prestasi tertinggi ialah ensembel Hutan Rawak menggunakan set data undersample dengan 30 fitur terpenting. Metrik yang dicapai ialah 0.531 skor kejituan, 0.657 skor panggil semula, 0.581 ROC AUC, 16.922 logistic loss dan 0.469 min ralat kuasa dua. Pengklusteran menggunakan K-Mean, peta semantik swaurus dan DBSCAN gagal memisah set data yang cukup diskrit untuk tujuan pengelasan. Pemahaman domain yang lebih menyeluruh dan pembentukan kluster yang lebih diskrit dalam set data LRGS TUA bagi mendapat sampel ideal untuk latihan model boleh meningkatkan prestasi pengelasan data.

Kata kunci: Perlombongan data, Pengklusteran, Ketakmampuan Kognitif Ringan

Pengenalan

Set data *Long-Term Research Grant Scheme project: Towards Usual Ageing* (LRGS TUA) merupakan jawapan soal-selidik kuantitatif kesihatan fizikal, kecerdasan mental, dan keadaan sosial daripada sampel warga emas berumur lebih 60 tahun. Tujuannya adalah untuk mengenalpasti kes ketakmampuan kognitif ringan (*Mild Cognitive Impairment*, MCI) dalam kalangan warga emas sihat bagi membantu usaha intervensi awal untuk menangani kemerosotan kecerdasan minda. Set data longitudinal LRGS TUA mempunyai jawapan responden daripada tiga detik masa, iaitu detik masa garis tapak, detik masa bulan ke-18, dan detik masa bulan ke -36.

Warga emas yang mengalami MCI didefinisikan sebagai mereka yang melaporkan diri mempunyai masalah ingatan, keputusan ujian kognitif rendah, tetapi tidak mengalami dementia dan tiada halangan dalam menjalankan aktiviti kehidupan harian.

Kajian ini menggunakan teknik perlombongan data terhadap data LRGS TUA untuk membina model pengelasan yang menggunakan metrik kesihatan fizikal, sejarah kesihatan dan keadaan sosial warga emas untuk mengenal pasti mereka yang ada MCI.

Penyataan Masalah

Ketakmampuan kognitif ringan (*Mild Cognitive Impairment*, MCI) merupakan petunjuk seseorang warga emas berisiko tinggi mengalami masalah kognitif serius pada masa akan datang (Petersen et al. 2014). Pengesanan kes MCI secara awal membolehkan tindakan diambil bagi memulih daya kognitif pesakit kepada paras asal, seterusnya menurunkan risiko kemerosotan daya kognitif pada masa akan datang. Pengesanan kes MCI adalah berdasarkan pelbagai ujian daya kognitif yang mengambil masa untuk dijalankan, dan jarang dimasukkan dalam rutin saringan kesihatan biasa. Sebaliknya, faktor lain seperti sejarah penyakit, keadaan sosial, dan ukuran metrik kesihatan lebih kerap disimpan dalam rekod pesakit. Tambahan pula,

mereka yang mempunyai MCI melaporkan diri mereka mempunyai masalah ingatan, disahkan mempunyai masalah ingatan melalui ujian kognitif, tetapi tidak mengalami dementia serta kurang masalah berfungsi dalam aktiviti seharian. Gejala yang mereka alami amat ringan, menyebabkan mereka sukar dibezakan daripada warga emas lain yang sihat. Kajian bertujuan menjalankan pra-pemprosesan data terhadap data kajian longitudinal LRGS TUA dan menilai algoritma pemodelan terbaik bagi membina model pengelasan MCI. Ada kaedah statistik dan pemodelan data yang tidak dapat diimplementasi sebab kurang kefahaman terhadapnya, dan kekangan masa untuk mengkajinya secara mendalam. Walaupun terdapat akses kepada penyelidik asal yang membina data TUA dan dua orang pensyarah perubatan am sebagai penasihat domain, projek ini tiada akses kepada pakar penyelidik dalam domain kajian yang dilakukan, maka pengesahan aliran pemprosesan data dan keputusan model tidak dapat dibuat secara sempurna. Laporan ini menerangkan metodologi kajian dan hasil kajian.

Metodologi Kajian

Pembinaan model pengelasan dibahagikan kepada dua fasa, iaitu fasa pemprosesan data dan fasa latihan model pengelasan. Penambahbaikan metodologi terhadap fasa tersebut dijalankan selepas setiap kali pembinaan model dalam usaha meningkatkan ketepatan pengelasan model. Terdapat fasa tambahan iaitu fasa pengklusteran bagi mewujudkan sampel ideal untuk kes Ada MCI dan Tiada MCI.

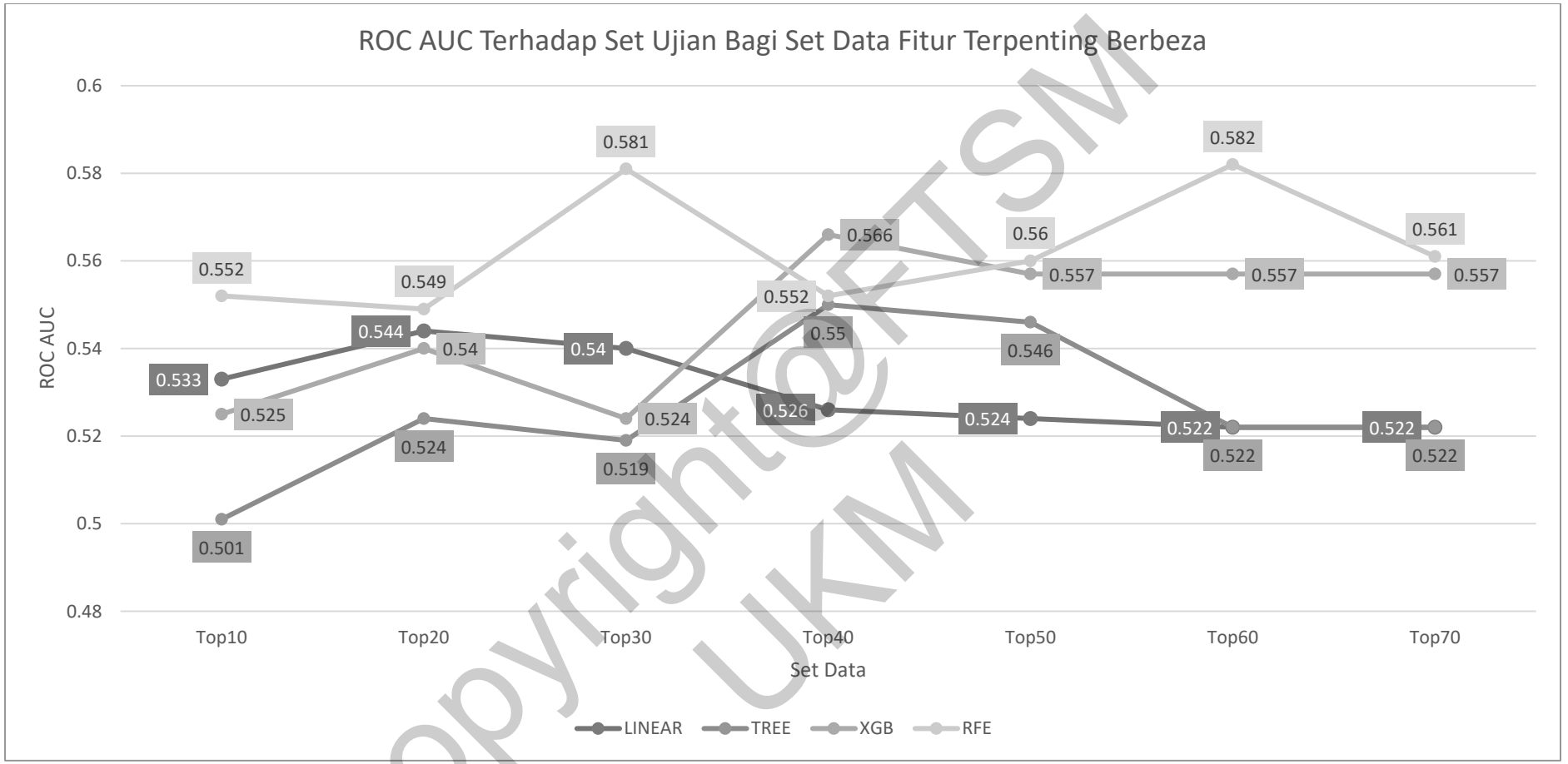
Dalam fasa pemprosesan data, pemahaman terhadap set data dilakukan. Ia juga fasa di mana data dibersihkan dan diformat bagi tujuan latihan model pengelasan. Langkah pemprosesan data yang dilakukan ialah pemilihan fitur secara konseptual, penapisan set data mentah, penggabungan fitur, pembetulan data, pengimputan data hilang, dan pengekodan semula fitur nominal dan ordinal. Eksperimen terhadap perubahan metodologi yang dialkukan ialah pengurangan kekardinalan data.

Dalam Fasa Latihan Model Pengelasan, pemodelan data dilakukan. Set data detik garis tapak LRGS TUA digunakan sebagai set data latihan, manakala set data detik bulan ke-18 LRGS TUA diguanakan sebagai set data ujian. Algoritma pembinaan model yang digunakan ialah model linear, pepohon keputusan, mesin vektor sokongan (SVM), *XGBoost* dan ensembel Hutan Rawak, dengan model palsu sebagai rujukan. Metrik pengelasan model yang dilihat ialah skor kejituan, skor panggil semula (*recall*), luas bawah lengkung ciri operasian penerima (ROC AUC), *logistic loss* dan min ralat kuasa dua. Dua set metrik pengelasan model direkod, satu untuk proses latihan model menggunakan purata skor semasa validasi silang, dan satu untuk proses ujian model. Eksperimen terhadap perubahan metodologi yang dilakukan ialah pengimbangan data dan analisis komponen utama (PCA).

Dalam fasa pengklusteran, algoritma pengklusteran dijalankan terhadap set data garis tapak LRGS TUA bagi mencari sampel ideal kes Ada MCI dan kes Tiada MCI. Algoritma pengklusteran yang digunakan ialah K-Mean, peta semantik swaurus (SOM) dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Parameter pengklusteran diubah bagi mengenal pasti konfigurasi terbaik bagi pengklusteran data. Kluster dinilai berdasarkan nisbah kes Ada MCI terhadap kes Tiada MCI.

Keputusan dan Perbincangan

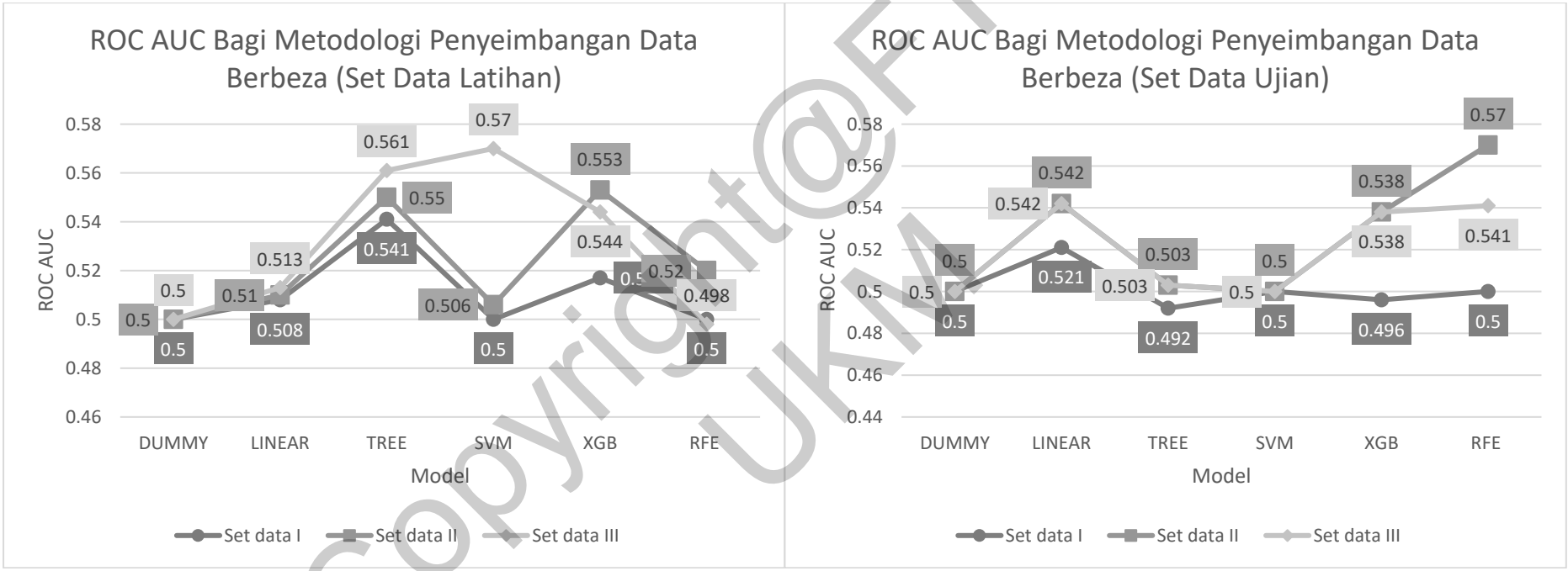
Model ensembel Hutan Rawak dengan menggunakan set data 30 fitur terpenting dalam pembinaan model mencapai prestasi yang paling tinggi antara kesemua kaedah pemodelan dan konfigurasi model yang diuji. Metrik pengelasan yang dicapai ialah 0.531 skor kejituan, 0.657 skor panggil semula, 0.581 skor ROC AUC, 16.922 skor *logistic loss* dan 0.469 skor min ralat kuasa dua. Hanya graf ROC AUC ditunjukkan sahaja disebabkan kekangan ruang.



* LINEAR ialah model linear, TREE ialah pepohon keputusan, XGB ialah *XGBoost*, RFE ialah ensembel Hutan Rawak

Rajah 1 ROC AUC Terhadap Set Ujian Bagi Set Data Fitur Terpenting Berbeza

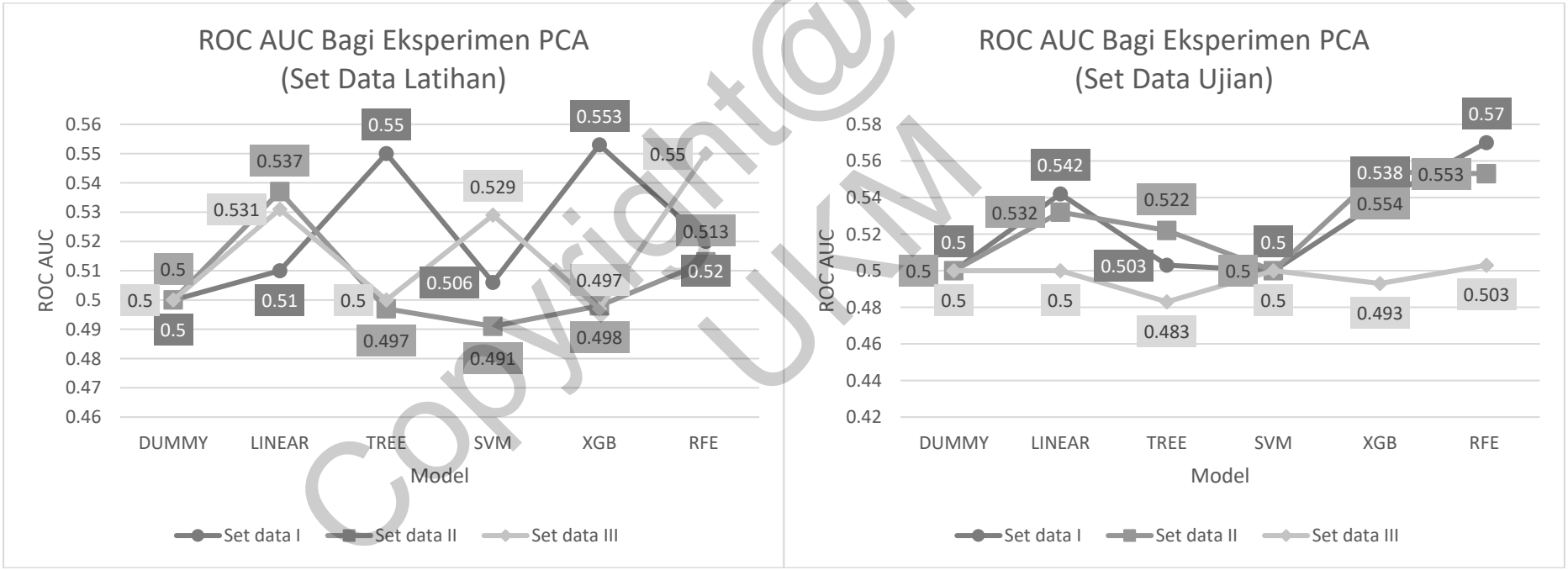
Kedua-dua kaedah penyeimbangan data secara *undersample* dan *oversample* (SMOTE) berjaya meningkatkan prestasi skor panggil semula model, dan seterusnya skor ROC AUC. Kaedah *undersample* dikenal pasti sebagai kaedah penyeimbangan data yang paling bagus kerana tidak memperkenalkan data buatan dalam set data latihan. Hanya graf ROC AUC ditunjukkan sahaja disebabkan kekangan ruang.



*DUMMY ialah model palsu, LINEAR ialah model linear, TREE ialah pepohon keputusan, SVM ialah mesin vektor sokongan, XGB ialah *XGBoost*, RFE ialah ensembel Hutan Rawak

Rajah 2 ROC AUC Bagi Metodologi Penyeimbangan Data Berbeza

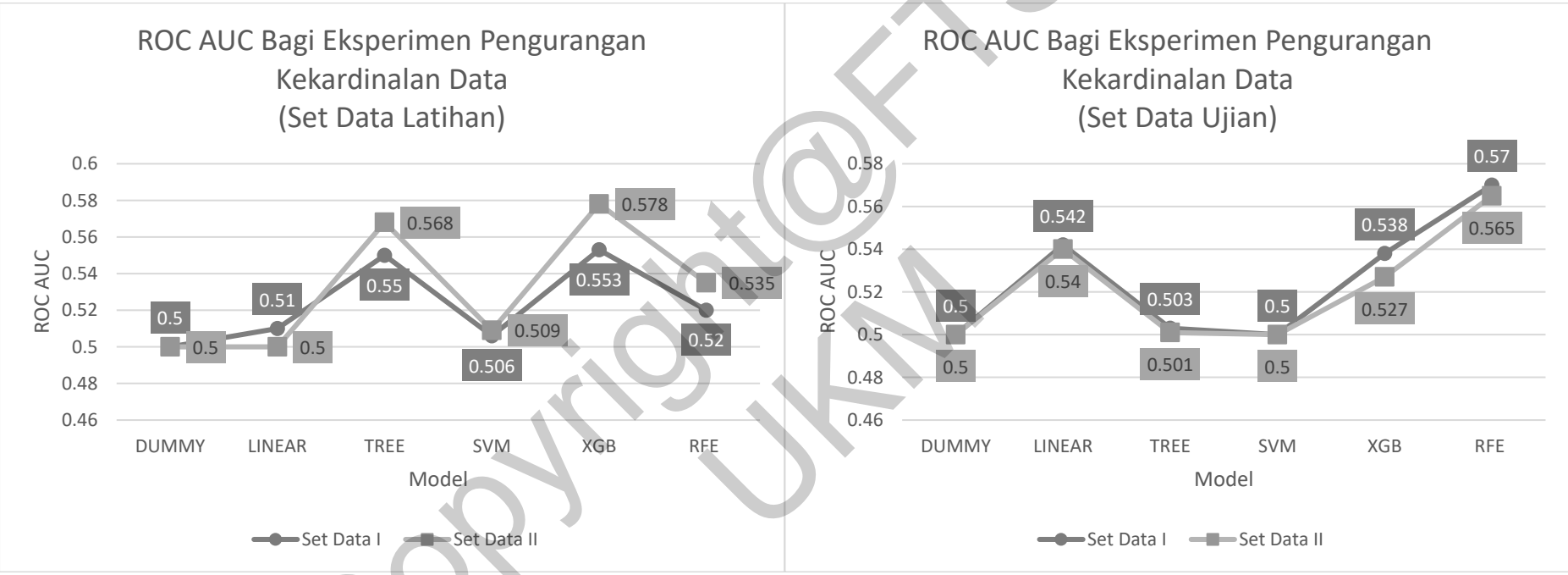
Analisis PCA mendapati pemilihan fitur berdasarkan komponen PCA mendapat peningkatan prestasi pengelasan model yang paling tinggi, diikuti dengan penggunaan teknik transformasi PCA. Namun, kedua-dua cara analisis PCA tidak dapat mencapai prestasi setinggi set data menggunakan fitur terpenting dalam gabungan pemodelan model linear, pepohon keputusan, *XGBoost* dan ensembel Hutan Rawak. Hanya graf ROC AUC ditunjukkan sahaja disebabkan kekangan ruang.



*DUMMY ialah model palsu, LINEAR ialah model linear, TREE ialah pepohon keputusan, SVM ialah mesin vektor sokongan, XGB ialah *XGBoost*, RFE ialah ensembel Hutan Rawak

Rajah 3 ROC AUC Bagi Eksperimen PCA

Pengurangan kerkadinalan data terhadap fitur status perkahwinan, status merokok, dan status penyakit telah menjatuhkan prestasi pengelasan bagi semua model. Hanya graf ROC AUC ditunjukkan sahaja disebabkan kekangan ruang.



*DUMMY ialah model palsu, LINEAR ialah model linear, TREE ialah pepohon keputusan, SVM ialah mesin vektor sokongan, XGB ialah *XGBoost*, RFE ialah ensembel Hutan Rawak

Rajah 4 ROC AUC Bagi Eksperimen Pengurangan Kekardinalan Data

Pengklusteran set data detik garis tapak LRGS TUA hanya dapat membina kluster dengan nisbah tertinggi 7:3 antara kes Ada MCI dengan kes Tiada MCI. Konfigurasi terbaik bagi algoritma pengklusteran ialah 5 pusat-bentuk bagi K-Mean, nilai menegak 6 bagi SOM, dan *epsilon* 6 bagi DBSCAN.

Jadual 1 Nisbah Rekod "MCI" ke Rekod "Tiada MCI" dalam Kluster Bagi K-Mean

Bil	Kluster 0 (MCI/Non)	Kluster 1 (MCI/Non)	Kluster 2 (MCI/Non)	Kluster 3 (MCI/Non)	Kluster 4 (MCI/Non)	Kluster 5 (MCI/Non)	Kluster 6 (MCI/Non)	Kluster 7 (MCI/Non)
2	0.586/0.414	0.444/0.555						
3	0.484/0.516	0.582/0.418	0.462/0.538					
4	0.588/0.412	0.563/0.438	0.477/0.523	0.424/0.580				
5	0.545/0.454	0.530/0.470	0.392/0.608	0.600/0.400	0.472/0.528			
6	0.447/0.553	0.389/0.611	0.439/0.561	0.700/0.300	0.556/0.444	0.550/0.450		
7	0.574/0.426	0.500/0.500	0.300/0.700	0.658/0.342	0.482/0.518	0.377/0.623	0.532/0.468	
8	0.583/0.417	0.500/0.500	0.300/0.700	0.658/0.342	0.481/0.519	0.377/0.623	0.532/0.468	0.333/0.666

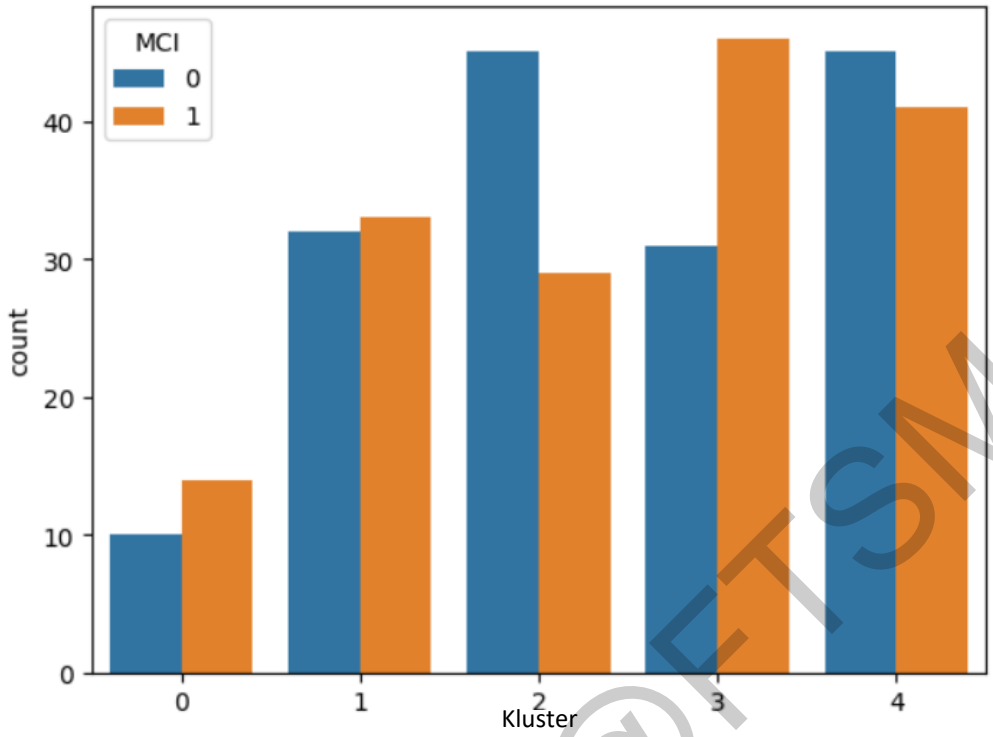
Jadual 2 Nisbah Rekod "MCI" ke Rekod "Tiada MCI" dalam Kluster Bagi SOM

Bil	Kluster 0 (MCI/Non)	Kluster 1 (MCI/Non)	Kluster 2 (MCI/Non)	Kluster 3 (MCI/Non)	Kluster 4 (MCI/Non)	Kluster 5 (MCI/Non)	Kluster 6 (MCI/Non)	Kluster 7 (MCI/Non)
2	0.532/0.468	0.465/0.535						
3	0.601/0.398	0.440/0.560	0.451/0.549					
4	0.462/0.538	0.478/0.522	0.506/0.494	0.552/0.448				
5	0.462/0.538	0.422/0.578	0.468/0.532	0.569/0.431	0.600/0.400			
6	0.583/0.417	0.583/0.417	0.600/0.400	0.340/.660	0.465/0.535	0.481/0.519		
7	0.333/0.666	0.456/0.544	0.472/0.528	0.472/0.528	0.566/0.434	0.621/0.379	0.555/0.444	
8	0.618/0.382	0.550/0.450	0.515/0.485	0.479/0.521	0.515/0.485	0.477/0.523	0.470/0.530	0.333/0.666

Jadual 3 Bilangan Kluster “MCI” dan Kluster “Tiada MCI” dalam DBSCAN

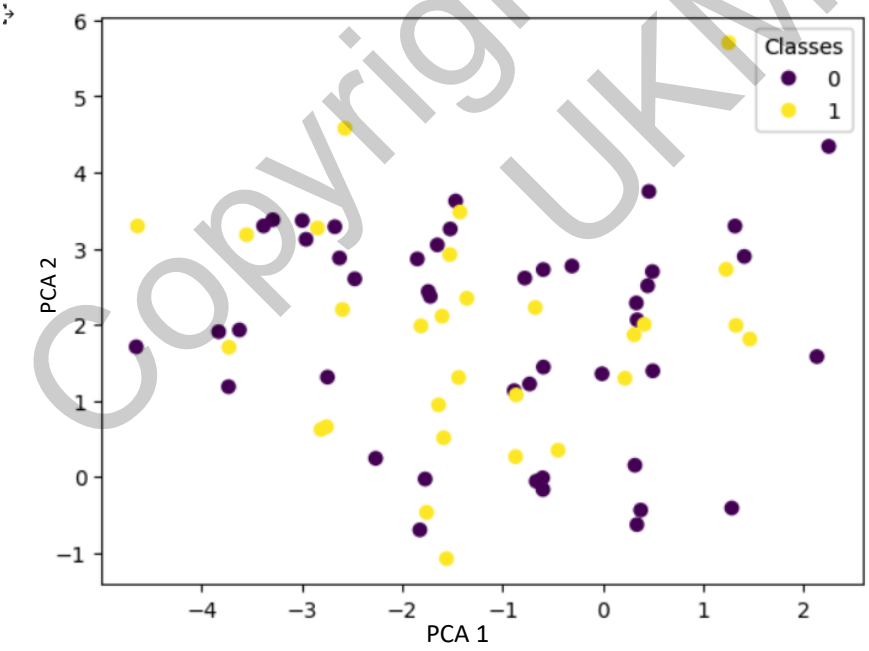
Epsilon	Bilangan titik gangguan	Bilangan Kluster MCI	Bilangan Titik dalam Kluster MCI	Bilangan Kluster Tiada MCI	Bilangan Titik dalam Tiada Kluster MCI	Bilangan Kluster Bercampur	Bilangan Titik dalam Kluster Bercampur
1	326	0	0	0	0	0	0
2	326	0	0	0	0	0	0
3	326	0	0	0	0	0	0
4	322	0	0	1	2	1	2
5	279	6	13	2	6	8	28
6	210	8	26	4	9	2	81
7	128	3	13	2	7	4	178
8	74	1	2	2	6	4	244
9	46	1	2	1	2	3	276
10	22	2	4	1	2	2	298

Kluster MCI mempunyai lebih 60% titik berlabel “MCI”; Kluster Tiada MCI mempunyai lebih 60% titik berlabel “Tiada MCI”; Kluster Bercampur merupakan kluster yang mempunyai 40% - 60% titik MCI atau Tiada MCI.



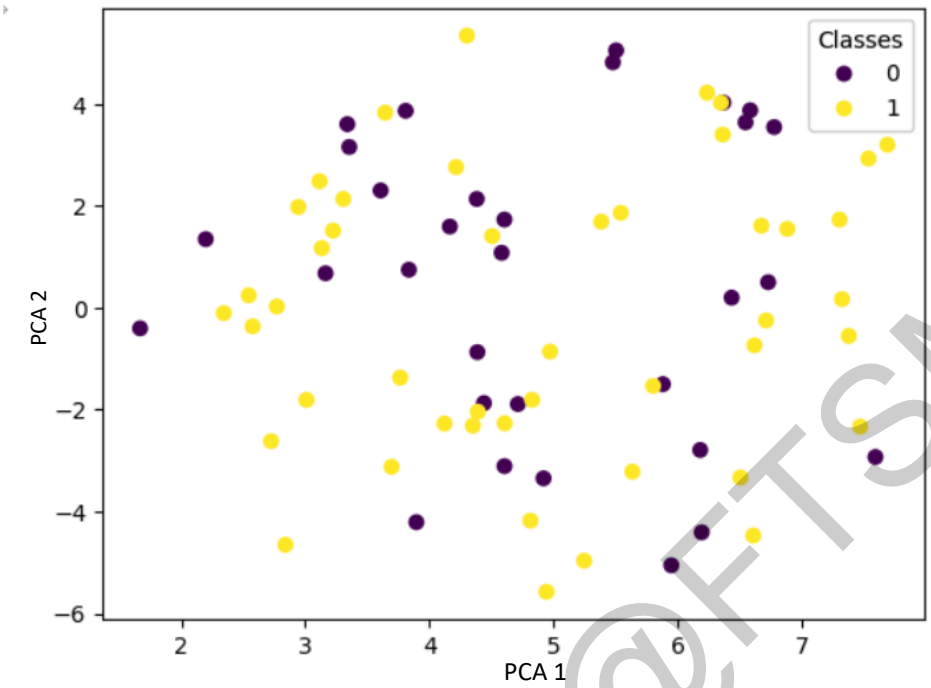
Rajah 21 Gugusan Data dengan Label Menggunakan K-Mean

```
vizSubCluster(output, 2)
```



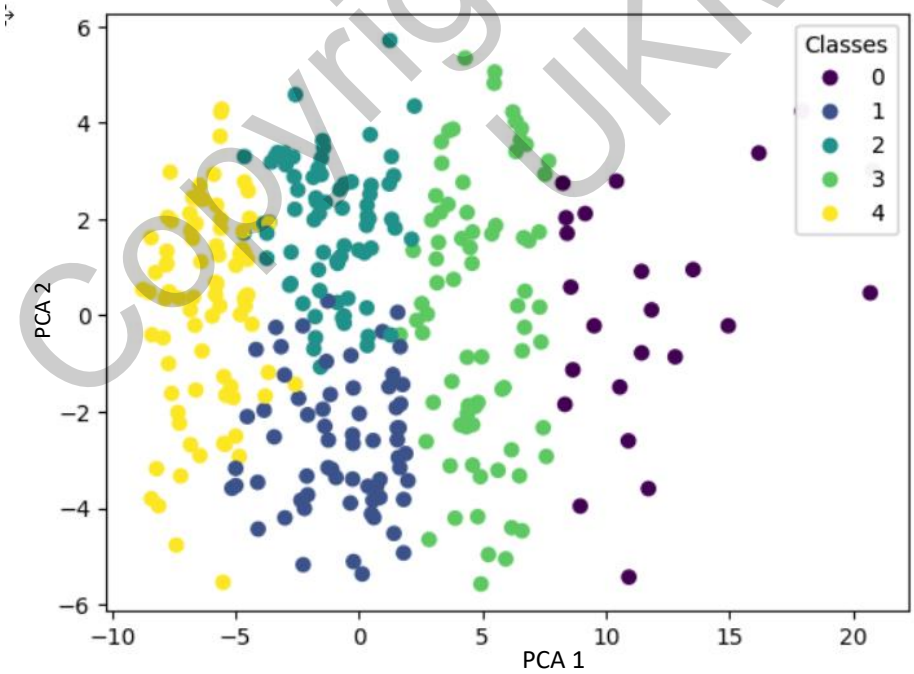
Rajah 22 Plot selerak Gugusan 2 Menggunakan K-Mean

```
vizSubCluster(output, 3)
```

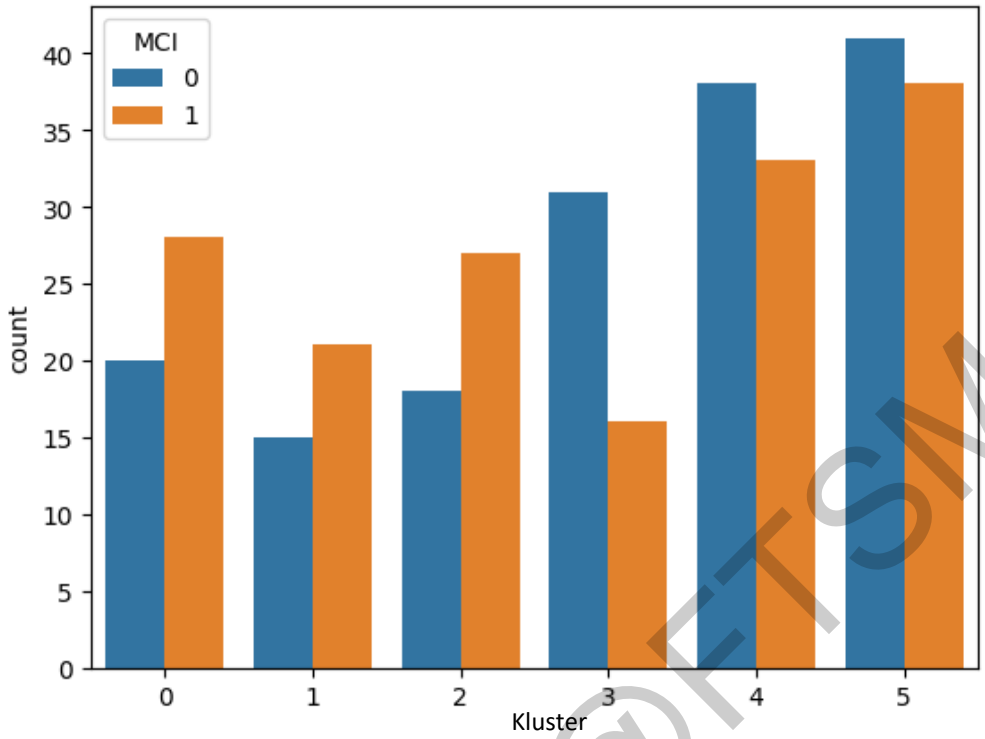


Rajah 23 Plot selerak Gugusan 3 Menggunakan K-Mean

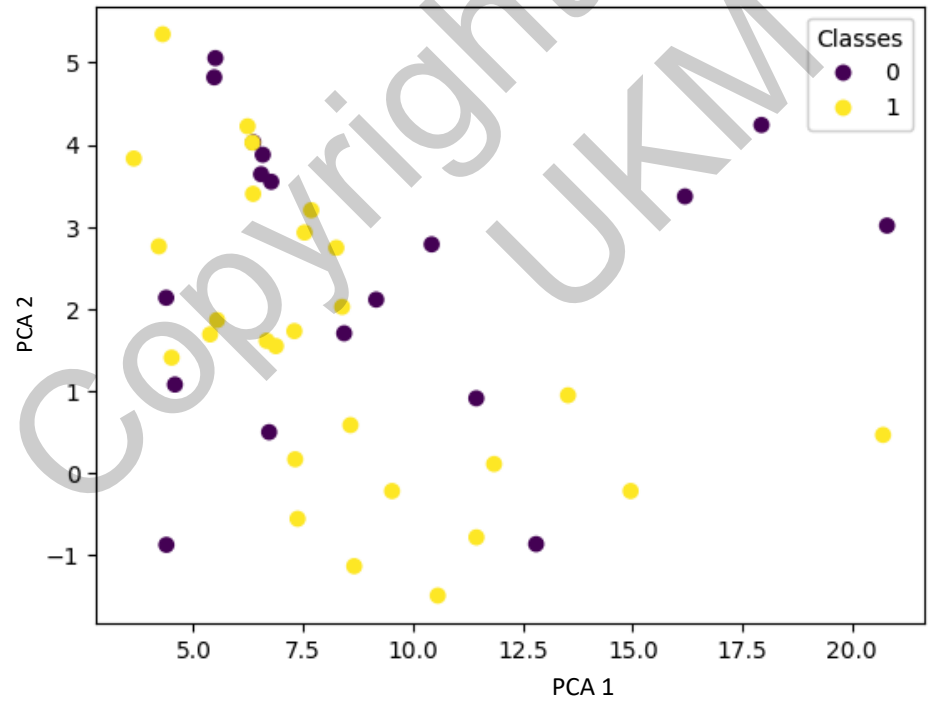
```
vizCluster(kmeans.labels_)
```



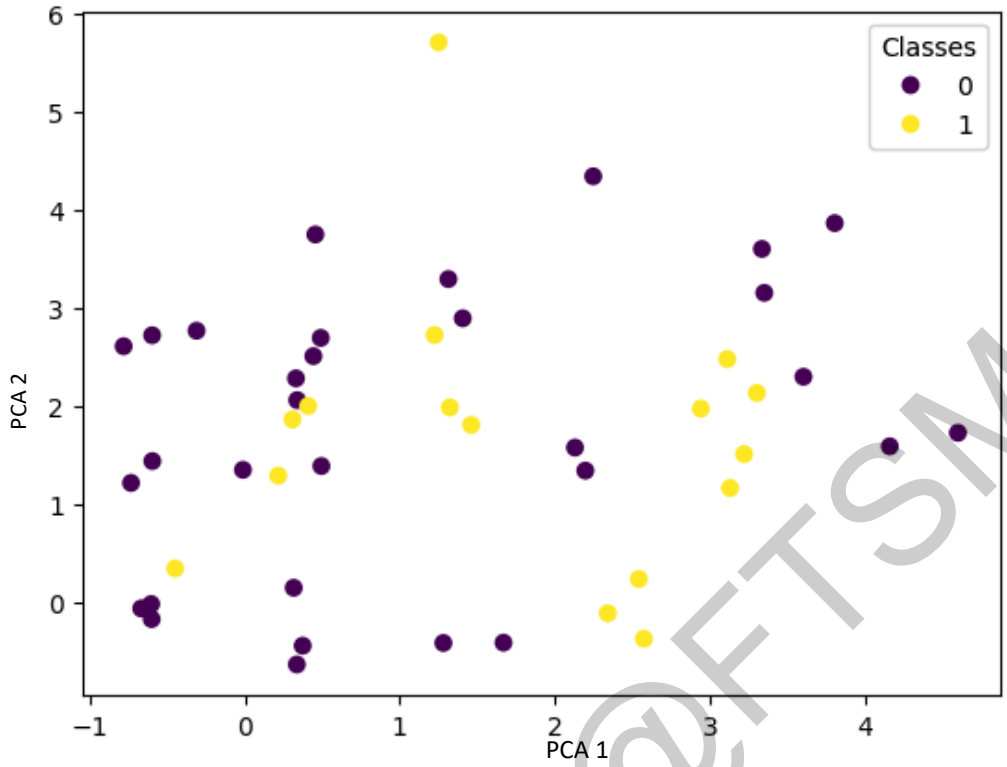
Rajah 24 Plot selerak Kesemua Gugusan Menggunakan K-Mean



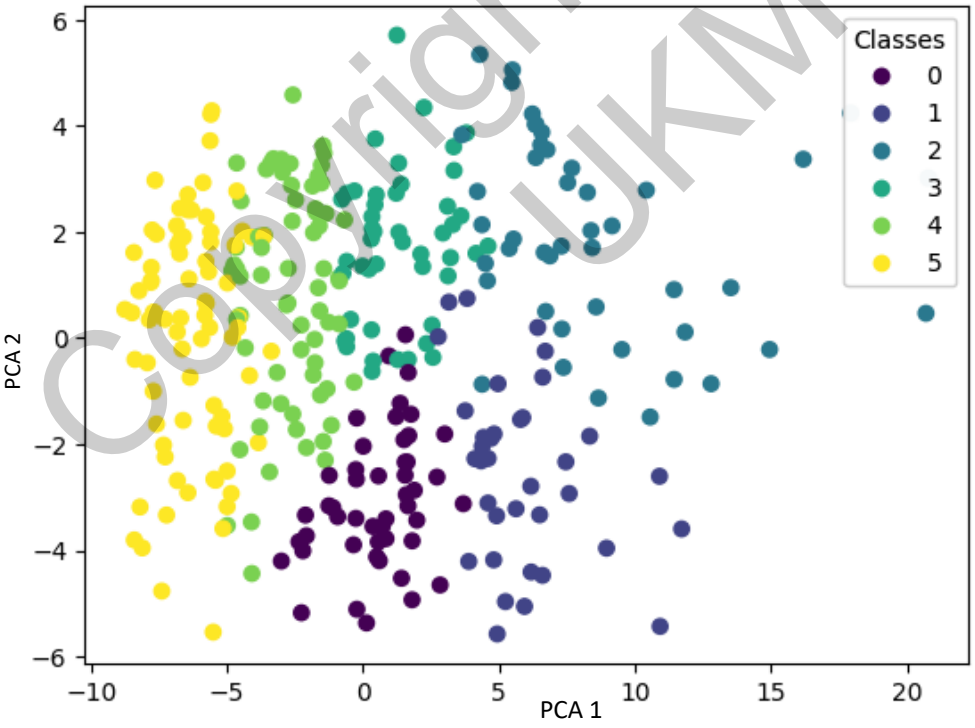
Rajah 25 Gugusan Data dengan Label Menggunakan SOM



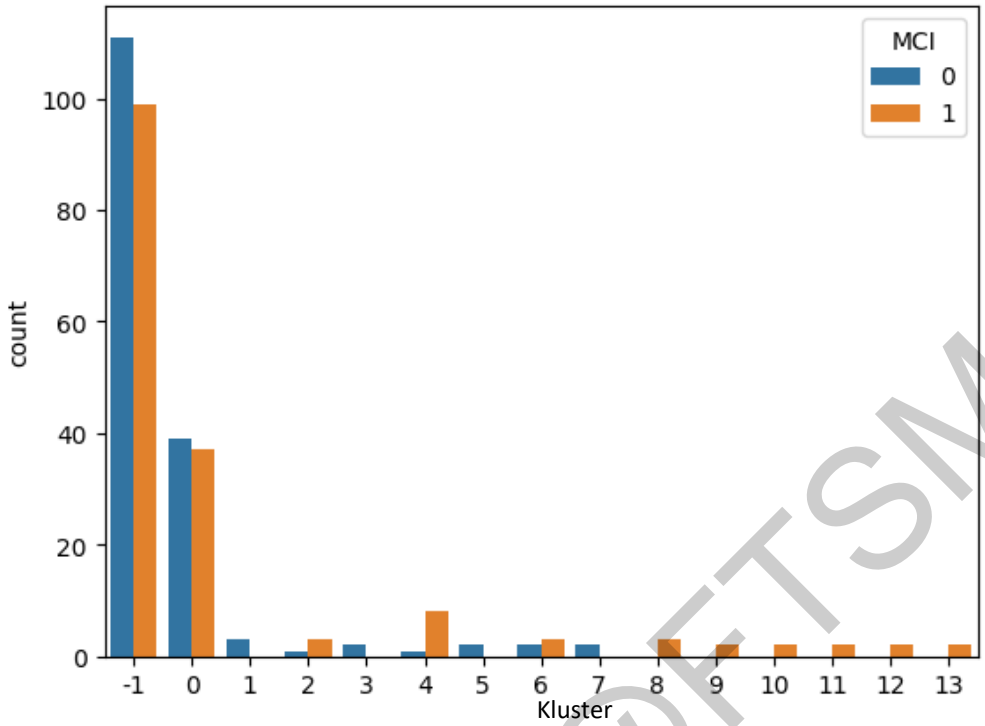
Rajah 26 Plot selerak Gugusan 2 Menggunakan SOM



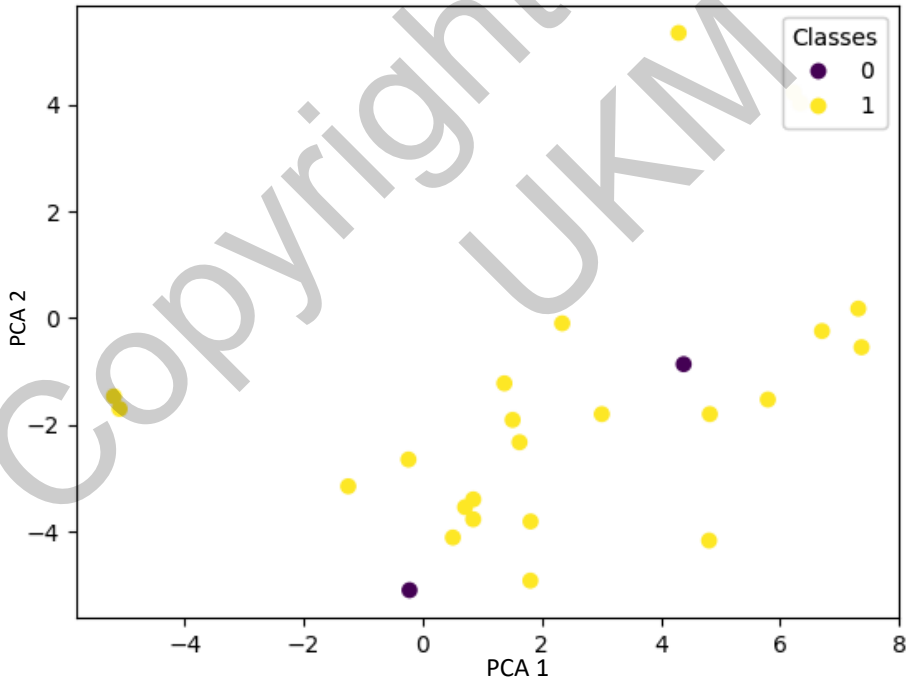
Rajah 27 Plot selerak Gugusan 3 Menggunakan SOM



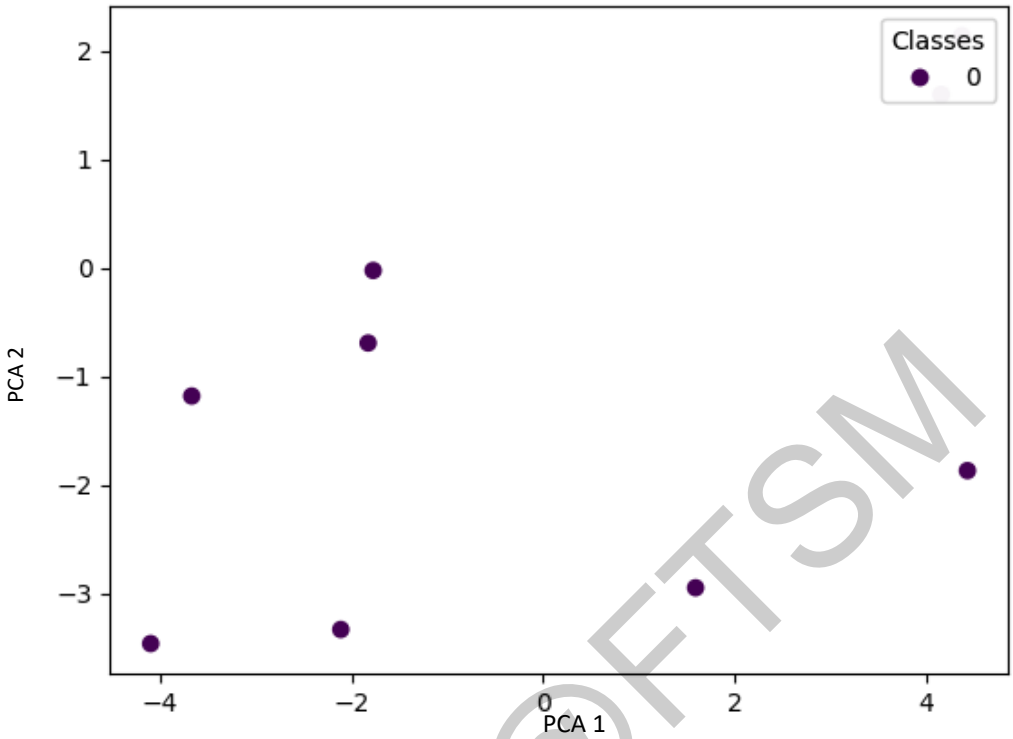
Rajah 28 Plot selerak Kesemua Gugusan Menggunakan K-Mean



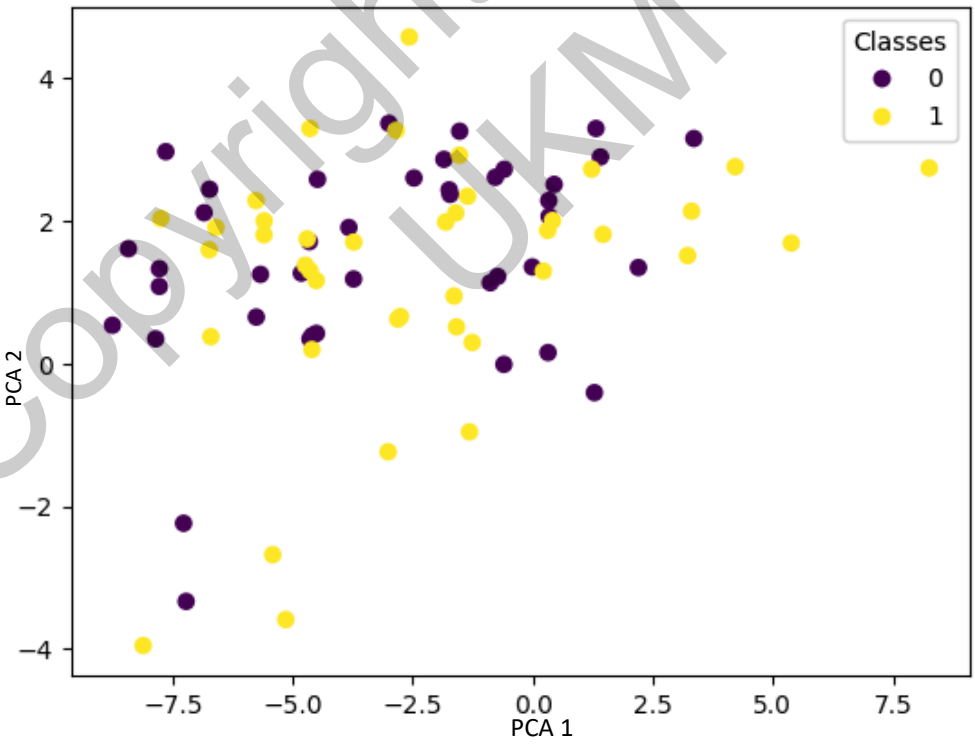
Rajah 29 Gugusan Data dengan Label Menggunakan DBSCAN



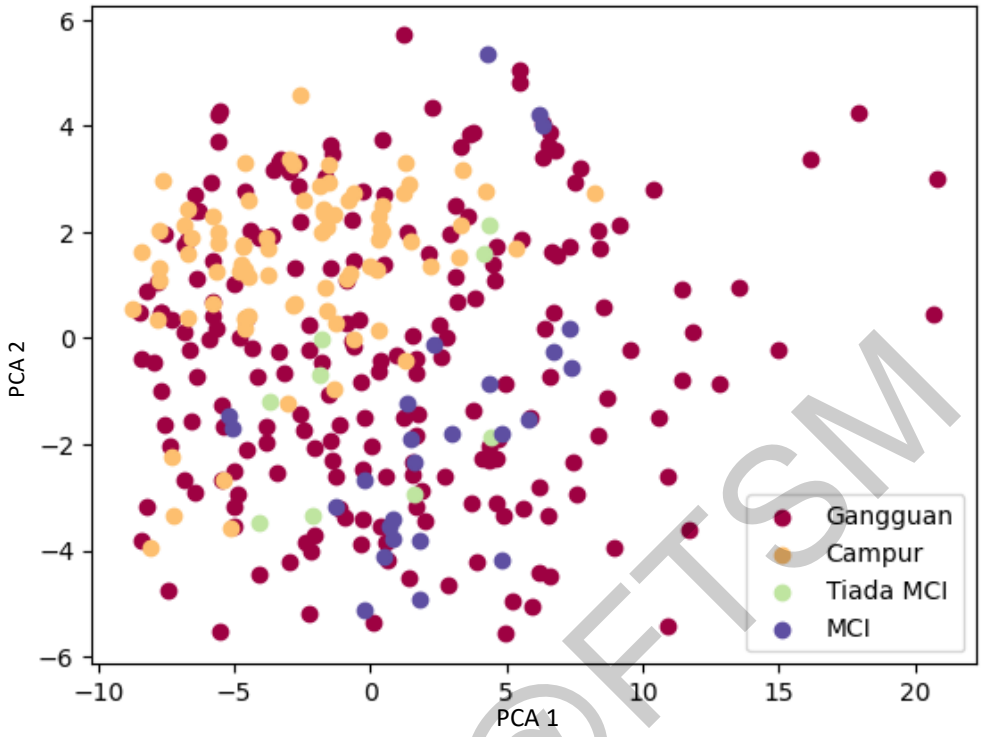
Rajah 30 Plot selerak Titik dalam Kluster “MCI” Menggunakan DBSCAN



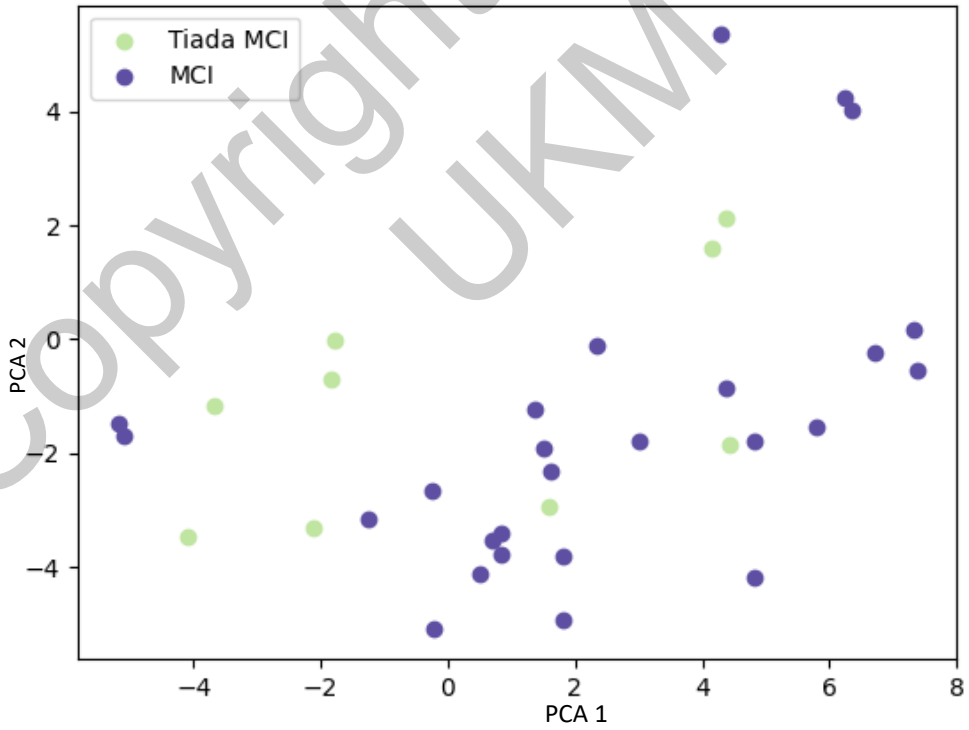
Rajah 31 Plot selerak Titik dalam Kluster “Tiada MCI” Menggunakan DBSCAN



Rajah 32 Plot selerak Titik dalam Kluster “Tiada MCI” Menggunakan DBSCAN



Rajah 33 Plot selerak Kesemua Kluster Menggunakan DBSCAN



Rajah 34 Plot selerak Kluster "MCI" dan "Tiada MCI" Menggunakan DBSCAN

Terdapat beberapa kekangan yang wujud sepanjang kajian. Pertama sekali, ketakmampuan kognitif merupakan suatu fenomena yang amat kompleks, dan boleh dipengaruhi oleh pelbagai faktor dari segi biologi, psikososial, klinikal, dan neurologi. Data TUA yang diambil hanya termasuk data sosial, metrik klinikal, dan gaya hidup mereka. Data daripada domain lain boleh dikutip dan dilombong bagi lebih memahami fenomena ketakmampuan kognitif, seperti darah, komposisi badan, genetik, dan pemakanan.

Keduanya, wujudnya kesukaran untuk membezakan antara rekod berlabel "MCI" dan "Tiada MCI". Label tersebut diberi berdasarkan keputusan ujian kecergasan minda sahaja, dan amat subjektif. Perlunya teknik pengklusteran yang lebih baik untuk mewujudkan kelompok yang lebih jelas antara rekod "MCI" dan "Tiada MCI", seperti dalam kajian oleh Ribeiro dan Zarate (2019). Hal ini juga dapat bantu memudahkan usaha mengenal pasti fitur yang boleh mengelirukan model dan fitur tersembunyi.

Kekangan ketiga, analisis data TUA adalah secara melintang, serta terhad kepada dua detik masa soal selidik. Oleh itu, faktor yang dikenal pasti merupakan petanda MCI yang boleh dikesan pada masa saringan kesihatan, dan bukan faktor seseorang mendapat masalah ketakmampuan kognitif pada jangka masa panjang. Usaha perlombongan data terhadap kajian longitudinal yang merentasi jangka masa yang lebih panjang, serta mempunyai bilangan detik masa soal selidik yang lebih banyak dapat bantu mengesan masalah ketakmampuan kognitif berdasarkan trend kecergasan minda dan kesihatan tubuh badan pada suatu jangka masa.

Kekangan keempat, set data TUA merupakan set data yang boleh mewakili warga emas di Malaysia sahaja. Kajian masa akan datang boleh membandingkan perbezaan trend

ketakmampuan kognitif antara negara akibat geografi, persekitaran, kebudayaan dan sikap masyarakat.

Kesimpulan

Projek ini mendapati algoritma pemodelan yang paling berkesan adalah ensembel Hutan Rawak dengan menggunakan set data undersample dengan 30 fitur paling berpengaruh pada fasa pemodelan data. Metrik tertinggi yang berjaya dicapai ialah 0.531 skor kejituan, 0.657 skor panggil semula, 0.581 ROC AUC, 16.922 logistic loss dan 0.469 min ralat kuasa dua. Usaha pengklusteran data bagi mendapat sampel unggul untuk mewakili kelas Ada MCI dan Tiada MCI gagal menunjukkan hasil akibat terlalu banyak pertindihan antara kedua-dua label data. Pertindihan dalam set data ini perlu ditanangi untuk mencapai prestasi yang lebih baik bagi model pengelasan.

Penghargaan

Terima kasih kepada penyelia saya Assoc. Prof. Dr. Suhaila Zainudin kerana memberi nasihat kepada saya dalam proses menyiapkan kajian ini, serta memberi sokongan ketika saya ada masalah motivasi. Terima kasih kepada ibu saya Dr. Cheong Ai Theng dalam memberi sokongan moral, serta nasihat penyiapan kajian daripada latar pensyarah perubatan.

RUJUKAN

Bakirarar B. & Elhan A. H. 2023. Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research. *Turkiye Klinikleri Journal of Biostatistics*. 15(1):19-29

Coppolino, G., Bolignano, D., Gareri, P. Ruberto C., Andreucci M., Ruotolo G., Rocca M. & Castagna A. 2018. Kidney function and cognitive decline in frail elderly: two faces of the same coin?. International Urology and Nephrology. 50:1505–1510.

Chen T. Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*. 785-794

Dodd J.W. 2015. Lung disease as a determinant of cognitive decline and dementia. Alzheimer's Research & Therapy. 7(1):32.

Engchuan W., Dimopoulos A. C., Tyrovolas S., Caballero F. F., Sanchez-Niubo A., Arndt H., Ayuso-Mateos J. L., Haro J. M., Chatterji S. & Panagiotakos D. B. 2019. Sociodemographic Indicators of Health Status Using a Machine Learning Approach and Data from the Enlglish Longitudinal Study of Aging (ELSA). *Health status using machine learning and ELSA*. 25:1994-2001.

Eshkoor S. A., Hamid T. A., Mun C. Y. 2016. Correlation of cognitive impairment with constipation and renal failure. Sains Malaysiana. 45(9):1357-1361.

Grant A., Aubin M. J., Buhrmann R., Kergoat M. J., Li G. * Freeman E. E. 2022. Visual Impairment, Eye Disease, and 3-Year Cognitive Decline: The Canadian Longitudinal Study on Aging. Ophthalmic Epidemiology. 29(5): 545-53.

Hall M. A. 1999. Feature Selection for Discrete and Numeric Class Machine Learning. Hamilton: University of Waikato.

Horgal A. L., Elliott A. L., Yang S. & Guo Y. 2022. Cross-sectional relationship between pain intensity and subjective cognitive decline among middle-aged and older adults with arthritis or joint conditions: Results from a population-based study. Sage Open Medicine. 10.

Jabatan Perangkaan Malaysia. 2021. *Perangkaan Penting, Malaysia, 2021*. https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=165&bul_id=UDlnQ05GMittVXJWZUVDYUFDcjVTZz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09 [30 November 2022].

Ju Y. J., Lee J. E. & Lee S.Y. 2021. Associations between Chewing Difficulty, Subjective Cognitive Decline, and Related Functional Difficulties among Older People without Dementia: Focus on Body Mass Index. The Journal of Nutrition, Health & Aging. 25(3):347-355.

Kurt P., Yener G. & Oguz M. 2011. Impaired digit span can predict further cognitive decline in older people with subjective memory complaint: a preliminary result. Aging & Mental Health. 15(3):364-9.

Larose D. T. & Larose C. D. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. Edisi ke-2. New Jersey: John Wiley & Sons, Inc.

Lau H., Shahar S., Hussin N., Kamarudin M.Z., Hamid T.A., Mukari S.Z., Rajab N.F., Din N.C., Omar A., Singh D.K., Haron H., Sharif R., Yahya H.M., Fitri A., Manaf Z.A., Mohammed Z. & Ishak W.S. 2019. Methodology approaches and challenges in population-based longitudinal study of a neuroprotective model for healthy longevity. *Geriatr Gerontol Int*. 19(3):233-239.

Lin L. H., Wang S. B., Xu W.Q., Hu Q., Zhang P., Ke Y. F., Huang J. H., Ding K. R., Li X. L., Hou C. L. & Jia F. J. 2022. Subjective cognitive decline symptoms and its association with socio-demographic characteristics and common chronic diseases in the southern Chinese older adults. BMC Public Health. 22(127).

Lipnicki D. M., Makkar S. R., Crawford J. D., Thalamuthu A., Kochan N. A., Lima-Costa M. F., Castro-Costa E., Ferri C. P., Brayne C., Stephan B., Libre-Rodriguez J. J., Llibre-Guerra J. J., Valhuerdi-Cepero A. J., Lipton R. B., Katz M. J., Derby C. A., Ritchie K., Ancelin M., Carriere I., Scarmeas N., Yannakoulia M., Hadjigeorgiou G. M., Lam L., Chan W., Fung A., Guaita A., Vaccaro R., Davin A., Kim K. W., Han J. W., Suh S. W., Riedel-Heller S. G., Roehr S., Pabst A., Boxtel M. V., Kohler S., Deckers K., Ganguli M., Jacobsen E. P., Hughes T. F., Anstey K. J., Cherbuin N., Haan M. N., Aiello A. E., Dang K., Kumagai S., Chen T., Narazaki K., Ng T. P., Gao Q., Nyunt M. S. Z., Scazufca M., Brodaty H., Numbers K., Trollor J. N., Meguro K., Yamaguchi S., Ishii H., Lobo A., Lopez-Anton R., Santabarbara J., Leung Y., Lo J. W., Popovic G. && Sachdev P. S. 2019. Determinants of cognitive performance and decline in 20 diverse ethno-regional groups: A COSMIC collaboration cohort study. PLoS Medicine. 16(7).

Liu M., He P., Zhou C., Zhang Z., Zhang Y., Li H., Ye Z., Wu Q., Yang S., Zhang Y., Liu C. & Qin X. 2022. Association of waist-calf circumference ratio with incident cognitive impairment in older adults. The American Journal of Clinical Nutrition.115(4):1005-1012.

Majlis Ekonomi dan Sosial Pertubuhan Bangsa-Bangsa Bersatu. 2017. *World population prospects: The 2017 revision.* https://www.un.org/development/desa/publications/world-population-prospects-the-2017-revision.html [30 November 2022].

Mei F., Dong S., Li J., Xing D. & Lin J. 2023. Preference of musculoskeletal pain treatment in middle-aged and elderly chinese people: a machine learning analysis of the China health and retirement longitudinal study. *BMC Musculoskeletal Disorders*. 24:528

Murukesu R. R., Singh D. K. A. & Shahar S. 2018. Faktor Risiko Inkontinens Urinari dalam Kalangan Warga Emas Dikomuniti. Jurnal Sains Kesihatan Malaysia Isu Khas. 2018: 227.

Nair A. K., Van Hulle C. A., Bendlin B. B., Zetterberg H., Blennow K., Wild N., Kollmorgen G., Suridjan I., Busse W. W. & Rosenkranz M. 2022. Asthma amplifies dementia risk: Evidence from CSF biomarkers and cognitive decline. Alzheimer's & Dementia: Translational Research & Clinical Interventions. 8(1).

Newsom J. T., Jones R. N. & Hofer S. M. 2012. *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. London: Routledge/Taylor & Francis Group.

Noronha M. D. M., Nobre C. N., Song M. A. J. & Zarate L. E. 2022. Interpreting the Human Longevity Profile Through Triadic Rules - A Case Study Based on the ELSA-UK Longitudinal Study. *Studies in health and technology informatics*. 290:782-786.

Ooi T. C., Singh D. K. A., Shahar S., Sharif R., Rivan N. F. M., Meramat A. & Rajab N. F. 2022. Higher Lead and Lower Calcium Levels Are Associated with Increased Risk of Mortality in Malaysian Older Population: Findings from the LRGS-TUA Longitudinal Study. International Journal of Environmental Research and Public Health. 19(12): 6955.

Paes, B. C., Plastino, A., & Freitas, A. A. 2013. In Proceedings of the Symposium on Knowledge Discovery, Mining and Learning. *Selecao de Fituros Aplicada a Classificacao Hierarquica*. 1–8.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplass J., Passos A., Cournapeau D., Brucher M., Perrot M. & Duchesna E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12:2825-2840.

Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V. & Fratiglioni, L. 2014. Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine*. 275:214–228.

Pomsuwan, T. & Freitas, A. A. 2017. Feature Selection for the Classification of Longitudinal Human Ageing Data. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 739-746.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann Publishers Inc.

Ribeiro, C.E., Brito, L.H.S., Nobre, C.N., Freitas, Alex A. & Zarate, L.E. 2017. A revision and analysis of the comprehensiveness of the main longitudinal studies of human ageing for data mining research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7 (3). e1202.

Ribeiro C. E. & Zarate L. E. 2016. Data preparation for longitudinal data mining: a case study on human ageing. *Journal of Information and Data Management*. 7:116-129.

Ribeiro C. E. & Zarate L. E. 2019. Classifying longevity profiles through longitudinal data mining. *Expert Systems With Applications*. 117: 75-89.

Saito T., Yamada T., Miyauchi Y., Emoto N. & Okajima F. 2022. Use of the Japanese Version of the Montreal Cognitive Assessment to Estimate Cognitive Decline in Patients Aged 75 Years or Older with and without Type 2 Diabetes Mellitus. Journal of Nippon Medical School. 89(2):196-202.

Su D., Zhang X., He K. & Chen Y. 2021. Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. *Journal of Affective Disorders*. 282:289-298.

Szlejf C., Suemoto C. K., Janovsky C. C. P. S., Bertola L., Barreto S. M., Lotufo P. A. & Benseñor I. M. 2021. Subtle Thyroid Dysfunction Is Not Associated with Cognitive

Decline: Results from the ELSA-Brasil. Journal of Alzheimer's Disease: JAD. 81(4):1529-1540.

Tableau. Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data. https://www.tableau.com/learn/articles/what-is-data-cleaning [2021-10-17]

Uzoigwe C. E., O'Leary L., Nduka J., Sharma D., Melling D., Simmons D. & Barton S. 2020. Factors associated with delirium and cognitive decline following hip fracture surgery. The Bone & Joint Journal. 102-B(12).

Vanoh D., Shahar S., Normah D., Omar A., Vyrn C., Razali R., Ibrahim R. & Hamid T. 2016. Predictors of poor cognitive status among older Malaysian adults: baseline findings from the LRGS TUA cohort study. Aging Clinical & Experimental Research 29(2):173-182.

Vanoh D., Shahar S., Yahya H. M., Che Din N., Mat Ludin A. F., Ajit Singh D. K., Sharif R. & Rajab N. F. 2021. Dietary Supplement Intake and Its Association with Cognitive Function, Physical Fitness, Depressive Symptoms, Nutritional Status and Biochemical Indices in a 3-Year Follow-Up Among Community Dwelling Older Adults: A Longitudinal Study. Clin Interventions in Aging. 16:161-175.

Vitorino L. M., Lucchetti A. L. G. & Lucchetti G. 2022. The role of spirituality and religiosity on the cognitive decline of community-dwelling older adults: a 4-year longitudinal study. Aging & Mental Health.

Wang H., Yang C. & Yao Y. 2022. Familial factors, depression and cognitive decline: A longitudinal mediation analysis based on latent growth modeling (LGM). International Journal of Methods in Psychiatric Research. 31(2).

Yang H. & Bath P. A. 2020. The Use of Data Mining Methods for the Prediction of Dementia: Evidence From the English Longitudinal Study of Aging. *IEEE Journal of Biomedical and Health Informatics*. 24(2):345-353.

Eileen Tong Hui Guan (A180693)
Suhaila Zainudin
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia