Asia-Pacific Journal of Information Technology and Multimedia 13(1) 2024: 72 - 92
https://doi.org/10.17576/apjitm-2024-1301-06

Combining Cluster Quality Index and Supervised Learning to Predict Students' Academic Performance

Gabungan Indeks Kualiti Kelompok dan Pembelajaran Terselia Untuk Meramal Prestasi Akademik Pelajar

Suhaila Zainudin, Rapi'ah Ibrahim, Hafiz Mohd Sarim

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor

suhaila.zainudin@ukm.edu.my

Received 18 September 2023
Accepted 4 December 2023, Available online 1 June 2024

ABSTRACT

Predicting students' academic performance can help the institution to take timely action, such as planning intervention measures to improve students' academic achievement. This study aims to identify the main factors contributing to the postgraduate student's academic performance. Preliminary predictions can be made to avoid student dropouts, especially for students studying at the postgraduate level. The results obtained from this study are significant for facilitating the institution in decision-making and formulating the best strategies for the primary stakeholder (students). This study employs a combination of data mining tasks, such as clustering and classification, to undertake the prediction task. First, the approach performed clustering with K-Means algorithm to identifies different student groups. Then, the clusters were evaluated with cluster quality indexes, namely, the Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index, to determine the best clusters. The best number of clusters is selected based on the Silhouette Coefficient score because the uniformity for this coefficient is between -1 and 1. The best cluster is further analysed using classification to predict students' academic performance. Three classification algorithms have been selected: Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT). The results show that the LR model best predicts students' academic performance levels compared to SVM and DT.

Keywords**:** Educational Data Mining (EDM), academic performance, postgraduate, Association Rule Mining, Clustering, Classification.

ABSTRAK

Meramal prestasi akademik pelajar boleh membantu sesebuah institusi mengambil tindakan tepat pada masanya seperti merancang langkah yang sesuai untuk meningkatkan pencapaian akademik pelajar. Kajian ini bertujuan untuk mengenal pasti faktor utama yang menyumbang kepada prestasi akademik pelajar pasca siswazah. Ramalan awal boleh dibuat bagi mengelakkan pelajar tercicir terutamanya bagi pelajar yang belajar di peringkat pasca siswazah. Keputusan yang diperolehi daripada kajian ini adalah signifikan untuk memudahkan

institusi dalam membuat keputusan dan merangka strategi terbaik untuk pemegang taruh utama (pelajar). Kajian ini menggunakan gabungan tugas perlombongan data seperti pengelompokan dan pengelasan untuk melaksanakan tugas ramalan. Pertama, pendekatan yang dilakukan pengelompokan dengan algoritma K-Means terutamanya mengenal pasti kumpulan pelajar yang berbeza. Kemudian, kluster dinilai dengan indeks kualiti kluster iaitu; Pekali Siluet, Indeks Calinski-Harabasz dan Indeks Davies-Bouldin, untuk menentukan kelompok terbaik. Bilangan kluster terbaik dipilih berdasarkan skor Pekali Siluet kerana keseragaman bagi pekali ini adalah antara -1 hingga 1. Kluster terbaik kemudiannya dianalisis menggunakan klasifikasi untuk meramal prestasi akademik pelajar. Tiga algoritma pengelasan telah dipilih; Regresi Logistik (LR), Mesin Vektor Sokongan (SVM) dan Pokok Keputusan (DT). Keputusan menunjukkan bahawa model LR adalah yang terbaik untuk meramal tahap prestasi akademik pelajar, berbanding SVM dan DT. Atribut kerap muncul dalam keputusan analisis adalah JANTINA, PROGRAM_PENGAJIAN, STATUS_KAHWIN, BIL_PELAJAR_SELIA, SEM_SESI_ HENTI, STATUS_BEKERJA, STATUS_TAJAAN dan SEM_SESI_MASUK. Atribut ini boleh diselidik dengan lebih lanjut untuk mendapatkan faktor lebih jelas untuk prestasi akademik pelajar.

Kata kunci**:** Perlombongan Data Pendidikan (EDM), prestasi akademik, pascasiswazah, Perlombongan Petua Sekutuan, Pengelompokan, Pengelasan.

PENGENALAN

Pendidikan memainkan peranan utama bagi pertumbuhan ekonomi dan pembangunan sesebuah negara. Dalam ekonomi global masa ini, kejayaan sesebuah negara amat bergantung pada ilmu pengetahuan, kemahiran, dan kompetensi yang dimiliki rakyat (Kementerian Pendidikan Malaysia 2013). Sejak beberapa tahun kebelakangan ini, semakin banyak kajian dibuat menggunakan perlombongan data bagi mengenal pasti faktor mempengaruhi prestasi akademik pelajar khususnya yang berada di institusi pengajian tinggi. Bidang kajian ini dikenali sebagai Perlombongan Data Pendidikan (Bakhshinategh et al. 2017). Kini, kajian yang menggunapakai teknik perlombongan data pendidikan dalam ramalan prestasi pelajar dan lain-lain isu berkaitan pendidikan semakin mendapat perhatian penyelidik (Zhang 2021).

Perlombongan Data Pendidikan telah menjadi alat yang berkesan untuk menyelidik hubungan tersembunyi di dalam data pendidikan dan meramal pencapaian pelajar (Yağcı 2022). Mana institusi pengajian mempunyai sistem maklumat pelajar yang menyimpan sejumlah data yang besar berkaitan pelajar iaitu demografi pelajar, kursus, pengajar, kehadiran pelajar, gred pencapaian pelajar dan lain-lain. Maklumat daripada sistem ini sangat berpotensi untuk digunakan dalam Perlombongan Data Pendidikan bagi meramal prestasi pelajar dari berbagai peringkat pengajian. Sebagai contoh, Mohd Khairy et al (2018) telah menggunakan analitik data iaitu visualisasi untuk menyerlahkan hubungan kuat di antara ujian aptitud tahun 6 (sekolah rendah) dengan potensi keputusan SPM yang cemerlang (sekolah menenngah). Manakala, Ali et al. (2022) pula mengkaji corak pembelajaran pelajar di dalam persekitaran atas talian. Kajian beliau telah mengenal pasti peranan perlombongan data pendidikan dalam mengekstrak corak tersebut khasnya untuk domain sistem tutoran pintar atau ITS.

Kajian ini pula menyumbang dalam mengenal pasti faktor utama yang mempengaruh tahap prestasi akademik pelajar. Seterusnya, ramalan awal boleh dibuat bagi mengawal kadar keciciran pelajar khususnya bagi pelajar di peringkat pascasiswazah. Hasil kajian ini sangat penting bagi memudahkan pihak berkepentingan untuk membuat keputusan dan merangka

strategi terbaik untuk kecemerlangan pelajar sebagai pemegang taruh. Seksyen seterusnya membincangkan kajian lepas tentang faktor yang mempengaruhi ramalan prestasi akademik pelajar.

FAKTOR MEMPENGARUHI RAMALAN PRESTASI AKADEMIK PELAJAR

Menurut Alyahyan & Dustegor (2020), pencapaian akademik terdahulu dan demografi pelajar merupakan dua faktor utama dibentangkan dalam 69% kertas kajian yang berkaitan. Pemerhatian ini selari dengan keputusan yang diperolehi daripada kajian lepas yang menyatakan bahawa gred penilaian dalaman dan Purata Nilai Gred Kumulatif (PNGK) merupakan faktor utama digunakan untuk meramal prestasi pelajar dalam Perlombongan Data Pendidikan (Shahiri et al. 2015), di mana lebih 40% kajian menyatakan bahawa pencapaian akademik terdahulu merupakan faktor paling penting. Pencapaian akademik terdahulu adalah sebagai sejarah pelajar dalam bentuk gred, atau apa sahaja indikator prestasi akademik yang pelajar perolehi sebelum ini. Sama ada data di peringkat sekolah menengah, pra-universiti atau data universiti yang membantu memahami konsistensi prestasi pelajar tersebut (Mesaric & Sebalj 2016). Selain itu, keputusan ujian kemasukan ke universiti juga boleh digunakan sebagai faktor pengaruh (Mesaric & Sebalj 2016; Aluko et al. 2018).

Data universiti pula terdiri daripada data gred yang diperolehi pelajar dari awal pengajian. Data ini termasukl Purata Nilai Gred (PNG) mengikut semester, PNGK (Almarabeh 2017; Hamoud et al. 2018; Mueen et al. 2016), markah kerja kursus (Almarabeh 2017; Hamoud et al. 2018; Mueen et al. 2016; Sivasakthi 2017), markah tugasan termasuk tugasan-tugasan makmal, rekod kehadiran pelajar (Almarabeh 2017; Mueen et al. 2016), dan juga markah kuiz pelajar (Almarabeh 2017).

Demografi pelajar pula menunjukkan keputusan yang berbeza-beza. Beberapa kajian menyatakan data tersebut memberi impak terhadap kejayaan pelajar, contohnya jantina (Hamoud et al. 2018; Sivasakthi 2017), umur dan status sosioekonomi (Hamoud et al. 2018; Mueen et al. 2016), bangsa atau etnik (Ahmad et al. 2015), dan latar belakang ibu bapa pelajar (Hamoud et al. 2018) dilaporkan penting, manakala kajian lain menunjukkan sebaliknya, khususnya jantina (Almarabeh 2017).

Beberapa faktor yang berkaitan dengan persekitaran pelajar didapati agak memberi kesan, seperti jenis program yang diikuti (Hamoud et al. 2018), kaedah pembelajaran (Mueen et al. 2016; Sivasakthi 2017), dan tempoh semester (Mesaric & Sebalj 2016). Kebanyakan kajian juga menggunakan maklumat daripada aktiviti e-pembelajaran seperti bilangan log masuk pelajar, bilangan pelajar memasuki ruang diskusi, dan jumlah masa pelajar melihat bahan pembelajaran yang disediakan (Hamoud et al. 2018). Aktiviti e-pembelajaran ini walaupun memberi impak yang kecil, tetapi masih dianggap penting.

Selain itu, kejayaan pelajar turut dikaitkan dengan faktor psikologi seperti minat dan sikap pelajar. Beberapa kajian menunjukkan bahawa minat pelajar, tahap tekanan dan kebimbangan yang dihadapi pelajar, aspek pengurusan diri dan masa pelajar (Hamoud et al. 2018), sikap terhadap pembelajaran (Hamoud et al. 2018; Mueen et al. 2016), dan motivasi diri (Mueen et al. 2016) menunjukkan antara faktor yang mempengaruhi kejayaan pelajar.

Menurut Alban & Mauricio (2019), perlombongan data yang digunakan untuk membuat ramalan keciciran pelajar universiti, diklasifikasikan kepada kepintaran buatan dan statistik.

Bagi kepintaran buatan, hampir 79% kertas kajian menggunakan algoritma pengelas Pohon Keputusan atau *Decision Tree* (DT). Algoritma ini dipilih berdasarkan fleksibiliti untuk memproses data dalam bentuk numerik (*numerical*), dan kategori (*categorical*), serta menghasilkan output yang mudah ditafsir. Selain itu, algoritma ini menunjukkan peratus ketepatan yang tinggi (Yasmin 2013). Algoritma pengelas Pohon Keputusan ID3 berkesan dalam pengelasan data berdasarkan daftar sejarah pelajar dan algoritma ini dikatakan lebih sensitif berbanding algoritma yang lain (Sivakumar et al. 2016). Algoritma pengelas NN atau *Neural Network* dan SVM (*Support Vector Machine*) pula menjadi algoritma kedua tertinggi digunakan kerana algoritma ini dikatakan baik dalam menyelesaikan masalah pengelasan (Liang et al. 2016), ringkas dan mudah difahami (Sivakumar et al. 2016).

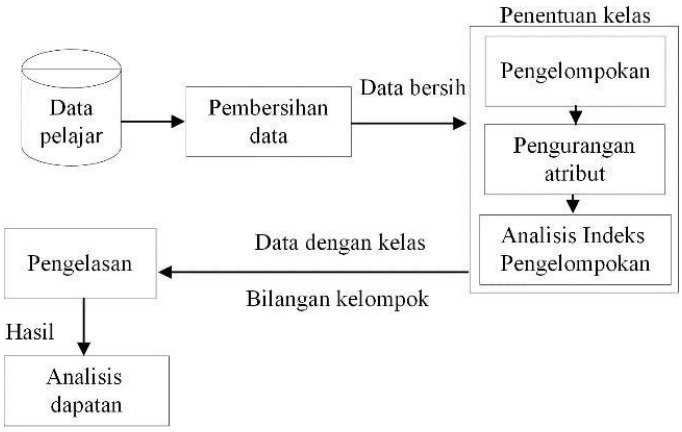
Dalam konteks lain, Ashraf et al. (2022) telah mencadangkan model berasaskan regresi untuk meramal nilai indeks prestasi utama atau KPI (*Key Performance Index*) bagi berbagai indikator pencapai sesuatu organisasi seperti bilangankemasukan pelajar dan bilangan pelajar bergraduasi. KPI sebegini penting untuk mencorak hala tuju jangka panjang sesebuah organisasi.

Kajian Zainal Rafit et al. (2021) berkisar tentang masalah ketidakhadirancalon STAM (Sijil Tinggi Agama Malaysia) yang tidak hadir peperiksaan awam berdasarkan analisis ke atas data untuk sepuluh tahun dari 2007 hingga 2016. Hasil kajian mendapati bahawa gred yang diperoleh untuk subjek Bahasa Inggeris, Matematik dan Sains dalam keputusan peperiksaan awam tingkatan lima iaitu Sijil Pelajaran Malaysia adalah faktor utama bagi menarik diri daripada peperiksaan STAM untuk seseorang calon. Faktor ini seterusnya diwakili dalam bentuk model berasaskan petua. Penilaian ke atas model membuktikan bahawa model berpotensi untuk meramal bilangan calon yang mungkin tidak hadir peperiksaan.

Dapatan dari kajian lepas yang telah dibincangkan memberi halatuju dalam pemilihan metodologi dan teknik penyelesaian bagi kajian ini. Seterusnya, seksyen berikut memperihalkan metodologi kajian yang melibatkan gabungan indeks kualiti pengelompokan dan pembelajaran terselia untuk meramal prestasi akademik pelajar.

METODOLOGI

Metodologi kajian yang digunakan adalah berasaskan metodologi Penemuan Pengetahuan dalam Pangkalan Data atau *Knowledge Discovery in Database* (KDD). KDD merupakan kaedah untuk mengekstrak pengetahuan seperti corak, hubungan atau ramalan daripada data (Wirawati dan Azuraliza, 2018). Rajah 1 menunjukkan metodologi kajian berasaskan KDD yang digunakan dalam kajian ini. Seterusnya, setiap fasa di dalam rajah diterangkan mengikut turutan.



RAJAH 1. Metodologi Kajian

FASA PEMBERSIHAN DATA

Set data pelajar yang digunakan mengandungi 17,967 sampel atau rekod pelajar, iaitu sebanyak 4,960 rekod pelajar pascasiswazah keciciran, dan juga sebanyak 13,007 rekod pelajar pascasiswazah yang telah bergraduat untuk 6 tahun. Pelajar keciciran ini diklasifikasi sebagai pelajar yang telah diberi status berhenti sama ada Diberhentikan, Gagal Peperiksaan/Pengajian atau Menarik Diri. Jadual 1 menunjukkan keterangan atribut bagi set data berkenaan yang mempunyai 32 atribut berjenis nominal atau numerik.

JADUAL 1. Keterangan Atribut Bagi Set Data Pelajar Pascasiswazah

Bil	Nama Atribut	Jenis Data	Keterangan
1	ID	Numerik	Nombor siri pelajar
2	SESI_HENTI	Nominal	Sesi berhenti
3	SEM_HENTI	Nominal	Semester berhenti
4	STATUS	Nominal	Status pengajian
5	SEBAB_HENTI	Nominal	Status sebab berhenti
6	SESI_MASUK	Nominal	Sesi mula daftar pengajian
7	SEM_MASUK	Nominal	Semester mula daftar pengajian
8	FAKULTI	Nominal	Nama fakulti atau institut
9	TAHAP_PENGAJIAN	Nominal	Tahap atau peringkat pengajian
10	PROGRAM_PENGAJIAN	Nominal	Nama program pengajian
11	JENIS_PENDAFTARAN	Nominal	Jenis pendaftaran
12	MOD_PENGAJIAN	Nominal	Jenis atau mod pengajian
13	TARIKH_PEMBENTANGAN	Nominal	Tarikh pembentangan cadangan penyelidikan
14	BIL_PENERBITAN	Numerik	Bilangan penerbitan
15	BIL_SEM_TANGGUH	Numerik	Bilangan semester yang ditangguh
16	STATUS_TAJAAN	Nominal	Status penajaan termasuk Pembantu Penyelidik Siswazah (GRA)
17	BIL_HADIR_SIDANG	Numerik	Bilangan menghadiri persidangan
18	JUM_AMARAN	Nominal	Jumlah peringatan/amaran Laporan Kemajuan Calon
19	HADIR_LATIHAN	Numerik	Kehadiran latihan 2018

20	BIL_PENYELIA	Numerik	Bilangan penyelia
21	ID_PENYELIA	Nominal	ID penyelia utama
22	NAMA_PENYELIA	Nominal	Nama penyelia utama
23	BIL_PELAJAR_SELIA	Numerik	Bilangan pelajar yang diselia (penyelia utama)
24	BIL_PELAJAR_GAGAL	Numerik	Bilangan pelajar yang gagal (penyelia utama)
25	PNGK_MASUK	Numerik	PNGK semasa kemasukan
26	PNGK_HENTI	Numerik	PNGK semasa berhenti
27	JANTINA	Nominal	Jantina
28	STATUS_KAHWIN	Nominal	Status perkahwinan
29	BIL_ANAK	Numerik	Bilangan anak
30	STATUS_BEKERJA	Nominal	Status bekerja (jika ada)
31	STATUS_KESIHATAN	Nominal	Status kesihatan
32	STATUS_KECACATAN	Nominal	Status kecacatan
33	STATUS_WARGANEGARA	Nominal	Status warganegara
34	NEGARA_ASAL	Nominal	Negara asal

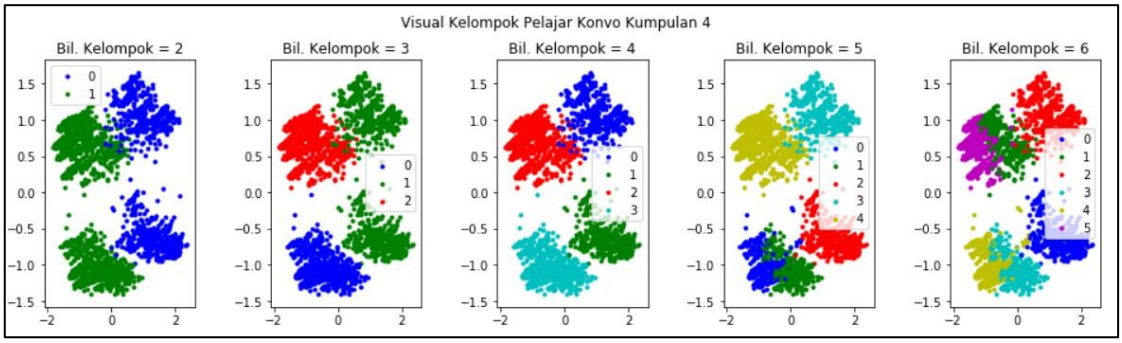
Proses seterusnya adalah menganalisis data yang diterima bagi memahami set data ini dengan lebih mendalam. Hasil penelitian, beberapa langkah perlu dilaksanakan dalam fasa ini seperti memperbaiki data yang hilang, mengeluarkan atribut yang tidak signifikan untuk digunakan dalam perlombongan data seperti nombor siri pelajar, atribut yang berkardinaliti 1 (*Distinct*: 1), atribut yang mempunyai lebih 80% data hilang, atribut yang berkolerasi dengan atribut yang lain, dan menggunakan kaedah *equal-depth (frequency) binning* terhadap atribut tertentu. Atribut yang berkolerasi bermaksud apabila terdapat dua atau lebih atribut yang sangat berhubung kait antara keduanya seperti ID_PENYELIA dan NAMA_PENYELIA, di mana salah satu daripada atribut tersebut boleh dikeluarkan tanpa menjejaskan data asal.

Kaedah *equal-depth (frequency) binning* pula merupakan kaedah untuk membahagikan data kepada beberapa selang supaya setiap selang mempunyai bilangan sampel yang hampir sama (Sarkar et al. 2018). Kaedah ini dibuat bagi mengelakkan model ramalan yang dihasilkan menjadi berat sebelah (*bias*). Begitu juga dengan beberapa atribut yang dikategorikan semula mengikut pemahaman domain, kerana mempunyai pecahan kategori yang terlalu banyak, atau julat nilai yang besar. Pakar domain juga turut berpandangan supaya set data pelajar ini dibahagikan kepada lapan kumpulan bagi pelajar cicir, dan lapan kumpulan bagi pelajar bergraduat Selain itu, mereka juga telah memuktamadkan atribut yang dipilih.

FASA PENENTUAN KELAS

Oleh kerana set data ini tidak mempunyai sebarang label kelas atau kumpulan, maka tugas pengelompokan perlu dilaksanakan terlebih dahulu dianalisis dengan lebih lanjut. Pengelompokan dengan Algoritma K-Min dipilih untuk mengenal pasti kelompok pelajar (Sarkar et al. 2018). Algoritma K-Min digunakan dengan menetapkan nilai K atau bilangan kelompok yang akan dibentuk kepada 2, 3, 4, 5 dan 6 kelompok. Kemudian, kaedah Analisis Komponen Utama atau *Principal Component Analysis* (PCA) (Alban & Mauricio 2019) dilaksanakan bagi mengurangkan dimensi data kepada dua komponen utama supaya kelompok yang terhasil tadi dapat divisualkan. PCA merupakan antara contoh algoritma yang popular untuk pengurangan dimensi atau *dimensionality reduction*. Rajah 2 menunjukkan contoh visual yang terhasil. Bagi menilai kualiti kelompok dan mengenal pasti berapa bilangan kelompok yang paling optimum perlu dipilih, tiga metrik tidak diselia digunakan iaitu Pekali Siluet,

Indeks Calinski-Harabasz dan Indeks Davies-Bouldin (Sarkar et al. 2018).



RAJAH 2. Contoh Visual Pengelompokan Menggunakan Algoritma K-Min Dengan Bilangan Kelompok 2 Hingga 6

FASA PENGELASAN

Pembelajaran terselia atau pengelasan menggunakan algoritma LR atau *Linear Regression*, SVM dan DT untuk meramal tahap prestasi akademik pelajar (Sarkar et al. 2018). LR merupakan algoritma pembelajaran mesin yang paling ringkas untuk pengelasan binari. Algoritma ini mudah dilaksanakan dan boleh digunakan sebagai asas untuk masalah pengelasan binari. Selain pengelasan binari, algoritma ini juga boleh membuat pengelasan multi-kelas atau dikenali sebagai regresi multinominal. Oleh kerana hasil pengelompokan yang diperolehi mempunyai dua atau lebih kelompok, kedua-dua jenis regresi digunakan iaitu binari untuk set data yang mempunyai pemboleh ubah sasaran (label kelas) sebanyak dua, manakala multinomial digunakan untuk set data yang mempunyai pemboleh ubah sasaran (label kelas) sebanyak tiga atau lebih.

FASA ANALISIS DAPATAN

Fasa ini adalah fasa terakhir dalam metodologi kajian iaitu proses menafsir atau menilai dapatan daripada pengelasan data. Proses ini akan menggambarkan corak dapatan daripada data atau model. Hasil kajian boleh diintegrasikan dengan sistem lain, didokumenkan atau dilaporkan kepada pihak yang berkaitan dengan domain kajian. Menurut Alyahyan & Dustegor (2020), lazimnya beberapa model pengelasan akan terhasil. Oleh itu, sangat penting untuk menilai dan memilih model yang paling sesuai dengan masalah. Matriks kekalutan (Jadual 2) digunakan untuk menilai prestasi algoritma pengelas (Sarkar et al. 2018).

JADUAL 2. Matriks Kekalutan

Kelas Sebenar	Kelas Yang Diramal	
	Cicir	Bergraduat
Cicir	Positif Betul (PB) *Bilangan pelajar yang cicir, dikelaskan betul sebagai "cicir"*	Negatif Salah (NS) *Bilangan pelajar yang bergraduat, dikelaskan betul sebagai "bergraduat"*
Bergraduat	Positif Salah (PS) *Bilangan pelajar yang bergraduat, dikelaskan salah sebagai "cicir"*	Negatif Betul (NB) *Bilangan pelajar yang bergraduat, dikelaskan betul sebagai "bergraduat"*

DAPATAN SETIAP FASA DAN PERBINCANGAN

Bahagian ini menghuraikan hasil keputusan dari setiap fasa di dalam metodologi kajian yang telah dijalankan ke atas set data pelajar pascasiswazah yang berstatus cicir atau bergraduat pada sesi 2014/2015 hingga 2019/2020. Pelajar cicir ini diklasifikasi sebagai pelajar yang telah diberi status Berhenti sama ada Diberhentikan, Gagal Peperiksaan/Pengajian atau Menarik Diri. Manakala, pelajar bergraduat adalah pelajar yang berjaya menamatkan pengajian dan dikurniakan ijazah.

INPUT PAKAR DOMAIN TENTANG DATA PELAJAR

Pakar domain telah bersetuju supaya set data pelajar dibahagikan kepada lapan kumpulan bagi pelajar cicir, dan lapan kumpulan bagi pelajar bergraduat. Keterangan kumpulan dan bilangan sampel bagi setiap kumpulan pelajar cicir dan bergraduat ini ditunjukkan dalam Jadual 3. Selain itu, atribut akhir yang dimuktamadkan untuk dipilih adalah seperti yang ditunjukkan dalam Jadual 4 bagi pelajar cicir, dan Jadual 5 bagi pelajar bergraduat.

JADUAL 3. Kumpulan Pelajar Cicir Dan Bergraduat

Kumpulan	Keterangan	Bilangan Cicir	Sampel Bergraduat
Pelajar yang mengikuti Program Diploma Pascasiswazah Secara Kerja Kursus		7	177
Pelajar yang mengikuti Program Diploma Siswazah Secara Kerja Kursus		18	172
Pelajar yang mengikuti Program Doktor Falsafah Secara Kerja Kursus dan Penyelidikan atau Mod Campuran		73	87
Pelajar yang mengikuti Program Doktor Falsafah Secara Penyelidikan		**1,296**	**2,545**
Pelajar yang mengikuti Program Sarjana Secara Kerja Kursus		1,351	6,120
Pelajar yang mengikuti Program Sarjana Secara Kerja Kursus dan Penyelidikan atau Mod Campuran		1,055	1,398
Pelajar yang mengikuti Program Sarjana Secara Kerja Kursus dan Praktikum atau Kerja Klinikal serta Kajian Kes dan/atau Penyelidikan		219	1175
Pelajar yang mengikuti Program Sarjana Secara Penyelidikan		937	1321

JADUAL 4. Atribut Pelajar Cicir

Bil.	Nama Atribut	Kumpulan 1	2	3	4	5	6	7	8
1.	SEM_SESI_HENTI	✓	✓	✓	✓	✓	✓	✓	✓
2.	STATUS				✓		✓	✓	✓
3.	SEBAB_HENTI	✓	✓	✓	✓	✓	✓	✓	✓
4.	SEM_SESI_MASUK	✓	✓	✓	✓	✓	✓	✓	✓
5.	FAKULTI				✓				✓
6.	PROGRAM_PENGAJIAN		✓	✓	✓	✓	✓		✓
7.	JENIS_PENDAFTARAN		✓	✓	✓	✓	✓	✓	✓
8.	BIL_SEM_TANGGUH		✓	✓	✓	✓	✓	✓	✓
9.	STATUS_TAJAAN			✓	✓	✓	✓	✓	✓
10.	BIL_PENYELIA			✓	✓		✓	✓	✓
11.	ID_PENYELIA			✓	✓		✓	✓	✓

12.	BIL_PELAJAR_SELIA			✓	✓		✓	✓	✓
13.	BIL_PELAJAR_GAGAL			✓	✓		✓	✓	✓
14.	PNGK_MASUK		✓	✓	✓	✓	✓		✓
15.	PNGK_HENTI	✓	✓	✓		✓	✓		
16.	JANTINA	✓	✓	✓	✓	✓	✓	✓	✓
17.	STATUS_KAHWIN	✓	✓	✓	✓	✓	✓	✓	✓
18.	NEGARA_ASAL			✓	✓	✓	✓	✓	✓
Jumlah Atribut		6	10	16	17	12	17	14	17

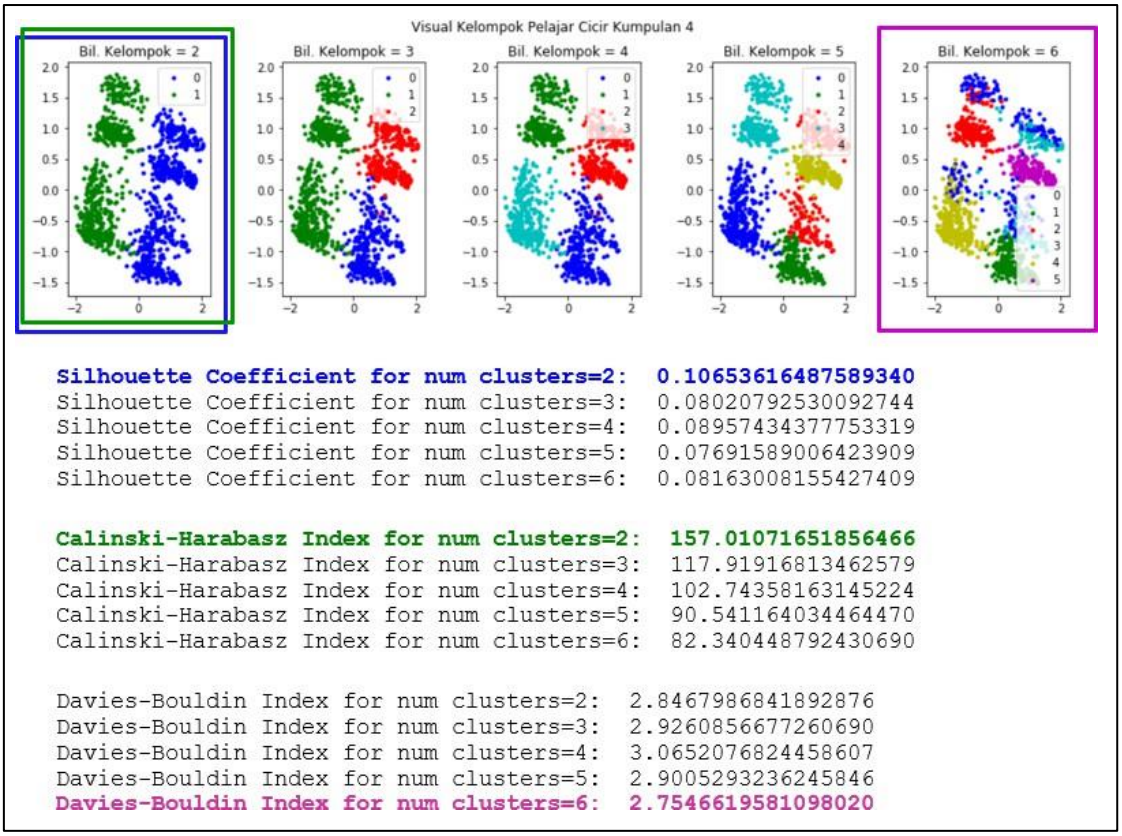
JADUAL 5. Atribut Pelajar Bergraduat

Bil.	Nama Atribut	Kumpulan							
		1	2	3	4	5	6	7	8
1.	SEM_SESI_HENTI	✓	✓	✓	✓	✓	✓	✓	✓
2.	SEM_SESI_MASUK	✓	✓	✓	✓	✓	✓	✓	✓
3.	FAKULTI				✓				✓
4.	PROGRAM_PENGAJIAN		✓	✓	✓	✓	✓		✓
5.	JENIS_PENDAFTARAN		✓	✓	✓	✓	✓	✓	✓
6.	BIL_PENERBITAN			✓	✓	✓	✓	✓	✓
7.	STATUS_TAJAAN	✓		✓	✓	✓	✓	✓	✓
8.	BIL_PENYELIA			✓	✓		✓	✓	✓
9.	ID_PENYELIA			✓	✓		✓	✓	✓
10.	BIL_PELAJAR_SELIA			✓	✓		✓	✓	✓
11.	BIL_PELAJAR_GAGAL			✓	✓		✓	✓	✓
12.	PNGK_MASUK	✓	✓		✓	✓	✓		✓
13.	PNGK_HENTI			✓		✓	✓		
14.	JANTINA	✓	✓	✓	✓	✓	✓	✓	✓
15.	STATUS_KAHWIN	✓	✓	✓	✓	✓	✓	✓	✓
16.	STATUS_BEKERJA	✓	✓	✓	✓	✓	✓	✓	✓
17.	NEGARA_ASAL		✓	✓	✓	✓	✓	✓	✓
Jumlah Atribut		7	9	15	16	12	16	13	16

Terdapat sejumlah 16 kumpulan dari Jadual 4 dan 5. Namun, untuk skop kajian ini perbincangan akan fokus kepada Pelajar Cicir Kumpulan 4 (pelajar yang mengikuti Program Doktor Falsafah Secara Penyelidikan) (Jadual 3).

HASIL PENGELOMPOKAN

Tugas pengelompokan perlu dilaksanakan terlebih dahulu untuk menentukan kelompok atau kumpulan yang wujud di dalam set data. Pengelompokan dengan Algoritma K-Min dipilih untuk mengenal pasti kelompok pelajar. Rajah 3 menunjukkan visual kelompok serta nilai skor Pekali Siluet, Indeks Calinski-Harabasz dan Indeks Davies-Bouldin (Sarkar et al. 2018) yang diperolehi untuk Pelajar Cicir Kumpulan 4 (pelajar yang mengikuti Program Doktor Falsafah Secara Penyelidikan). Nilai K atau bilangan kelompok ditetapkan kepada 2, 3, 4, 5 dan 6 kelompok. Tiga metrik tidak diselia iaitu Pekali Siluet, Indeks Calinski-Harabasz dan Indeks Davies- Bouldin digunakan untuk menilai kualiti kelompok dan mengenal pasti berapa bilangan kelompok yang paling sesuai untuk dipilih.



RAJAH 3. Visual Kelompok Serta Nilai Pekali Siluet, Indeks Calinski-Harabasz Dan Indeks Davies-Bouldin Yang Diperolehi Untuk Pelajar Cicir Kumpulan 4

Berdasarkan Rajah 3, skor terbaik untuk Pekali Siluet adalah 0.1065 iaitu skor yang paling menghampiri nilai 1, skor terbaik untuk Indeks Calinski-Harabasz pula adalah 157.0107 iaitu nilai yang paling besar, manakala skor Indeks Davies-Bouldin yang terbaik adalah 2.7547, iaitu nilai yang paling kecil. Berdasarkan skor terbaik ini, bilangan kelompok paling optimum untuk Pekali Siluet dan Indeks Calinski-Harabasz adalah 2 kelompok, manakala bilangan kelompok paling optimum untuk Indeks Davies-Bouldin adalah 6 kelompok.

Jadual 6 dan Jadual 7 menunjukkan ringkasan bilangan kelompok paling optimum dan nilai skor ketiga-tiga metrik penilaian bagi kumpulan pelajar cicir dan bergraduat. Kajian mendapati bahawa ketiga-tiga metrik penilaian memperolehi bilangan kelompok paling optimum iaitu sama ada 2, 3, 5 atau 6 kelompok. Pekali Siluet menunjukkan skor yang seragam kerana nilai skor bagi pekali ini adalah antara -1 hingga 1, di mana nilai yang menghampiri 1 merupakan kelompok yang terbaik. Oleh itu, bilangan kelompok paling optimum yang diperolehi menggunakan Pekali Siluet akan digunakan bagi menetapkan label kelas pelajar cicir dan bergraduat untuk tugas perlombongan data seterusnya iaitu pengelasan.

JADUAL 6. Ringkasan Bilangan Kelompok Paling Optimum Dan Nilai Skor 3 Metrik Penilaian Bagi Kumpulan Pelajar Cicir

Kump.	Pekali Siluet		Indeks Calinski_Harabasz		Indeks Davies-Bouldin	
	Bilangan Kelompok Paling Optimum	Skor	Bilangan Kelompok Paling Optimum	Skor	Bilangan Kelompok Paling Optimum	Skor
1	3	0.1488079632	6	3.742857143	6	0.308869843
2	2	0.1611372699	2	4.246655032	6	1.257116718
3	2	0.1724297930	2	16.44443223	5	1.931175051
4	2	0.1065361649	2	157.0107165	6	2.754661958
5	2	0.1311704728	2	195.4420123	6	2.452682567
6	3	0.1122162003	2	113.6364551	5	2.512495818
7	2	0.2276780730	2	63.36670995	2	1.779564218
8	2	0.1087500892	2	105.8864894	2	2.763435343

JADUAL 7. Ringkasan Bilangan Kelompok Paling Optimum Dan Nilai Skor 3 Metrik Penilaian Bagi Kumpulan Pelajar Bergraduat

Kump.	Pekali Siluet		Indeks Calinski_Harabasz		Indeks Davies-Bouldin	
	Bilangan Kelompok Paling Optimum	Skor	Bilangan Kelompok Paling Optimum	Skor	Bilangan Kelompok Paling Optimum	Skor
1	6	0.3653577200	6	46.52298732	5	1.277284770
2	6	0.4600524426	5	61.73909882	6	1.073822579
3	3	0.1556216069	3	13.29662772	6	1.894737006
4	2	0.1105417179	2	309.2047370	2	2.768096700
5	2	0.1521525394	2	999.2615156	6	2.297696767
6	2	0.1030207273	2	150.3079799	5	2.555246096
7	2	0.1930317928	2	262.9626526	2	1.879651668
8	2	0.1080690364	2	137.8375015	3	2.656053787

Seterusnya, berdasarkan hasil dari Jadual 6 dan 7, label kelas untuk set data yang mempunyai 2, 3 dan 6 kelompok paling optimum berdasarkan nilai skor Pekali Siluet dihasilkan seperti Jadual 8.

JADUAL 8. Label Kelas Untuk Bilangan Kelompok Paling Optimum 2, 3 Dan 6

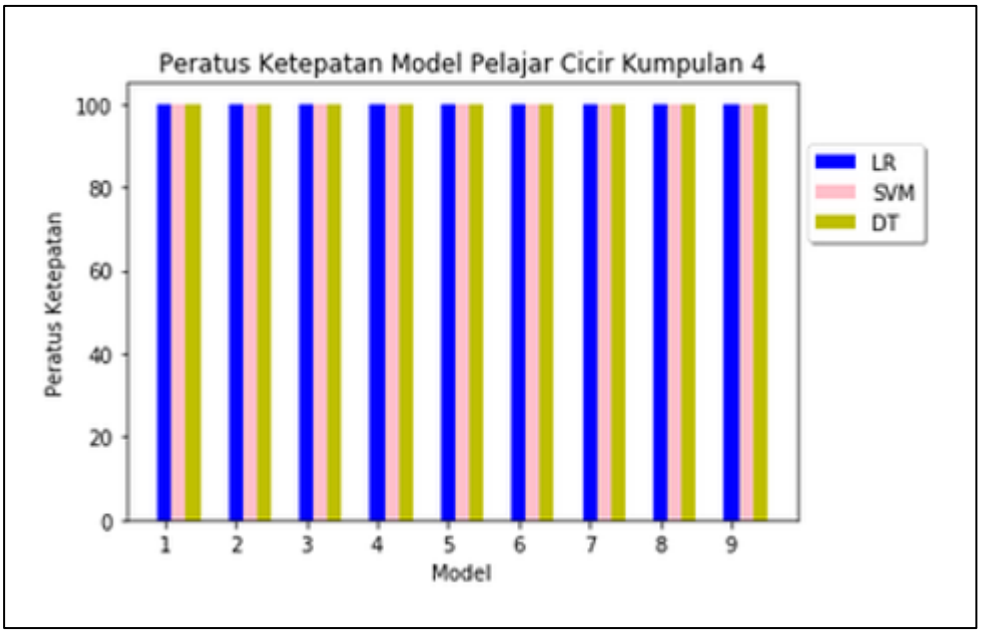
Bilangan Kelompok	Label Kelas
2	Rendah Tinggi
3	Rendah Sederhana Tinggi
6	Sangat Rendah Rendah Sederhana Rendah Sederhana Tinggi Tinggi Sangat Tinggi

Jadual 8 menunjukkan bilangan kelompok dan label yang digunakan untuk bilangan kelompok.

Sebagai contoh, untuk bilangan 2 kelompok, satu kelompok dilabel sebagai Rendah dan satu lagi sebagai Tinggi. Untuk 3 kelompok pula, satu kelompok dilabel sebagai Rendah, diikuti kelompok seterusnya sebagai Sederhana dan kelompok akhir dilabel Tinggi. Manakala bagi 6 kelompok, satu kelompok dilabel sebagai SangatRendah, kelompok seterusnya dilabel Rendah dan diikuti kelompok seterusnya sebagai SederhanaRendah. Kelompok seterusnya dilabel sebagai SederhanaTinggi dan dua kelompok terakhir dialbel sebagai Tinggi dan Sangat Tinggi.

PENGELASAN

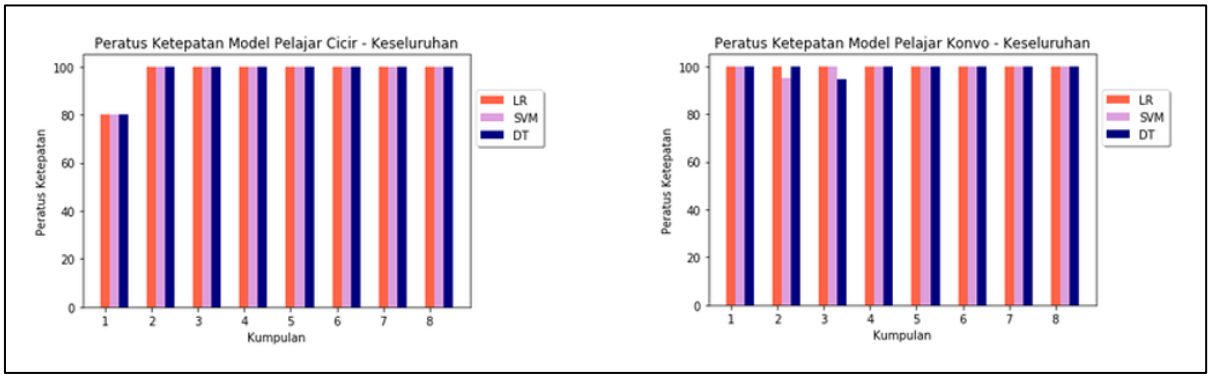
Pengelasan menggunakan algoritma pengelas untuk meramal tahap prestasi akademik pelajar. Tiga algoritma pengelas dipilih iaitu LR, SVM dan DT kerana algoritma pengelas ini banyak digunakan dalam kajian-kajian lepas yang berkaitan Perlombongan Data Pendidikan. Pengelasan dilaksana menggunakan kaedah *Percentage Split* iaitu set data bagi setiap kumpulan dipecahkan kepada 9 Model untuk setiap algoritma pengelas. Model 1 – 90:10 iaitu 90% data untuk Data Latihan, dan 10% data lagi untuk Data Ujian, Model 2 – 80:20, Model 3 – 70:30, Model 4 – 60:40, Model 5 – 50:50, Model 6 – 40:60, Model 7 – 30:70, Model 8 – 20:80, dan Model 9 – 10:90. Kemudian, model terbaik untuk setiap kumpulan dipilih mengikut ketiga-tiga algoritma pengelas tadi. Rajah 4 menunjukkan carta palang peratus ketepatan Model Pelajar Cicir Kumpulan 4 bagi tiga algoritma pengelas.



RAJAH 4. Carta Palang Peratus Ketepatan Model Pelajar Cicir Kumpulan 4 Bagi Tiga Algoritma Pengelas

Oleh kerana semua peratus ketepatan yang diperolehi adalah sama (100%), maka Model 3 dipilih sebagai model terbaik untuk Pelajar Cicir Kumpulan 4 kerana pembahagian 70% data untuk Data Latihan, dan 30% data lagi untuk Data Ujian adalah model yang biasa dipilih untuk pengelasan (Jimenez et al. 2018; Lee et al. 2018; Sahar et al. 2020; Verma & Illes 2019). Kemudian, kesemua model terbaik untuk setiap kumpulan dipilih mengikut ketiga-tiga algoritma pengelas tadi. Rajah 5 menunjukkan carta palang peratus ketepatan model terbaik setiap kumpulan pelajar cicir dan bergraduat bagi tiga algoritma pengelas. Jadual 9 mewakilkan beberapa metrik penilaian seperti peratus ketepatan, *precision*, *recall*, F1 dan ROC untuk SVM,

LR dan DT. Jadual 10 menunjukkan senarai atribut penting untuk model pelajar cicir kumpulan 4 bagi SVM, LR dan DT. Manakala, Jadual 11 menunjukkan set petua janaan DT. Kesemua Jadual 12 hingga Jadual 14 adalah bagi Model Pelajar Cicir Kumpulan 4.



RAJAH 5. Carta Palang Peratus Ketepatan Model Terbaik Setiap Kumpulan Pelajar Cicir Dan Bergraduat Bagi Tiga Algoritma Pengelas

JADUAL 9. Metrik Penilaian Model Pelajar Cicir Kumpulan 4 Bagi Tiga Algoritma Pengelas

		LR					SVM					DT				
	Model	*Accuracy* (%)	*Precision*	*Recall*	F1	ROC	*Accuracy* (%)	*Precision*	*Recall*	F1	ROC	*Accuracy* (%)	*Precision*	*Recall*	F1	ROC
1	90:10	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
2	80:20	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
3	**70:30**	**100.00**	**1.0000**	**1.0000**	**1.0000**	**1.00**	**100.00**	**1.0000**	**1.0000**	**1.0000**	**1.00**	**100.00**	**1.0000**	**1.0000**	**1.0000**	**1.00**
4	60:40	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
5	50:50	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
6	40:60	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
7	30:70	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
8	20:80	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00
9	10:90	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00	100.00	1.0000	1.0000	1.0000	1.00

JADUAL 10. Senarai Atribut Penting Bagi Model Pelajar Cicir Kumpulan 4 Bagi Tiga Algoritma Pengelas

		Atribut penting (LR)	Atribut penting (SVM)	Atribut penting (DT)
1	90:10	PROGRAM_PENGAJIAN, FAKULTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
2	80:20	PROGRAM_PENGAJIAN, FAKULTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
3	**70:30**	**PROGRAM_PENGAJIAN, FAKULTI**	**PROGRAM_PENGAJIAN**	**PROGRAM_PENGAJIAN**
4	60:40	PROGRAM_PENGAJIAN, FAKULTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
5	50:50	PROGRAM_PENGAJIAN, FAKULTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
6	40:60	PROGRAM_PENGAJIAN, FAKULTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
7	30:70	PROGRAM_PENGAJIAN, FAKULTI, JENIS_PENDAFTARAN, NEGARA_ASAL	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
8	20:80	PROGRAM_PENGAJIAN, FAKULTI, JENIS_PENDAFTARAN, NEGARA_ASAL, BIL_PELAJAR_GAGAL, SEM_SESI_HENTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN
9	10:90	PROGRAM_PENGAJIAN, FAKULTI, JENIS_PENDAFTARAN, BIL_PELAJAR_GAGAL, BIL_PELAJAR_SELIA, NEGARA_ASAL, JANTINA, PNGK_MASUK, STATUS_TAJAAN, SEM_SESI_HENTI	PROGRAM_PENGAJIAN	PROGRAM_PENGAJIAN

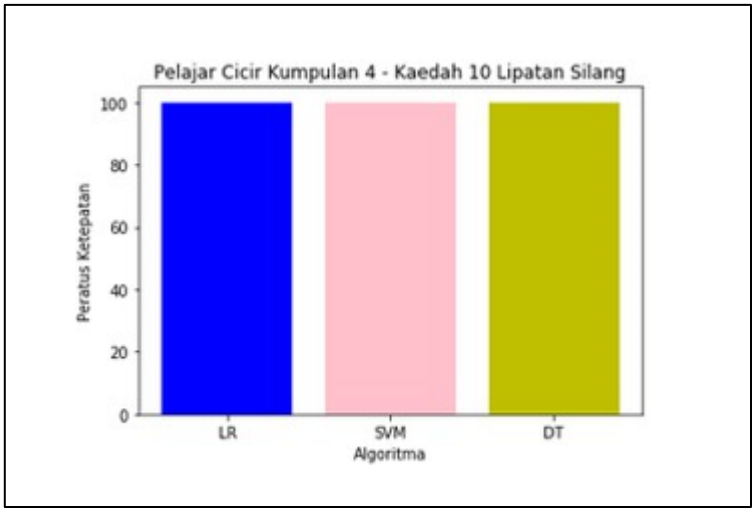
JADUAL 11. Set Petua Janaan DT Bagi Model Pelajar Cicir Kumpulan 4

	Petua Janaan DT		
	Model	Set Petua Asal	Set Petua Diterjemah
1	90:10	▪ Jika PROGRAM_PENGAJIAN <= -0.121 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.121 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
2	80:20	▪ Jika PROGRAM_PENGAJIAN <= -0.117 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.117 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
3	**70:30**	▪ **Jika PROGRAM_PENGAJIAN <= -0.119 Maka KLUSTER = 1** • **Jika PROGRAM_PENGAJIAN > -0.119 Maka KLUSTER = 0**	▪ **Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah** ▪ **Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi**
4	60:40	▪ Jika PROGRAM_PENGAJIAN <= -0.131 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.131 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
5	50:50	▪ Jika PROGRAM_PENGAJIAN <= -0.15 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.15 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
6	40:60	▪ Jika PROGRAM_PENGAJIAN <= -0.16 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.16 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
7	30:70	▪ Jika PROGRAM_PENGAJIAN <= -0.125 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.125 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah ▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi
8	20:80	▪ Jika PROGRAM_PENGAJIAN <= -0.074 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.074 Maka KLUSTER = 0	
9	10:90	▪ Jika PROGRAM_PENGAJIAN <= -0.149 Maka KLUSTER = 1 • Jika PROGRAM_PENGAJIAN > -0.149 Maka KLUSTER = 0	

Berdasarkan Rajah 5, model LR sesuai digunakan untuk meramal tahap prestasi akademik pelajar kerana semua output yang dihasilkan mencapai peratus ketepatan maksimum iaitu 100% (hanya Pelajar Cicir Kumpulan 1 sahaja mencapai peratus ketepatan sebanyak 80% bagi semua model pengelas). Metrik penilaian lain seperti ketepatan, *precision*, *recall*, F1 dan ROC juga menunjukkan nilai maksimum diperolehi. Antara atribut penting yang kerap ditemui adalah JANTINA, PROGRAM_PENGAJIAN, STATUS_KAHWIN, SEM_SESI_HENTI, ID_PENYELIA, BIL_PELAJAR_SELIA, STATUS_BEKERJA dan SEM_SESI_MASUK.

Oleh kerana kebanyakan peratus ketepatan yang diperolehi mencapai 100%, model mungkin mempunyai masalah *overfitting* di mana model begitu spesifik atau terikat dengan data sehingga gagal melakukan generalisasi apabila menerima set data yang belum pernah dilihat. Bagi mengelakkan masalah ini, Kaedah 10 Lipatan Silang pula dilaksanakan. Menurut Sani et al. (2018), lipatan silang adalah kaedah statistik untuk menilai model ramalan dengan membahagi sampel asal menjadi dua bahagian. Satu set latihan untuk belajar atau melatih model, dan satu set ujian untuk melakukan penilaian ke atasnya. Kaedah yang dipanggil k-Lipatan Silang dilaksanakan apabila data diasingkan menjadi k lipatan bersaiz sama. Latihan dan pengesahan dilakukan berulang kali sebanyak k-lelaran. Dalam setiap lelaran, lipatan data yang berlainan disimpan untuk pengesahan, manakala baki k-1 lipatan disimpan untuk belajar. Ini bagi memastikan dalam setiap lelaran, model dilatih pada subset data yang berbeza. Oleh itu, teknik lipatan silang sangat berkesan dipraktikkan untuk pemilihan model (Sarkar et al. 2018).

Rajah 6 menunjukkan carta palang peratus ketepatan Model Pelajar Cicir Kumpulan 4 bagi tiga algoritma pengelas, menggunakan Kaedah 10 Lipatan Silang. Jadual 12 pula menunjukkan beberapa metrik penilaian seperti peratus ketepatan dan sisihan piawai. Jadual 13 menunjukkan atribut penting janaan DT, dan Jadual 14 menunjukkan set petua janaan DT. Kesemua Jadual 12 hingga Jadual 14 adalah bagi Model Pelajar Cicir Kumpulan 4, menggunakan Kaedah 10 Lipatan Silang.



RAJAH 6. Carta Palang Peratus Ketepatan Model Pelajar Cicir Kumpulan 4 Bagi Tiga Algoritma Pengelas, Menggunakan Kaedah 10 Lipatan Silang

JADUAL 12. Metrik Penilaian Model Pelajar Cicir Kumpulan 4 Bagi Tiga Algoritma Pengelas, Menggunakan Kaedah 10 Lipatan Silang

LR		SVM		DT	
Ketepatan (%)	Sisihan Piawai	Ketepatan (%)	Sisihan Piawai	Ketepatan (%)	Sisihan Piawai
100.00	+/- 0.0000	100.00	+/- 0.0000	100.00	+/- 0.0000

JADUAL 13. Atribut Penting Janaan DT Bagi Model Pelajar Cicir Kumpulan 4, Menggunakan Kaedah 10 Lipatan Silang

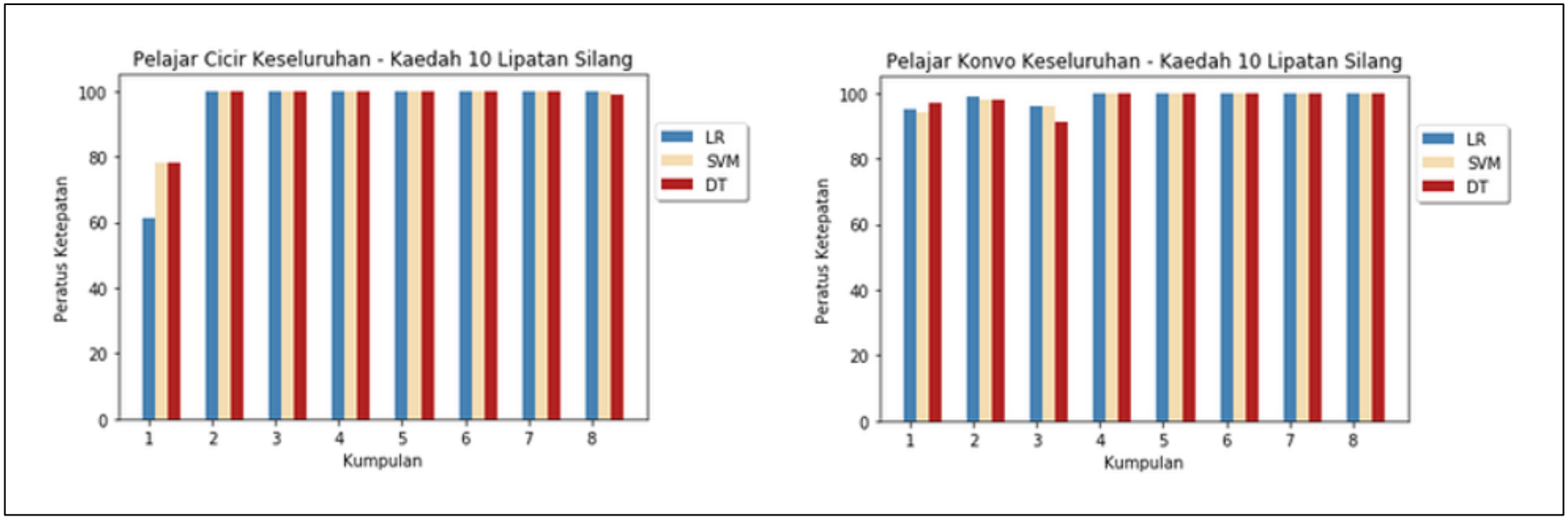
LR	SVM	DT
Atribut Penting	Atribut Penting	Atribut Penting
-	-	PROGRAM_PENGAJIAN

JADUAL 14. Set Petua Janaan DT Bagi Model Pelajar Cicir Kumpulan 4, Menggunakan Kaedah 10 Lipatan Silang

Petua Janaan DT	
Set Petua Asal	Set Petua Diterjemah
Jika PROGRAM_PENGAJIAN <= -0.102 Maka KLUSTER = 1	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahS&T Maka CICIR = Rendah
Jika PROGRAM_PENGAJIAN > -0.102 Maka KLUSTER = 0	▪ Jika PROGRAM_PENGAJIAN = DoktorFalsafahBukanS&T Maka CICIR = Tinggi

Seterusnya, Rajah 7 menunjukkan carta palang peratus ketepatan model setiap kumpulan pelajar cicir dan bergraduat bagi tiga algoritma pengelas, menggunakan Kaedah 10 Lipatan Silang. Berdasarkan Rajah 9 dan Rajah 11, analisis mendapati bahawa peratus ketepatan untuk Kaedah *Percentage Split* dan Kaedah 10 Lipatan Silang tidak jauh berbeza. Terdapat sedikit penurunan keputusan sahaja untuk Pelajar Cicir Kumpulan 1 dan Pelajar Bergraduat Kumpulan 1 – 3. Ini menunjukkan kebanyakan keputusan yang memperolehi peratus ketepatan tinggi, bukanlah disebabkan masalah *overfitting* kerana keputusan bagi Kaedah 10 Lipatan Silang menunjukkan bahawa model boleh digeneralisasi dengan baik apabila menerima set data yang belum pernah dilihat. Pelaksanaan Kaedah 10 Lipatan Silang juga menunjukkan antara atribut penting yang kerap ditemui adalah JANTINA, PROGRAM_PENGAJIAN, STATUS_KAHWIN, BIL_PELAJAR_SELIA, SEM_SESI_ HENTI, STATUS_BEKERJA, STATUS_TAJAAN dan SEM_SESI_MASUK. Kekerapan atribut penting dinilai menerusi kewujudan atribut itu di dalam petua yang dijana menerusi algoritma DT.

Jadual 15 pula menunjukkan jumlah kedudukan penilaian prestasi untuk setiap model pengelas berdasarkan Kaedah 10 Lipatan Silang. Peratus ketepatan yang paling tinggi diletakkan pada kedudukan pertama, kedua tertinggi pada kedudukan ke-2, dan paling rendah pada kedudukan ke-3. Kemudian, kedudukan ini dijumlahkan bagi melihat secara keseluruhan, model pengelas mana yang sesuai digunakan untuk meramal tahap prestasi akademik pelajar. Kedudukan menunjukkan model LR berada di kedudukan pertama, berbanding SVM dan DT yang kedua-duanya berada di kedudukan kedua. Seperti dibincangkan sebelum ini, LR merupakan algoritma pengelas yang paling ringkas dan dijadikan asas untuk menyelesaikan masalah pengelasan. Keputusan ini menunjukkan bahawa model LR memadai untuk dipilih kerana set data ini tidak terlalu kompleks, dan mengandungi bilangan sampel yang sedikit (Jadual 3).



18

RAJAH 7. Carta Palang Peratus Ketepatan Model Terbaik Setiap Kumpulan Pelajar Cicir Dan Bergraduat Bagi Tiga Algoritma Pengelas, Menggunakan Kaedah 10 Lipatan Silang

JADUAL 15. Atribut Penting Janaan DT Bagi Model Pelajar Cicir Kumpulan

Model	Kumpulan Pelajar Cicir								Kumpulan Pelajar Bergraduat								Jumlah Kedudukan
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	
LR	2	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	18
SVM	1	1	1	1	1	1	1	1	3	2	1	1	1	1	1	1	19
DT	1	1	1	1	1	1	1	2	1	2	2	1	1	1	1	1	19

KESIMPULAN

Set data pelajar pascasiswazah yang berstatus berhenti/bergraduat pada sesi 2014/2015 hingga 2019/2020 telah diperolehi dan diproses sebelum digunakan dalam algoritma perlombongan data. Pengelompokan dengan Algoritma K-Min pula dipilih untuk mengenal pasti kelompok pelajar. Kualiti kelompok telah diuji menggunakan tiga metrik penilaian iaitu Pekali Siluet, Indeks Calinski-Harabasz dan Indeks Davies-Bouldin bagi mengenal pasti berapa bilangan kelompok paling optimum perlu dipilih. Bilangan kelompok paling optimum dipilih berdasarkan skor Pekali Siluet kerana nilai skor bagi pekali ini adalah seragam iaitu antara -1 hingga 1. Pada mulanya bilangan kelompok ditetapkan kepada 2, 3, 4, 5 dan 6 kelompok. Hasil yang diperolehi menunjukkan bilangan kelompok paling optimum yang diperolehi menggunakan Pekali Siluet adalah 2 kelompok untuk 11 kumpulan, 3 kelompok untuk tiga kumpulan, dan 6 kelompok untuk dua kumpulan.

Pengelasan pula merupakan tugas terakhir yang digunakan untuk meramal tahap prestasi akademik pelajar. Tiga algoritma pengelas telah dipilih iaitu LR, SVM dan DT. Pengelasan dibuat menggunakan Kaedah Percentage Split dan Kaedah 10 Lipatan Silang. Beberapa metrik penilaian digunakan untuk menilai model bagi setiap pengelas. Keputusan menunjukkan bahawa model LR sesuai digunakan untuk meramal tahap prestasi akademik pelajar, berbanding SVM dan DT kerana mendapat kedudukan pertama dalam kedudukan penilaian prestasi untuk setiap model pengelas.

Dari konteks faktor yang mempengaruh prestasi akademi pelajar, kajian menemui lapan atribut yang kerap muncul dalam analisis iaitu JANTINA, PROGRAM_PENGAJIAN, STATUS_KAHWIN, BIL_PELAJAR_SELIA, SEM_SESI_ HENTI, STATUS_BEKERJA, STATUS_TAJAAN dan SEM_SESI_MASUK. Tiga atribut dari senarai ini iaitu JANTINA, PROGRAM_PENGAJIAN dan STATUS_BEKERJA telah dinyatakan sebagai faktor paling mempengaruhi prestasi akademik pelajar dalam kajian Alyahyan & Dustegor (2020). Kajian lebih mendalam dicadangkan untuk menilai lima atribut yang berbaki sama ada sesuai untuk menilai prestasi akademik pelajar.

Oleh itu, bagi menambah baik kajian pada masa hadapan, adalah dicadangkan supaya skop kajian ini diperluaskan kepada semua pelajar merangkumi semua tahap pengajian termasuk Asasi dan Prasiswazah. Faktor-faktor penting yang mempengaruhi ramalan prestasi akademik pelajar seperti yang dibincangkan dalam topik sebelum ini boleh diambil kira bagi mendapatkan hasil kajian yang lebih komprehensif. Selain itu, penemuan pengetahuan yang lebih menarik mungkin diperolehi apabila data mengandungi maklumat yang mencukupi, contohnya data sejarah yang diambil dalam tempoh 10 tahun atau lebih kerana data ini mewakili banyak kumpulan pelajar berbanding sekiranya data sejarah diambil dalam tempoh yang pendek, di mana pelajar tersebut berkemungkinan hanya terdiri daripada kumpulan pelajar yang sama sahaja.

PENGHARGAAN

Kajian ini merakamkan ucapan terima kasih kepada Universiti Kebangsaan Malaysia dan Kementerian Pengajian Tinggi atas sokongan di dalam kajian ini menerusi geran penyelidikan Fundamental Research Grant Scheme bertajuk An Ensemble Multimodal Decision Analytics Approach For Collaborative Pandemic dengan kod FRGS/1/2022/ICT02/UKM/02/7.

RUJUKAN

Ahmad, F., Ismail, N.H. & Aziz, A.A. 2015. The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. *Applied Mathematical Sciences* 9(129): 6415– 6426.

Ashraf Abdelhadi, Suhaila Zainudin and Nor Samsiah Sani. 2022. A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments. *International Journal of Advanced Computer Science and Applications* 13(1) DOI: http://dx.doi.org/10.14569/IJACSA.2022.0130156.

Alban, M. & Mauricio, D. 2019. Predicting University Dropout through Data Mining: A Systematic Literature. *Indian Journal of Science and Technology* 12(4): 1–12.

Ali, M. A., Sahari, N., Fadzilah, S., & Zainudin. (2022). Identifying students' learning patterns in online learning environments: A literature review. International Journal of Emerging Technologies in Learning (Online), 17(8), 189-205. DOI:https://doi.org/10.3991/ijet.v17i08.29811.

Almarabeh, H. 2017. Analysis of Students' Performance by Using Different Data Mining Classifiers. *International Journal of Modern Education and Computer Science* 9(8): 9–15.

Aluko, R.O., Daniel, E.I., Oshodi, O., Aigbavboa, C.O. & Abisuga, A.O. 2018. Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology* 16(3): 385–397.

Alyahyan, E. & Dustegor, D. 2020. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education* 17(3): 1–21.

Mohd Khairy, Azhar & Adam, Afzan & Yaakub, Mohd Ridzwan. 2018. Data Analytics In Malaysian Education System: Revealing The Success Of Sijil Pelajaran Malaysia From Ujian Aptitud Sekolah Rendah. *Asia-Pacific Journal of Information Technology and Multimedia*. 7(2). 29-45. DOI: 10.17576/apjitm-2018-0702- 03.

Bakhshinategh, B., Zaiane, O.R., ElAtia, S. & Ipperciel, D. 2017. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies* 23(1): 537–553.

Hamoud, A.K., Hashim, A.S. & Awadh, W.A. 2018. Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence* 5(2): 26–31.

Jimenez, F., Martinez, C., Miralles-Pechuan, L., Sanchez, G. & Sciavicco, G. 2018. Multi-Objective Evolutionary Rule-Based Classification with Categorical Data. *Entropy 2018* 20(9): 684–716.

Kementerian Pendidikan Malaysia. 2013. *Pelan Pembangunan Pendidikan Malaysia 2013-2025 (Pendidikan Prasekolah hingga Lepas Menengah)*.

Lee, L.C., Liong, C.Y. & Jemain, A.A. 2018. Validity of the best practice in Splitting Data for Hold-out Validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal* 139: 125–142.

Liang, J., Yang, J., Wu, Y., Li, C. & Zheng, L. 2016. Big Data Application in Education: Dropout Prediction in Edx MOOCs. *IEEE Second International Conference on Multimedia Big Data,* hlm 440–443.

Mesaric, J. & Sebalj, D. 2016. Decision trees for predicting the academic success of students. *Croatian Operational Research Review* 7(2): 367–388.

Mueen, A., Zafar, B. & Manzoor, U. 2016. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science* 8(11): 36–42.

Sahar, M.W., Beaver, A., Keyserlingk, M.A.G. & Weary, D.M. 2020. Predicting Disease in Transition Dairy Cattle Based on Behaviors Measured Before Calving. *Animals 2020* 10(6): 928–942.

Sani, N.S., Rahman, M.A., Bakar, A.A., Sahran, S. & Sarim, H.M. 2018. Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification. *International Journal on Advanced Science Engineering Information Technnology* 8(4–2): 1698–1705.

Sarkar, D., Bali, R. & Sharma, T. 2018. Practical Machine Learning with Python: A Problem Solver's Guide to Building Real-World Intelligent Systems. Berkely: Apress.

Shahiri, A.M., Husain, W. & Rashid. N.A. 2015. A Review on Predicting Student's Performance using Data Mining Techniques. *Procedia Computer Science* 72: 414–422.

Sivakumar, S., Venkataraman, S. & Selvaraj, R. 2016. Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. *Indian Journal of Science and Technology* 9(4): 1–5.

Sivasakthi, M. 2017. Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance. *Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)*, hlm 346–350.

Universiti Kebangsaan Malaysia. 2018. *Laporan Tahunan UKM 2018*.

Verma, C. & Illes, Z. 2019. Attitude Prediction towards ICT and Mobile Technology for the Real- Time: An Experimental Study using Machine Learning. *The 15th International Scientific Conference eLearning and Software for Education* 3: 247–254.

Wirawati Dewi, Ahmad & Azuraliza, Bakar. 2018. Classification Models for Higher Learning Scholarship Award Decisions. *Asia-Pacific Journal of Information Technology and Multimedia*. 07. 131-145. 10.17576/apjitm-2018-0702-10.

Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11 2022. https://doi.org/10.1186/s40561-022- 00192-z

Yasmin, D. 2013. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education* 34(2): 218–231.

Zainal Rafit, Z. H., Zainudin, S., & Ali Othman, Z. 2021. Model Peramalan Bilangan Calon Tarik Diri dari Peperiksaan Awam Malaysia Menerusi Pendekatan Perlombongan Data dan Petua: Non-Attendance Candidates' Prediction Model for Malaysia Public Exam Using Data Mining and Rules Approach. *Journal of ICT in Education*, 8(1), 26–42. https://doi.org/10.37134/jictie.vol8.1.3.2021.

Zhang Y, Yun Y, An R, Cui J, Dai H, Shang X. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. Front Psychol. 2021 Dec 7;12:698490. doi: 10.3389/fpsyg.2021.698490. Erratum in: Front Psychol. 2022 Jan 21;12:842357. PMID: 34950079; PMCID: PMC8688359.