Journal of ICT in Education (JICTIE)
ISSN 2289-7844 / Vol. 8 / Issue 1 / 2021 / 26-42

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

Zaiyinul Hayat Zainal Rafit[1], Suhaila Zainudin[2*], Zulaiha Ali Othman[3]

[1]*Lembaga Peperiksaan, Kementerian Pendidikan Malaysian; zaiyinulhayat@gmail.com*
[2]*Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia; suhaila.zainudin@ukm.edu.my*
[3]*Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia; zao@ukm.edu.my*

** Corresponding author*

To cite this article (APA): Zainal Rafit, Z. H., Zainudin, S., & Ali Othman, Z. (2021). Model peramalan bilangan calon tarik diri dari peperiksaan awam menerusi pendekatan perlombongan data dan petua. *Journal of ICT in Education, 8*(1), 26-42. https://doi.org/10.37134/jictie.vol8.1.3.2021

To link to this article: https://doi.org/10.37134/jictie.vol8.1.3.2021

Abstrak

Peningkatan jumlah data tersimpan tentang hal akademik pelajar bermula dari tahap sekolah rendah, sekolah menengah, kolej hingga universiti telah dibantu dengan perkembangan teknologi storan data. Data yang pelbagai ini wajar diekstrak ke dalam pengetahuan yang membantu dalam pembuatan keputusan dari pelbagai peringkat. Isu calon yang tidak hadir peperiksaan umum adalah masalah berkala dan memberi kesan kepada usaha mengoptimumkan kos pengurusan dalam era perbelanjaan berhemah. Sebanyak 15 peratus hingga 30 peratus calon tidak hadir peperiksaan awam berdasarkan analisis ke atas data untuk sepuluh tahun dari 2007 hingga 2016. Memandangkan masalah ini wujud saban tahun, kajian ini mencadangkan penyelesaian menggunakan perlombongan data bagi membangunkan model untuk meramal calon yang berpotensi tidak hadir peperiksaan awam (Sijil Tinggi Agama Malaysia atau STAM). Pendekatan yang dicadang terdiri dari enam langkah; bermula dari pemahaman bisnes, pemahaman data, penyediaan data, permodelan, penilaian dan pengerahan. Keputusan mendapati 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. Ramalan ini telah disimulasi dan didapati boleh menjimatkan 10 peratus dari kos percetakan kertas soalan dan buku jawapan untuk peperiksaan STAM. Hasil kajian ini boleh dimanfaat oleh Lembaga Peperiksaan dengan membuat unjuran perbelanjaan setiap tahun bagi operasi peperiksaan dengan menggunakan model ramalan bilangan calon tarik diri. Manakala, pihak sekolah dan ibu bapa boleh memanfaatkan kajian ini untuk mencari kaedah bagi menambah baik pencapaian akademik pelajar.

Kata Kunci: perlombongan data, peramalan, model, petua, CRISP-DM.

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Abstract

Advances in data storage technology have spurred growth in stored data on students' academic for primary, secondary, college, and university levels. The Education Ministry could extract the rich data into knowledge to aid decision making at different levels. The issue of candidates who did not attend public exams is a recurrent problem and affects the efforts to optimize management cost in the era of prudent spending. Around 15 percent to 30 percent candidates did not attend the public exam based on a ten-year analysis from 2007 until 2016. Since the problem recurs yearly, this research proposes a solution based on data mining to develop a model that predicts candidate who has a high potential of not attending a public examination exam (i.e. Sijil Tinggi Agama Malaysia or STAM). The proposed approach has six steps; business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The results discovered that the grades obtained for English, Mathematics and Science in the public examination taken in Form 5 (Sijil Pelajaran Malaysia) are the main factors for the non-attendance of a candidate for the STAM examination. The research then implemented a rule-based model based on these factors. Model evaluation proves that the model can predict the number of candidates that may not attend the examination. The prediction has been simulated and can save up to 10 percent of the printing cost for exam papers and answer books for the STAM examination. These results can benefit Lembaga Peperiksaan by projecting the yearly expenditure for exam operations using the non-attendance candidates' prediction model. Additionally, the school and parents may use this research to improve the students' academic performance.

Keywords: data mining, prediction, model, rules, CRISP-DM.

PENGENALAN

Perkembangan penyelidikan teknologi perlombongan data merangkumi pelbagai bidang untuk meneroka pengetahuan baharu daripada set data yang sangat besar. Kualiti pengurusan dan pentadbiran dapat ditingkatkan untuk menhasilkan prosedur pembuatan keputusan yang lebih berkesan (Fischer et al., 2020). Maklumat dan pengetahuan dari sumber data pendidikan amat membantu pihak pengurusan dalam pembuatan keputusan dari pelbagai peringkat. Perkembangan teknologi domain perlombongan data turut memacu kajian perlombongan data pendidikan dan seterusnya membaiki kualiti proses pendidikan (Asif et al., 2017).

Perlombongan Data Pendidikan (PDP) mengaplikasi kaedah dan teknik dari domain perlombongan data, pembelajaran mesin, statistik, capaian maklumat, psikologi kognitif, pedagogi-psiko dan sistem pencadang ke atas set data pendidikan dan menyelesaikan isu-isu pendidikan (Dutt et al., 2017). Manakala, perkembangan terkni dalam Big Data dalam pendidikan mencetuskan kaedah baru yang memanfaatkan data bagi menyokong pembuatan keputusan berasaskan maklumat dan meningkatkan keberkesanan kaedah pendidikan (Fischer et al., 2020).

Sebagai contoh untuk pembuatan keputusan bagi domain pendidikan dari konteks penempatan guru di sekolah; Saniron dan Ali Othman (2019) telah membuat ujikaji ke atas teknik pengelasan terbaik dan penerokaan pengetahuan baharu ke atas data permohonan penempatan di Kementerian Pendidikan Malaysia (KPM). Hasil kajian mendapati teknik Kstar adalah teknik terbaik untuk pengelasan

penempatan guru dan dapat membantu pembuatan keputusan tentang penempatan guru di peringkat KPM.

Dari konteks pelajar, kajian perlombongan data bidang pendidikan meliputi pelbagai aspek seperti penggunaan perisian pendidikan, pembelajaran berpandukan komputer, peramalan akademik pelajar dan juga isu berkaitan minat pelajar di dalam jurusan yang diikuti (Zainudin & Ab Aziz; 2015). Di peringkat pengajian tinggi, Ahmad dan Abu Bakar (2018) telah membangunkan model pengelasan untuk pembuatan keputusan bagi biasiswa pengajian tinggi untuk membantu pemegang taruh di kementerian yang terbabit dalam penentuan polisi baharu di masa depan.

LATAR BELAKANG KAJIAN

Sijil Tinggi Agama Malaysia atau STAM merupakan peperiksaan awam bagi pelajar dari aliran agama di Malaysia. STAM dilaksana sejak tahun 2000 hasil kerjasama antara Kementerian Pendidikan Malaysia dan Universiti al-Azhar, Kaherah, Mesir. Syarat utama peperiksaan ini adalah calon STAM perlu mempunyai kelulusan peperiksaan Sijil Pelajaran Malaysia (SPM). Pemegang sijil STAM berpeluang untuk melanjutkan pelajaran ke Universiti al-Azhar, Kaherah, Mesir atau Universiti Awam di Malaysia.

Lembaga Peperiksaan (LP) adalah badan yang menguruskan pembekalan kertas soalan dan buku jawapan calon peperiksaan STAM berdasarkan data pendaftaran calon. Pada bulan Mei setiap tahun, LP akan membuat tempahan percetakan buku soalan dan buku jawapan dengan pencetak berdasarkan bilangan calon berdaftar. LP akan melaksana urusan penyemakan dan penghantaran soalan bercetak ke bilik kebal di setiap Jabatan Pendidikan Negeri (JPN) seluruh negara pada bulan Mei juga. Peperiksaan bertulis STAM diadakan pada bulan Oktober. Pihak sekolah akan menghantar surat makluman calon tarik diri dari menduduki peperiksaan STAM kepada LP selepas bulan Mei. Pencetak akan melakukan cetakan berdasarkan rekod asal bilangan calon yang berbeza dengan bilangan sebenar calon yang bakal menduduki peperiksaan.

Definisi calon tarik diri adalah calon yang tidak hadir menduduki kesemua kertas peperiksaan STAM. Definisi calon tidak tarik diri adalah calon yang hadir dan menduduki peperiksaan STAM. Calon yang tidak hadir untuk sebilangan kertas peperiksaan STAM tidak termasuk di dalam kategori calon tarik diri kerana calon tersebut akan mendapat keputusan dan pangkat persijilan berdasarkan kertas peperiksaan yang diduduki. Data peperiksaan bagi tahun 2007 sehingga tahun 2016 yang diperoleh dari LP menunjukkan bilangan calon yang tidak hadir menduduki peperiksaan STAM adalah tinggi iaitu pada kadar 15 peratus sehingga 30 peratus seperti dalam Jadual 1.

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Jadual 1: Bilangan dan peratus kehadiran calon peperiksaan STAM dari tahun 2007 sehingga tahun 2016.

Tahun	Bilangan Pendaftaran Calon	Calon Hadir		Calon Tarik Diri	
		Bilangan calon	Peratus %	Bilangan calon	Peratus %
2007	5,140	3,719	72.35	1,421	27.65
2008	5,103	3,669	71.90	1,434	28.10
2009	5,107	3,546	69.43	1,561	30.57
2010	6,345	4,913	77.43	1,432	22.57
2011	6,258	4,541	72.56	1,717	27.44
2012	6,505	4,828	74.22	1,677	25.78
2013	6,627	4,976	75.09	1,651	24.91
2014	6,784	5,057	74.54	1,727	25.46
2015	6,588	5,223	79.28	1,365	20.72
2016	6,396	5,327	83.31	1,067	16.69

Untuk memahami implikasi kos tentang calon tidak hadir peperiksaan, anggaran kos yang terlibat dibuat seperti berikut. Anggaran kos percetakan satu buku soalan pada harga RM3.00 dan buku jawapan adalah RM3.00. Katakan seorang calon STAM akan menduduki 13 kertas bertulis bagi 10 subjek. Berdasarkan kiraan berikut, pihak LP akan membelanjakan sebanyak RM78.00 untuk menyediakan buku soalan dan buku jawapan.

Perbelanjaan seorang calon, B
Harga buku soalan, H1
Harga buku jawapan, H2
Bilangan kertas bertulis, K
Perbelanjaan;

B = (harga buku soalan + buku jawapan) x bilangan kertas bertulis seorang calon
B = (H1 + H2) x K
= (RM3 + RM3) x 13
= RM6 x 13
= RM78

Berdasarkan Jadual 1, bilangan calon yang mendaftar untuk peperiksaan STAM pada tahun 2016 ialah seramai 6,396 orang.

Bilangan calon STAM tahun 2016, C2016 = 6,396 calon
Maka, perbelanjaan bagi tahun 2016, B2016 ialah
B2016 = (B) x (C2016)
= (RM78) x (6,396)
= RM498888.00

Sekiranya perbelanjaan dibuat hanya untuk calon hadir, Bhadir2016 sahaja,

Bhadir2016 = (B) x (Chadir2016)
= (RM78) x (5,327)
= RM415506.00

Maka kos yang dapat dijimatkan oleh LP ialah;

Jimat = RM498,888 – RM415,506
= RM83382.00

Kerugian yang ditanggung oleh LP untuk kos percetakan calon tidak hadir peperiksaan STAM adalah signifikan iaitu RM83382.00 dan satu inisiatif perlu dilaksana untuk menangani kerugian ini. Sehubungan itu, kajian perlombongan data dilaksana bertujuan untuk meramal faktor yang berkaitan dengan calon tarik diri. Faktor yang diramal akan dimanfaatkan sebagai elemen dalam model perlombongan data untuk meramal calon yang berpotensi tidak hadir peperiksaan STAM.

KAJIAN RAMALAN PRESRASI AKADEMIK PELAJAR PELBAGAI PERINGKAT

Analisis peramalan yang digunakan oleh pihak pengurusan pelajar mengandungi data yang terdiri daripada campuran pelbagai maklumat tentang pelajar. Berdasarkan maklumat ini, kajian dalam PDP secara majoriti bertumpu kepada meramal pencapaian pelajar menggunakan teknik perlombongan data (Aydoğdu, 2020). Selain dari itu, teknik pembelajaran mesin banyak digunakan untuk menganalisis aktiviti pembelajaran dalam talian internet bagi meramal prestasi pembelajaran pelajar di institusi pengajian tinggi. Antara kajian yang telah dilaksana adalah meramal keciciran pelajar Massive Open Online Courses (MOOC) (Koedinger et al., 2015). Hasil analisis peramalan pelajar MOOC yang tercicir mencadangkan bahawa faktor hubungan sosial memberi kesan kepada keciciran pelajar. Ini membuahkan idea agar rekabentuk MOOC menerapkan aspek keterlibatan sosial yang menarik komitmen pelajar untuk menyempurnakan pembelajaran menerusi MOOC.

PDP lazimnya mengaplikasi algoritma pembelajaran mesin terhadap data keputusan pencapaian murid terdahulu untuk menjana model peramalan bagi meramal keputusan pembelajaran murid lain pada masa hadapan (Uswatun Khasanah & Harwati, 2017). Menurut Uswatun Khasanah dan Harwati (2017), kajian PDP membantu pihak pengajar mengenal pasti murid tercicir dalam pembelajaran atau tarik diri dari meneruskan pembelajaran bagi sesuatu kursus dan membantu murid dengan memberikan bantuan pengajaran yang sewajarnya. PDP juga boleh mendapatkan ramalan awal keciciran dan pemberhentian pembelajaran ketika awal pembelajaran kursus dijalankan dengan penghasilan model pengelasan yang boleh dipercayai (Márquez-Vera et al., 2016).

Algoritma Pokok Keputusan telah digunakan dalam perlombongan data pencapaian terdahulu dalam kajian Uswatun Khasanah dan Harwati (2017), untuk menjana model yang digunakan untuk meramal pencapaian murid. Manakala, Asif et al. (2017) menggunakan teknik Pokok Keputusan, Rangkaian Neural dan Naif Bayes untuk menentukan ramalan keputusan pelajar di universiti awam di Pakistan. Kajian mereka mendapati teknik Naif Bayes memberikan ketepatan ukuran yang lebih baik. Dapatan utama adalah perlu untuk fokus kepada sebilangan kecil kursus yang menjadi indikator kepada prestasi akademik pelajar supaya amaran dan sokongan boleh diberikan kepada pelajar berprestasi rendah manakala nasihat dan peluang diberikan kepada pelajar berprestasi tinggi.

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Dari konteks Malaysia, Buniyamin et al., (2016) telah melaksana kajian untuk meramal pencapaian akademik terhadap 391 orang pelajar dari Fakuti Kejuruteraan Elektrik, Universiti Teknologi Mara (UiTM) dengan menggunakan teknik pengelasan rangkaian neuro-fuzzy. Mereka menetapkan bilangan aturan adalah sama dengan kluster pengeluaran. Kajian yang dijalankan oleh Abu Saa (2016) untuk mengetahui faktor-faktor yang mempengaruhi pencapaian keputusan pelajar di Universiti Sains dan Teknologi Ajman, Emiriah Arab Bersatu (UAE) turut menggunakan teknik Pokok Keputusan dalam perlombongan data.

Kajian yang dijalankan oleh Costa et al. (2017) adalah untuk menentukan teknik yang paling tepat dan berkesan bagi mendapatkan ramalan pelajar yang akan gagal dalam kursus pengenalan pangaturcaraan. Dengan menggunakan empat teknik perlombongan data iaitu Naïf Bayes, Pokok Keputusan, Rangkaian Neural Buatan dan Mesin Sokongan Vektor terhadap data ujian dan latihan yang didapatkan daripada dalam talian sehingga model pembelajaran mesin diperoleh. Model dan algoritma itu diuji terhadap data pelajar universiti awam Brazil. Kajian menentukan teknik dan algoritma terbaik bagi membuat ramalan awal calon yang akan gagal dalam pembelajaran yang dijalankan oleh Márquez-Vera et al., (2016) dengan menggunakan teknik pengelasan terhadap 419 orang murid tahun pertama sekolah tinggi di Mexico.

Kajian yang menggunakan atribut gred keputusan peperiksaan sahaja turut dijalankan menggunakan PDP. Kajian untuk meramal keputusan peperiksaan calon SPM terhadap murid di Maktab Rendah Sains MARA telah dijalankan oleh Makhtar et al. (2017) terhadap subjek yang didaftarkan untuk peperiksaan SPM dengan menggunakan teknik Naïf Bayes. Kajian terhadap gred keputusan peperiksaan juga dijalankan oleh Marbouti et al. (2016) terhadap 120 pelajar tahun pertama pengajian kejuruteraan di universiti bahagian selatan Amerika Syarikat menggunakan teknik Mesin Sokongan Vektor, Pokok Keputusan, Rangkaian Neural, Naïf Bayes dan K-Jiran Terdekat.

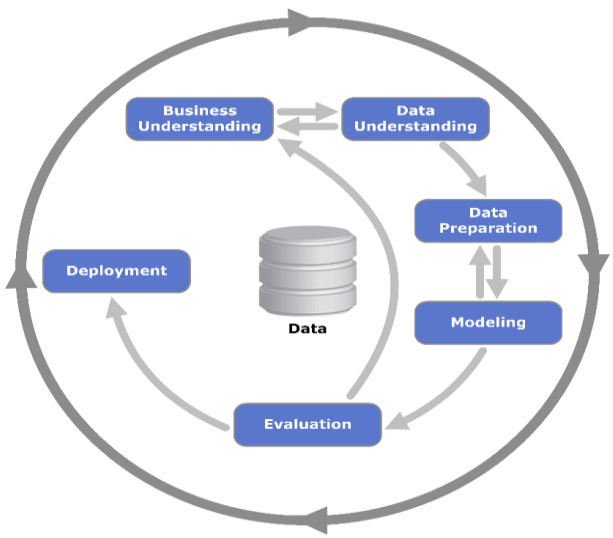
Uswatun Khasanah dan Harwati (2017) melaksanakan kajian meramal bilangan pelajar universiti yang kandas pengajian di Indonesia. Bilangan pelajar kandas penting kerana menjadi satu aspek dalam akreditasi universiti Indonesia. Kajian ini membuat ramalan menggunakan Rangkaian Bayesian dan Pokok Keputusan. Ismail et al. (2021) mengkaji prestasi teknik Pokok Keputusan, Naïf Bayes, Rangkaian Neural Buatan, Mesin Sokongan Vektor dan Pohon Rawak berdasarkan ketepatan, F-measure dan tempoh perlaksanaan. Kajian Ismail (2021) menyimpulkan bahawa teknik Mesin Sokongan Vektor linear, Mesin Sokongan Polinomial dan Naif Bayes mengatasi teknik lain untuk set data bersaiz kecil (480 sampel). Manakala, Pokok Keputusan dan Pohon Rawak adalah terbaik untuk set data bersaiz besar (1044 sampel).

Teknik-teknik yang telah digunakan dalam kajian lepas ini digunakan sebagai asas bagi pemilihan teknik untuk menganalisis data di dalam kajian ini. Antara teknik yang difokuskan berdasarkan kajian lepas adalah Pokok Keputusan, Hutan Rawak, Naif Bayes. dan Rangkaian Neural.

METODOLOGI KAJIAN

Kajian dilaksana mengikut metodologi perlombongan data yang digunakan secara meluas di dalam kajian perlombongan data yang dinamakan Cross Industry Standard Process for Data Mining (CRISP-DM) (Martínez-Plumed et al., 2019). CRISP-DM adalah pendekatan umum untuk menjawab persoalan dengan berpandukan data. Kajian ini memilih CRISP-DM berdasarkan kelebihan kaedah ini yang menyediakan kerangka kerja lengkap yang terdiri dari proses perlombongan data yang matang dan sesuai dilaksana dalam projek dari pelbagai domain.

Terdapat 6 fasa dalam CRISP-DM akan diterangkan dalam konteks penghasilan model calon tarik diri dari peperiksaan awam (Rajah 1).



Rajah 1: Fasa dalam CRISP-DM (Martínez-Plumed et al., 2019).

Fasa 1 adalah memahami objektif dan keperluan dari perspektif bisnes (*business understanding*). Kemudian mengubah pengetahuan ke dalam definisi masalah perlombongan data. Rekabentuk perancangan projek dilaksana untuk mencapai objekif. Masalah digarap sebagai membangunkan model calon yang berpotensi untuk tidak hadir peperiksaan awam tingkatan enam (Sijil Tinggi Agama Malaysia atau STAM) dalam konteks kajian ini.

Fasa 2 adalah memahami data (*data understanding*) yang bermula dengan pengumpulan dan pemahaman data, mengenalpasti masalah kualiti data, mendalami pemahaman diri (*insight*) berkenaan data. Fasa 2 berkait rapat dengan Fasa 1 khususnya dalam meformulasi masalah perlombongan data dan perancangan projek yang berkait dengan pemahaman data yang sedia ada. Data diperolehi dengan membuat permohonan rasmi kepada Lembaga Peperiksaan.

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Permohonan dibuat untuk mendapatkan set data mentah keputusan peperiksaan SPM dan STAM dari tahun 2006 sehingga 2017. Set data yang dimohon kepada pihak LP ialah set data keputusan peperiksaan SPM dan STAM bagi sebelas kohort iaitu dari tahun 2006 sehingga tahun 2017. Data keputusan SPM bermula dari tahun 2006 sehingga tahun 2015 dan data keputusan STAM bermula dari tahun 2007 sehingga 2016 dijadikan set data latihan dan data ujian model. Manakala, keputusan peperiksaan SPM tahun 2016 dan keputusan peperiksaan STAM tahun 2017 dijadikan data pengesahan model.

Oleh kerana data peperiksaan adalah data sulit, proses mendapatkan data daripada pangkalan data peperiksaan kendalian LP hanya boleh dilakukan oleh pegawai LP. Semasa permohonan data dibuat, senarai atribut yang diperlukan telah disertakan (Jadual 2). Ini memudahkan pihak LP mengekstrak data untuk kajian.

Jadual 2: Senarai atribut yang diminta daripada LP.

Set Data SPM	**Set Data STAM**
Keputusan Peperiksaan SPM	**Keputusan Peperiksaan STAM**
{Nama}	{Nama}
{Jantina}	{Jantina}
{No. Kad Pengenalan}	{No. Kad Pengenalan}
{Kod Pusat}	{Kod Pusat}
{Jumlah Mata Pelajaran}	{Pangkat Sijil}
{Kod Mata Pelajaran dan Gred} x 20	{Tahun Peperiksaan}
{Tahun Peperiksaan}	

Kajian ini mengkaji data keputusan peperiksaan SPM dan STAM dari tahun 2006 sehingga tahun 2016 untuk dilombong dengan tujuan meramal calon yang akan menarik diri dari menduduki peperiksaan STAM bagi data tahun 2017. Jumlah data yang diterima daripada pihak LP adalah sebanyak 67,947 calon dengan 43 atribut bagi kedua-dua set data keputusan peperiksaan. Data calon bagi kedua-dua peperiksaan adalah calon yang berpadanan. Contohnya, calon A yang berada di set data peperiksaan STAM 2007 mesti mempunyai data di set data peperiksaan SPM 2006.

Fasa 3 (*data preparation*) adalah menyediakan data yang melibatkan semua aktiviti untuk mengubah data mentah ke set data akhir. Aktiviti penyediaan data termasuk pemilihan jadual, rekod dan atribut, pembersihan data, membangunkan atribut baharu, mengubah data ke dalam format yang sesuai untuk alat permodelan dan lain-lain. Aktiviti ini berkemungkinan dilaksana berulang kali. Aktiviti utama yang telah dijalankan untuk set data ini diterangkan seperti berikut:

i. *Penyepaduan Set Data SPM dengan Set Data STAM*.
   Penyepaduan dua set data ini dilakukan dengan padanan nombor kad pengenalan calon. Nombor kad pengenalan yang sepadan dari dua set data (SPM dan STAM) akan disepadukan menjadi satu set data kajian.

ii. *Penyusunan Subjek SPM dan Gred Mengikut Atribut Pelajaran.*
Kod subjek yang sama disusun ke dalam satu atribut mengikut kod subjek bersekali dengan gred.

iii. *Pembersihan Data untuk Pengecilan Saiz Data.*
Proses ini membuang atribut dan data yang tidak relevan atau tidak lengkap. Proses juga mengenal pasti calon yang mempunyai data yang tidak lengkap seperti bilangan subjek yang didaftarkan kurang daripada 6 subjek. Data calon ini dikeluarkan memandangkan subjek dan gred tidak diboleh diganti dengan sebarang nilai rawak. Subjek yang tidak menjadi subjek utama yang didaftarkan oleh calon dikeluarkan. Subjek yang kurang daripada 50 peratus daripada jumlah bilangan calon dikeluarkan bagi mengelakkan data subjek yang tidak seimbang.

iv. *Penukaran Gred Keputusan SPM dan STAM dari Gred Huruf Kepada Gred Nombor.*
Penukaran gred kepada bentuk nominal membolehkan pembelajaran mesin dengan mudah menjalankan algoritma pembelajaran.

v. *Pengagregatan Subjek dan Gred Subjek SPM.*
Beberapa subjek akan diagregat kepada satu subjek mengikut bidang aliran. Subjek bidang agama digabungkan dan ditambah satu kod subjek. Sebagai contoh, subjek Tasawwur Islam, Pendidikan Al-Quran dan Al-Sunnah dan diagregat atau digabungkan menjadi SYAQS. Proses sama dilaksana untuk beberapa subjek bidang sains tulen seperti Fizik, Kimia dan Biologi diagregat kepada satu kod iaitu SCTULEN. Manakala, pengagregatan data gred subjek dilakukan dengan menggunakan nilai purata bagi semua gred tersebut. Pengagregatan data atribut gred subjek dilakukan dengan membuat purata bagi semua gred tersebut. Gred subjek SPM juga diagregatkan bagi mengurangkan bilangan nod yang akan diaplikasikan dengan algoritma teknik pengelasan. Gred asal yang berjumlah 11 gred iaitu gred A+, A, A-, B+, B, C+, C, D, E, G dan T diagregatkan kepada 1, 2, 3, 4, 5 dan 6.

vi. *Mengekstrak Data Mengikut Aliran.*
Pengekstrakan data dilakukan dengan mengasingkan data calon mengikut pakej subjek aliran. Dua set data diwujudkan iaitu set data aliran agama dan set data aliran agama sains. Set data aliran agama mempunyai atribut BM, BI, SEJ, MATHS, SAINS, BA dan SYAQS dari data peperiksaan SPM dan pangkat STAM dari data peperiksaan STAM. Manakala set data aliran agama sains mempunyai atribut BM, BI, SEJ, MATHS, SCTULEN, BA dan SYAQS dari data peperiksaan SPM dan pangkat STAM dari data peperiksaan STAM. Jadual 3 adalah contoh untuk set data aliran agama.

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Jadual 3: Set data aliran agama.

BM	BI	SEJ	MATHS	SAINS	BA	SYAQS	Pangkat STAM
3	5	4	5	4	2	3	2
2	5	5	5	4	5	4	4
4	5	5	5	6	6	4	5
5	5	5	5	5	5	5	6
4	5	3	5	4	4	3	3

Atribut pangkat STAM akan dijadikan Atribut Kelas Ramalan sama ada calon akan hadir atau tidak ke peperiksaan STAM. Pangkat STAM ditukar kepada atribut kelas TARIK DIRI yang mempunyai kelas YA atau TIDAK. Pangkat STAM antara 1 hingga 5 ditukar kepada kelas TIDAK mewakili calon yang tidak tarik diri manakala pangkat STAM bersamaan 6 ditukar kepada kelas YA mewakili calon yang tarik diri. Setelah penukaran ini dilaksana, hasil terakhir untuk set data aliran agama dalam Jadual 4.

Jadual 4: Set data aliran agama dengan atribut Tarik Diri.

BM	BI	SEJ	MATHS	SAINS	BA	SYAQS	Pangkat STAM
3	5	4	5	4	2	3	TIDAK
2	5	5	5	4	5	4	TIDAK
4	5	5	5	6	6	4	TIDAK
5	5	5	5	5	5	5	YA
4	5	3	5	4	4	3	TIDAK

Fasa 4 (*modeling*) adalah permodelan di mana beberapa teknik permodelan komputeran cerdas dipilih, diaplikasi dan parameter dikalibrasi kepada nilai optimum. Fasa 4 berkait rapat dengan Fasa 3 di mana masalah ditemui semasa permodelan atau idea baharu tercetus untuk membentuk data baharu. Langkah pertama dalam membina model adalah menggunakan algoritma CfsSubsetEval-BestFirst sebagai penilai Pemilihan Atribut di dalam aplikasi WEKA (https://www.cs.waikato.ac.nz/ml/weka/). Bagi set data aliran agama, atribut yang utama ialah atribut BI, MATHS dan SAINS manakala bagi set data aliran agama sains, atribut yang utama ialah MATHS dan SCTULEN.

Justifikasi untuk pemilihan teknik pengelasan adalah memilih teknik yang paling banyak digunakan bagi menghasilkan faktor-faktor penting peramalan dari kajian lepas. Teknik yang dipilih berdasarkan justifikasi ini adalah Pokok Keputusan. Algoritma pengelasan teknik Pokok Keputusan yang boleh menghasilkan petua pengelasan dan diaplikasikan terhadap data pendidikan yang terdapat dalam perisian WEKA adalah algoritma J48 dan JRip. Sehubungan itu, kajian ini dijalankan dengan menjadikan kajian oleh Márquez-Vera et al., (2016) dan Makhtar et al., (2017) sebagai rujukan utama.

Kajian ini menjalankan dua kali ujian bagi setiap set data. Ujian pertama dijalankan dengan menggunakan semua atribut. Ujian kedua dijalankan hanya menggunakan 3 atribut utama. Bagi set data aliran agama, atribut utama ialah BI, MATHS dan SAINS manakala bagi set data aliran agama sains pula, atribut yang utama ialah MATHS dan SCTULEN. Akan tetapi ujian menggunakan atribut utama bagi set data aliran agama sains tetap memasukkan atribut BI bagi mengelakkan ujian berat sebelah dalam kajian.

Fasa 5 (*evaluation*) adalah menilai model terbaik dari beberapa model berkualiti tinggi yang telah dibangunkan. Sebelum memasuki Fasa 6 (pengerahan), model dinilai dari segi langkah-langkah pembangunan model supaya objektif untuk membangunkan model calon yang berpotensi untuk tidak hadir peperiksaan awam tingkatan enam (Sijil Tinggi Agama Malaysia atau STAM) tercapai.

Fasa 6 adalah pengerahan (*deployment*). Pengetahuan yang didapati dari fasa sebelum perlu diselaraskan dan dipersembahkan supaya boleh digunakan oleh pengguna akhir. Hasil akhir fasa 6 boleh berbentuk menjana laporan atau melaksana proses perlombongan data yang boleh berulang. Pemahaman tentang tindakan yang perlu dilaksana untuk memanfaatkan model yang terhasil.

HASIL DAN KEPUTUSAN KAJIAN

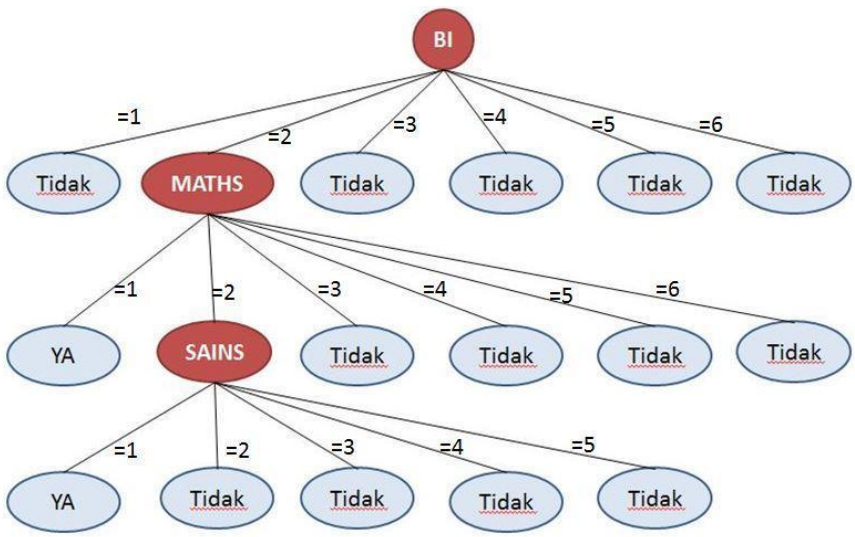
Jadual 5 menunjukkan keputusan ujian pembelajaran mesin terhadap set data aliran agama dan set data aliran agama sains. Keputusan ujian menunjukkan ketepatan pengelasan model pengelasan merujuk kepada nilai kadar TP atau *True Positive* (bagi kelas calon hadir menduduki peperiksaan STAM atau TIDAK) dan peratus pengelasan yang betul. Teknik pengelasan dari domain perlombongan data yang digunakan ialah teknik Pokok Keputusan (J48), Petua Pengelasan (JRip), Hutan Rawak (RF), Naif Bayes (NB) dan Rangkaian Neural (NN) yang terdapat dalam perisian WEKA (https://www.cs.waikato.ac.nz/ml/weka/). Teknik-teknik ini juga antara teknik yang lazim digunakan dan telah diulas dalam kajian lepas**.**

Jadual 5: Keputusan ujian teknik pengelasan.

	Pengujian Model	**Set data aliran agama**		**Set data agama sains**	
		Semua atribut	3 atribut terbaik	Semua atribut	3 atribut terbaik
Kadar TP	J48	0.981	0.984	0.806	0.794
	JRip	0.980	0.982	0.795	0.777
	RF	0.936	0.977	0.721	0.784
	NB	0.838	0.930	0.651	0.723
	NN	0.982	0.988	0.816	0.799

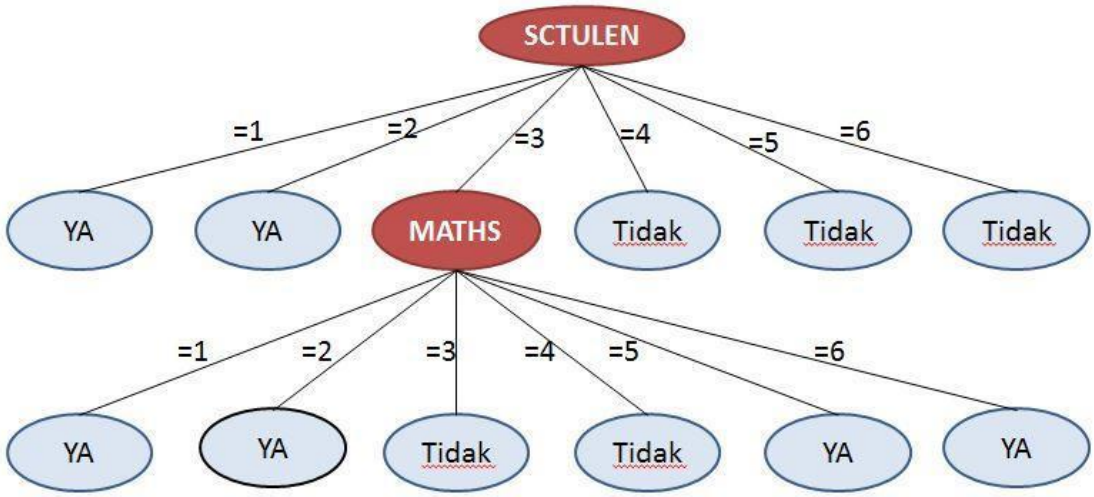
Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

Keputusan dalam Jadual 5 menunjukkan algoritma J48 (Pokok Keputusan) menghasilkan peratus pengelasan yang terbaik dan konsisten dengan nilai kadar TP setinggi 0.981 untuk semua atribut dan 0.984 bagi 3 atribut terbaik (set data aliran agama) dan 0.806 untuk semua atribut (set data aliran agama sains). Oleh itu, petua pengelasan yang terhasil daripada algoritma Pokok Keputusan digunakan sebagai model ramalan calon tarik diri daripada menduduki peperiksaan STAM.



Rajah 2: Model bagi set data aliran agama.

Berdasarkan model Pokok Keputusan dalam Rajah 2 menunjukkan atribut BI, MATHS dan SAINS (nod merah) menjadi faktor utama bagi bagi set data aliran agama. Atribut BI merupakan nod terawal pokok dan faktor paling dominan. Untuk keputusan Tidak yang mewakili calon tidak tarik diri, beberapa petua telah dijana. Pada lapisan pertama pokok keputusan, calon tidak tarik diri apabila mendapat keputusan gred mata pelajaran BI=1, 3, 4, 5 dan 6. Manakala pada lapisan kedua, calon tidak tarik diri apabila mendapat BI = 2 dan MATHS=3, 4, 5 dan 6. Pada lapisan ketiga, calon tidak tarik diri apabila mendapat BI = 2 dan MATHS = 2 dan SAINS = 2,3,4,5. Terdapat 3 petua dijana menerusi pokok keputusan berkaitan calon tidak tarik diri. Pokok keputusan menunjukkan YA iaitu calon tarik diri menerusi 2 petua berikut. Pertama, calon mendapat keputusan gred mata pelajaran BI=2 dan MATHS=2 atau kedua, calon mencapai keputusan gred mata pelajaran BI= 2 dan MATHS=1 dan SAINS=1.



Rajah 3: Model bagi set data aliran agama sains.

Rajah 3 menunjukkan atribut SCTULEN dan MATHS telah menjadi faktor utama bagi bagi set data aliran agama sains. Atribut SCTULEN merupakan nod terawal pokok dan faktor paling dominan. Pokok keputusan menunjukkan TIDAK iaitu calon tidak tarik diri apabila mendapat keputusan gred mata pelajaran SCTULEN=4, 5 dan 6. Calon tidak tarik diri apabila mendapat keputusan gred mata pelajaran SCTULEN=3 dan MATHS=2, 3 dan 4. Manakala pokok keputusan menunjukkan YA iaitu calon tarik diri apabila mendapat keputusan gred mata pelajaran SCTULEN=1 dan 2. Calon juga tarik diri apabila keputusan gred mata pelajaran SCTULEN=gred2 dan MATHS=gred 1, 5 dan 6. Pengetahuan yang telah dilombong ini digarap dalam petua berbentuk JIKA-MAKA. Terdapat 2 petua terhasil daripada set data aliran agama iaitu:

JIKA (BI=2 AND DAN MATHS=2) MAKA TarikDiri=YES

JIKA (BI=2 AND DAN MATHS=2 DAN SAINS=1) MAKA TarikDiri=YES

Manakala 3 petua terhasil daripada set data aliran agama sains iaitu:

JIKA (SCTULEN$\leqslant$2) MAKA TarikDiri=YES

JIKA (SCTULEN=3 DAN MATHS=1) MAKA TarikDiri=YES

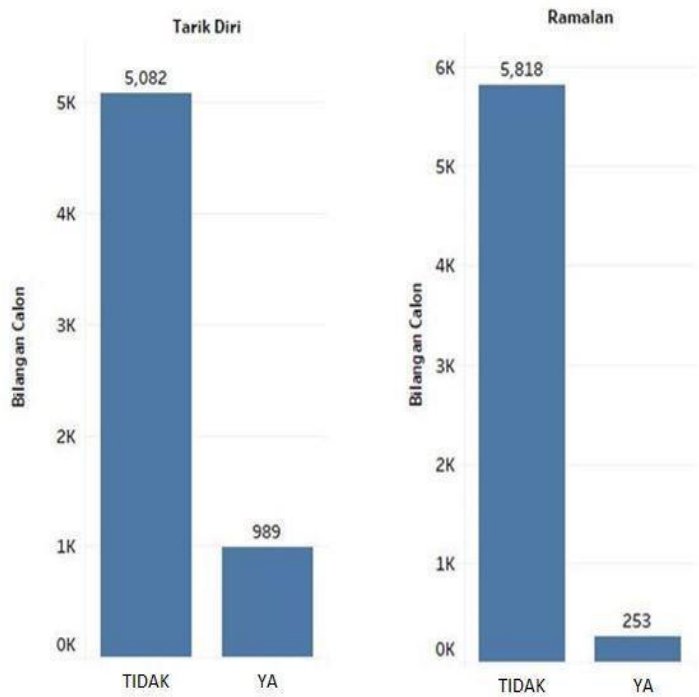
JIKA (SCTULEN=3 DAN MATHS$\geqslant$5) MAKA TarikDiri=YES

Received Date: 31 Jan 2021; Accepted Date: 31 May 2021; Published Date: 09 June 2021

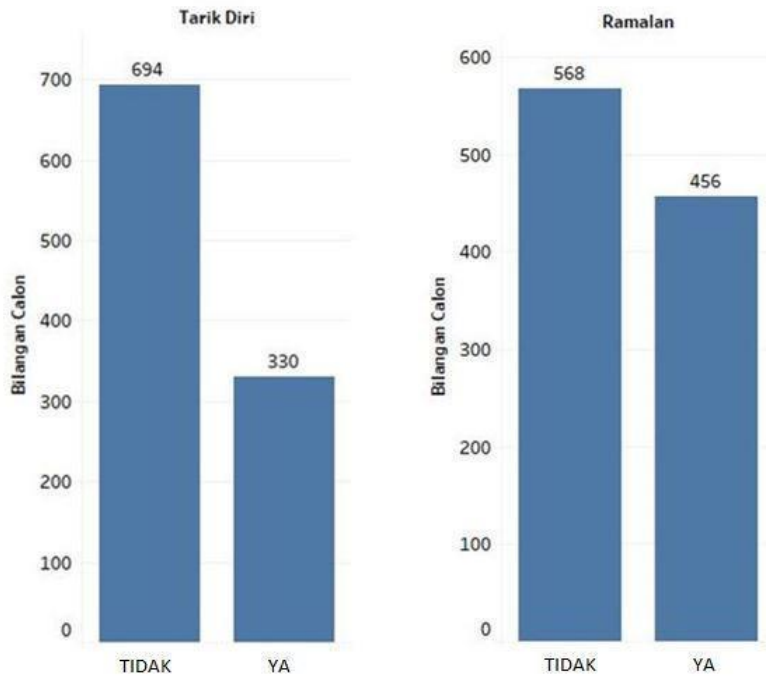
Petua JIKA-MAKA ini adalah model ramalan calon tarik diri. Model kemudiannya diaplikasikan terhadap set data pengesahan. Set data pengesahan adalah set data keputusan peperiksaan SPM tahun 2016 untuk meramal calon tarik diri bagi peperiksaan STAM tahun 2017. Keputusan yang terhasil dibandingkan dengan keputusan sebenar calon bagi peperiksaan STAM 2017.

Terdapat perbezaan bilangan calon di antara bilangan calon yang diramal menggunakan keputusan peperiksaan SPM tahun 2016 dengan bilangan calon sebenar keputusan peperiksaan STAM tahun 2017. Calon yang tidak tarik diri merupakan calon yang hadir menduduki peperiksaan dan mendapat keputusan STAM tahun 2017. Model telah meramal 253 orang calon dari aliran agama yang dijangka menarik diri dari peperiksaan STAM 2017, namun bilangan sebenar calon menarik diri adalah 989 orang. Peratus perbezaan antara ramalan dan sebenar adalah 74.42 peratus.

Untuk aliran agama sains, seramai 456 orang calon diramal akan menarik diri, namun sebenarnya seramai 330 orang calon telah menarik diri. Peratus perbezaan adalah 38.18 peratus. Terdapat perbezaan yang agak ketara antara nilai ramalan dan nilai sebenar (Rajah 4 dan Rajah 5).



Rajah 4: Perbandingan bilangan sebenar dengan ramalan bagi set data aliran agama.



Rajah 5: Perbandingan bilangan sebenar dengan ramalan bagi set data aliran agama sains.

Berpandukan Rajah 4 dan Rajah 5, bilangan ramalan calon yang tidak tarik diri bagi kedua-dua aliran ialah 5818 + 568 = 6386 orang calon. Manakala bilangan ramalan calon yang tarik diri bagi kedua-dua aliran ialah 253 + 456 = 709 orang calon. Berdasarkan keputusan kajian, pengiraan perbelanjaan mengikut data pendaftaran adalah seperti berikut:

Katakan perbelanjaan untuk seorang calon, B ialah RM78.
Bilangan calon STAM 2017, C2017 = 7,095 orang calon.
Maka, perbelanjaan sebenar bagi tahun 2017, B2017 ialah

B2017 = (B) (C2017)
= (RM78) (7,095)
= RM553410

Perbelanjaan ramalan calon yang hadir, Bhadir = RM78(6386) = RM498108
Penjimatan perbelanjaan = perbelanjaan sebenar – perbelanjaan Bhadir
= RM553410 – RM498108
= RM55302
Peratus penjimatan LP = Penjimatan / (Perbelanjaan Sebenar) x100%
= 55,302/553,410 x 100% = 10%

Maka LP boleh menjimatkan perbelanjaan sebanyak 10 peratus bagi tahun 2017.

KESIMPULAN

Kajian ini secara amnya berjaya menghasilkan faktor penentu iaitu atribut subjek Bahasa Inggeris, Sains, Fizik, Kimia, Biologi dan Matematik bagi keputusan peperiksaan SPM yang menjadi faktor utama dalam model meramal potensi calon akan menarik diri atau tidak daripada menduduki peperiksaan STAM. Atribut ini adalah faktor peramal utama bagi dua data aliran calon SPM yang diuji. Pengaplikasian algoritma Pokok Keputusan J48 berpotensi membantu pihak LP dan sekolah dalam meramal bilangan calon tarik diri daripada peperiksaan STAM. Pihak LP dapat menjimatkan perbelanjaan bagi tahun 2017 sekurang-kurangnya 10 peratus berdasarkan simulasi.

Bagi pihak JPN, sekolah dan ibu bapa, mereka juga boleh memanfaatkan kaedah kajian ini untuk menambah baik pencapaian murid di dalam keputusan peperiksaan awam terutama peperiksaan STAM. Ramalan awal untuk mengenal potensi diri dapat membantu perancangan pembelajaran dan membaiki kaedah pengajaran dalam membantu meningkatkan prestasi pembelajaran murid.Pada masa akan datang, kajian yang selanjutnya dicadangkan untuk mendapatkan data lain yang lebih komprehensif seperti melihat kepada data demografi, keputusan tawaran ke institusi pengajian tinggi, sosio budaya murid dan tempatan, taraf ekonomi keluarga, aktiviti sosial, penglibatan dalam ko-kurikulum di dalam serta luar sekolah atau insitusi pengajian tinggi, psikologi, psikometrik dan interaksi jaringan sosial.

Bagi mengelakkan data yang tidak seimbang, dicadangkan juga kajian dilakukan terhadap set data mengikut negeri. Set data mengikut negeri boleh dijadikan sampel dengan mengambil kira bilangan calon yang tarik diri dengan tidak tarik diri adalah tidak terlalu banyak perbezaannya. Selain itu, juga dicadang mengambil sampel data yang seimbang atau sama banyak bagi kelas tarik diri dan tidak tarik diri menduduki peperiksaan STAM.

PENGHARGAAN

Penghargaan kepada dana penyelidikan GUP-2020-089 dan TT-2020-015 atas sokongan kepada kajian ini.

RUJUKAN

Abu Saa, A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications, 7*(5), 212–220. https://doi.org/10.14569/IJACSA.2016.070531

Ahmad, W. D., & Abu Bakar, A. (2018). Classification Models for Higher Learning Scholarship Award Decisions. *Asia-Pacific Journal of Information Technology & Multimedia, 7*(2). Penerbit Universiti Kebangsaan Malaysia (UKM Press), Dec. 2018, pp. 131–145. Crossref, doi:10.17576/apjitm-2018-0702-10.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Aydoğdu, Ş. (2020). Educational Data Mining Studies in Turkey: A Systematic Review. *Turkish Online Journal of Distance Education, 21*(3), 170-185. DOI: 10.17718/tojde.762046

Buniyamin, N., Mat, U. Bin, & Arshad, P. M. (2016). Educational data mining for prediction and classification of engineering students' achievement. In proceedings of the *IEEE 7th International Conference on Engineering Education*, ICEED 2015,

49–53. https://doi.org/10.1109/ICEED.2015.7451491

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*, *44*(1), 130–160. https://doi.org/10.3102/0091732X20903304

Ismail, L., Materwala H., & Hennebelle A. (2021) Comparative Analysis of Machine Learning Models for Students' Performance Prediction. In: Antipova T. (eds) Advances in Digital Science. ICADS 2021. *Advances in Intelligent Systems and Computing, 1352*. Springer, Cham. https://doi.org/10.1007/978-3-030-71782-7_14

Koedinger, K.R., D'Mello, S., McLaughlin, E.A., Pardos, Z.A., & Rosé, C.P. (2015), Data mining and education. *WIREs Cogn Sci*, *6*, 333-353. https://doi.org/10.1002/wcs.1350

Makhtar, M., Nawang, H., Nor, S., & Shamsuddin, W. A. N. (2017). Analysis On Students Performance Using Naïve Bayes Classifier. *Journal of Theoretical and Applied Information Technology*, *95*(16), 3993-4000.

Marbouti, F., Diefes-dux, H. A., & Madhavan, K. (2016). Computers & Education Models for Early Prediction of At-risk Students. In: A Course Using Standards-based Grading. *Computers & Education*, *103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N. J. A. H., Ramírez-Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2019.2962680

Saniron, S., & Ali Othman, Z. (2019). Model Penempatan Guru Berasaskan Perlombongan Data. *Journal of ICT in Education*, *3*, 13-23. Retrieved from https://ejournal.upsi.edu.my/index.php/JICTIE/article/view/2605.

Uswatun Khasanah, A., & Harwati, (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *Materials Science and Engineering Conference Series*, *215*(1), 12036. doi:10.1088/1757-899X/215/1/012036.

Zainudin, S., & Ab Aziz, M. A. (2015). *Perlombongan data bagi meramal prestasi akademik pelajar*. *Inovasi pengajaran dan pembelajaran dalam teknologi maklumat*. Pusat Pengajaran dan Teknologi Pembelajaran, UKM.