



ANALISIS RAMALAN PRESTASI AKADEMIK PELAJAR MENGUNAKAN APLIKASI MyCGPA

PREDICTION ANALYSIS OF STUDENT ACADEMIC PERFORMANCE USING MyCGPA APPLICATION

Muhammadd Yazid Abdul Mutalib¹, Zuraini Zainol^{2*}, Ummul Fahri Abdul Rauf³, Puteri NE Nohuddin⁴

¹ Department of Computer Science, Universiti Pertahanan Nasional Malaysia, Malaysia

Email: muhammadyazid250@gmail.com

² Department of Computer Science, Universiti Pertahanan Nasional Malaysia, Malaysia

Email: zuraini@upnm.edu.my

³ Pusat Asasi Pertahanan, Universiti Pertahanan Nasional Malaysia, Malaysia

Email: ummul@upnm.edu.my

⁴ Faculty of Business, Higher College of Technology, Sharjah, UAE

Email: pnohuddin@hct.ac.ae

* Corresponding Author

Article Info:

Article history:

Received date: 15.03.2023

Revised date: 10.04.2023

Accepted date: 23.05.2023

Published date: 01.06.2023

To cite this document:

Mutalib, M. Y. A., Zainol, Z., Rauf, U. F. A., & Nohuddin, P. N. (2023). Analisis Ramalan Prestasi Akademik Pelajar Menggunakan Aplikasi MyCGPA. *Journal of Information System and Technology Management*, 8 (31), 80-94.

DOI: 10.35631/JISTM.831006

This work is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)



Abstrak:

Cabaran utama yang dihadapi oleh ahli akademik ialah mengenalpasti pelajar yang berprestasi lemah pada peringkat awal sebelum mereka bergraduasi. Disebabkan masalah ini berlaku pada setiap semester, kajian ini mencadangkan penyelesaian dengan membangunkan satu aplikasi web yang menggunakan model ramalan bagi membantu penasihat akademik atau pihak pengurusan fakulti untuk mengenal pasti pelajar-pelajar yang diramalkan berprestasi lemah dalam bidang akademik pada peringkat lebih awal. Set data yang digunakan di dalam kajian ini diperolehi daripada (i) profil demografi pelajar tahun 3 sarjana muda sains komputer dan (ii) keputusan peperiksaan daripada Bahagian Pengurusan Akademik (BPA), UPNM. Algoritma *Multiple Linear Regression* (MLR) digunakan untuk membangunkan model ramalan PNGK pelajar berdasarkan tiga atribut utama: 'Activity', 'Absence' dan GPA. Kajian ini menggunakan metodologi CRISP-DM yang terdiri daripada enam fasa utama: pemahaman kajian, pemahaman data, penyediaan data, pemodelan, penilaian dan penyerahan. Hasil kajian ini boleh dimanfaatkan oleh pihak universiti untuk mengenalpasti pelajar-pelajar dari awal semester yang diramalkan berprestasi akademik lemah supaya pelbagai aktiviti pembelajaran dapat dilakukan untuk melonjakkan prestasi akademik pelajar tersebut.

Kata Kunci:

Model Ramalan, Prestasi Akademik Pelajar, *Multiple Linear Regression*, Perlombongan Data

Abstract:

The major challenge faced by academicians is identifying low-performing students at an earlier stage before they graduate. Since this problem occurs every semester, this study proposes a solution by developing a web application that applies a predictive model to assist academic advisors or faculty management in identifying students who are predicted to perform poorly in academics at an earlier stage. The data set used in this study was obtained from (i) the demographic profile of 3rd year undergraduate computer science students and (ii) the examination results from the Academic Management Division (BPA), UPNM. The Multiple Linear Regression (MLR) algorithm is used to develop this academic performance prediction model based on selected 'Activity', 'Absence' and GPA attributes. Meanwhile, the CRISP-DM methodology is used to solve the development problem of this application consisting of six main phases: research understanding, data understanding, data preparation, modelling, evaluation and deployment. The findings of this study can be used by the university to identify students who are expected to have poor academic performance at the beginning of the semester so that various learning activities can be carried out to enhance the student's performance.

Keywords:

Prediction Model, Student Academic Performance, Multiple Linear Regression, Data Mining

Pengenalan

Kualiti pengajaran dan pembelajaran di anggap sebagai salah satu pembangunan yang sangat penting di dalam bidang pendidikan. Kualiti tersebut dapat diukur dan dinilai berdasarkan tahap prestasi akademik pelajar sesebuah universiti itu. Oleh itu, institusi pendidikan perlu mewujudkan satu strategi yang cekap bagi membantu dalam meningkatkan prestasi akademik pelajar mereka. Strategi tersebut boleh dirancang dan diaplikasikan apabila pihak yang berkaitan seperti fakulti, jabatan ataupun pensyarah dapat membuat ramalan prestasi akademik pelajar dengan lebih awal agar pelbagai program dapat dirancang dengan baik. Antara contoh program yang boleh membantu meningkatkan kecemerlangan akademik pelajar ialah sesi kaunseling, aktiviti ilmiah, penglibatan dalam program-program pembangunan kecemerlangan dan sebagainya (Radwan et al., 2020).

Salah satu cara untuk meramal prestasi akademik pelajar adalah dengan menggunakan teknik perlombongan data (PD) ke atas set data akademik pelajar yang dikumpulkan di institusi pengajian tersebut. Bidang PD ini akan mengaplikasikan algoritma pembelajaran mesin ke atas set data yang diekstrak bagi mendapatkan corak, trend atau penjanaan ramalan ke atas data baru. Menurut Alyahyan dan Düşteğör (2020), terdapat pelbagai jenis faktor (atribut) yang digunakan untuk meramal prestasi pelajar. Antaranya ialah Purata Nilai Gred Semester (PNGS atau GPA), Purata Nilai Gred Keseluruhan (PNGK atau CGPA), maklumat demografi pelajar seperti status keluarga, pekerjaan ibu dan bapa serta banyak lagi. Salah satu teknik yang biasa digunakan oleh penyelidik dalam meramal prestasi pelajar adalah regresi. Regresi merupakan

teknik pembelajaran mesin yang menggunakan konsep yang sama seperti teknik pengelasan dalam menggunakan data yang dilatih untuk membuat ramalan. Terdapat beberapa perbezaan di antara kedua-dua teknik ini. Antaranya adalah output pemboleh ubah bagi kaedah regresi adalah dalam bentuk berangka (El Aissaoui et al., 2020). Perbezaan lain ialah fungsi pengelasan meramal label kelas data yang berbentuk kategori manakala fungsi peramalan pula meramal terus fungsi nilai selanjar (Hamdan et al., 2018). Dalam kajian ini, algoritma *Multiple Linear Regression* (MLR) akan digunakan untuk meramal prestasi akademik pelajar di Jabatan Sains Komputer, UPNM.

Latar Belakang Kajian

Perlombongan Data (PD) merupakan proses meneroka pengetahuan yang tersembunyi dalam sejumlah besar data melalui algoritma atau teknik tertentu (Han et al., 2011). PD banyak digunakan dalam pelbagai bidang penyelidikan seperti pendidikan, kesihatan dan perubatan, keselamatan siber, pengesanan penipuan, pencegahan jenayah dan banyak lagi. Menurut Dunham (2006), PD menggunakan pelbagai algoritma untuk melengkapkan pelbagai tugas. Penerangan ringkas berkenaan algoritma-algoritma ini telah dibincangkan dalam kajian Nohuddin et al. (2018a). Kesemua algoritma-algoritma ini mampu untuk membina sebuah model daripada data. Jadual 1 menunjukkan PD terbahagi kepada dua model utama iaitu prediktif dan deskriptif. Model prediktif akan membuat ramalan berkenaan nilai data menggunakan keputusan yang diketahui daripada pelbagai data. Permodelan prediktif boleh dibuat berdasarkan penggunaan set data yang lain. Manakala model deskriptif mengenal pasti corak atau hubungan dalam data. Model deskriptif berfungsi sebagai langkah untuk meneroka sifat data yang diperiksa, bukan untuk meramalkan sifat baharu. Empat teknik terkenal bagi tugas prediktif termasuklah pengelasan, regresi, analisis siri masa dan ramalan. Manakala tugas bagi deskriptif ialah pengelompokan, petua sekutuan, *summarization* dan penemuan urutan.

Jadual 1: Model dan *Tasks* dalam Perlombongan Data (Dunham, 2006)

	Model	Tugasan (<i>Task</i>)
Perlombongan Data	Prediktif	Pengelasan Regresi Analisis Siri Masa Ramalan
	Deskriptif	Pengelompokan <i>Summarization</i> Petua Sekutuan Penemuan Urutan

Multiple Linear Regression (MLR) merupakan model pengelasan dan digunakan secara meluas dalam analisis peramalan. MLR mengkaji hubungan di antara satu pemboleh ubah bersandar berterusan dan dua atau lebih pemboleh ubah tidak bersandar. Berikut ialah persamaan untuk MLR:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Di mana:

Y ialah pemboleh ubah bersandar.

X_i sehingga X_p merupakan pemboleh ubah bersandar, juga dikenali sebagai *regressor* dan peramal.

β_0 ialah nilai pintasan Y apabila semua pemboleh ubah tidak bersandar (X_i sehingga X_p) adalah sama dengan sifar.

B_1 sehingga B_p merupakan anggaran pekali regresi

Perlombongan Data Pendidikan (PDP) merupakan menggunakan pendekatan perlombongan data untuk menyelesaikan masalah berkaitan data pendidikan (Yağcı, 2022). PDP menggunakan pelbagai teknik seperti pembelajaran mesin, statistik, PD, psikologi pendidikan, psikologi kognitif, dan teori serta kaedah lain untuk menganalisis data pendidikan (Romero & Ventura, 2010). Antara teknik PD yang popular ialah pengelasan (Alsaman et al., 2019; Alsariera et al., 2022; Jauhari & Supianto, 2019; Rahayu et al., 2018), pengelompokan (Abdo et al., 2021; Fida et al., 2022; Iatrellis et al., 2021; Nohuddin et al., 2018b) dan perlombongan perhubungan (Nohuddin et al., 2021). Antara contoh kajian yang dijalankan dalam bidang PDP ialah meramal prestasi akademik pelajar, mengenal pasti pelajar yang berada dalam kedudukan bahaya, gagal atau tercicir di peringkat awal dan menyediakan sokongan yang diperlukan untuk mereka yang berisiko.

Pembangunan Aplikasi Ramalan Prestasi Akademik Pelajar menggunakan Teknik Perlombongan Data dalam Bidang Pendidikan

PD banyak digunakan dalam bidang pendidikan khususnya dalam penerokaan ramalan prestasi akademik pelajar. Pemodelan ramalan menggunakan teknik PD dan algoritma pembelajaran mesin untuk mengenal pasti corak dalam data dan pengesanan awal prestasi akademik pelajar. Sehingga kini, banyak kajian penyelidikan lebih tertumpu kepada pembinaan model ramalan prestasi akademik pelajar (Alsariera et al., 2022; Baig et al., 2023; Pallathadka et al., 2023; Qiu et al., 2022) dan tidak banyak kajian ramalan prestasi akademik yang memfokuskan kepada pembangunan aplikasi atau sistem. Terdapat beberapa kajian yang dijalankan untuk membangunkan aplikasi peramal prestasi akademik pelajar khususnya untuk membantu dalam pengesanan awal pelajar yang berprestasi lemah dan tercicir daripada pembelajaran.

Sebagai contoh penyelidik Alboaneen et al. (2022) telah membangunkan satu aplikasi berasaskan web yang dapat meramal prestasi akademik pelajar dan mengenal pasti risiko kegagalan berdasarkan faktor akademik dan demografi. Kajian ini menggunakan pelbagai algoritma pembelajaran mesin seperti *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), *Random Forest* (RF), *K-Nearest Neighbors* (KNN) and *Linear Regression* (LR). Kajian ini menggunakan set data pelajar perempuan di Jabatan Sains Komputer, Imam Abdulrahman bin Faisal Universiti (IAU). Set data ini mengandungi maklumat akademik dan demografi pelajar yang terdiri daripada 168 orang pelajar dan 842 *instance* atribut secara keseluruhannya. Berdasarkan kajian mereka, pemodelan ramalan RF telah dipilih kerana ia memberikan nilai *Mean Absolute Percentage Error* (MAPE) yang terendah iaitu sebanyak 6.34% berbanding dengan algoritma lain. Untuk pembangunan aplikasi web, penyelidik telah mengintegrasikan model ramalan RF untuk meramal prestasi akademik pelajar menggunakan rangka kerja web mikro Flask yang berasaskan bahasa pengaturcaraan Python. Dalam aplikasi ini, pengguna perlu memasukkan 10 input iaitu ID pelajar, nama kursus, level, *mid-term score*, *mid activities score*, *lab score*, *city*, *marital status*, *children* dan *health issues*. Selepas memasukan input yang sah, aplikasi ini akan meramal skor markah dan gred bagi subjek yang dipilih.

Dalam kajian lain, Sökkhey dan Okazaki (2020) juga telah membangunkan sistem bernama APPS yang dapat membuat ramalan awal prestasi akademik pelajar. Kajian ini menggunakan empat algoritma pembelajaran mesin iaitu KNN, *Hybrid C5.0*, *Hybrid RF* dan *Improved Deep Belief Network* (IDBN). Kajian ini menggunakan set data pelajar di Cambodia. Keputusan eksperimen menunjukkan Hybrid RF memberikan peratus ketepatan ramalan sebanyak 99.98% dan RMSE sebanyak 0.008%. APPS merupakan sistem yang dibangunkan menggunakan *Shiny*. Sistem ini dibina menggunakan bahasa pengaturcaraan R, HTML, CSS dan *JavaScript*.

Moharekar and Pol (2019) membangunkan aplikasi mudah alih iaitu *Academic Performance Prediction Application* (APPA) yang dapat meramal prestasi akademik pelajar. Empat algoritma pembelajaran mesin telah digunakan dalam kajian ini iaitu DT, KNN, RF dan *XG Boost Multi-Classification*. Hasil eksperimen menunjukkan pemodelan *XG Boost Multi-Classification* telah memberikan peratus ketepatan yang paling tinggi iaitu sebanyak 86% berbanding dengan algoritma yang lain. Pembangunan aplikasi ini menggunakan Django Python REST Framework sebagai *front-end* and bahasa pengaturcaraan Python sebagai *back-end*.

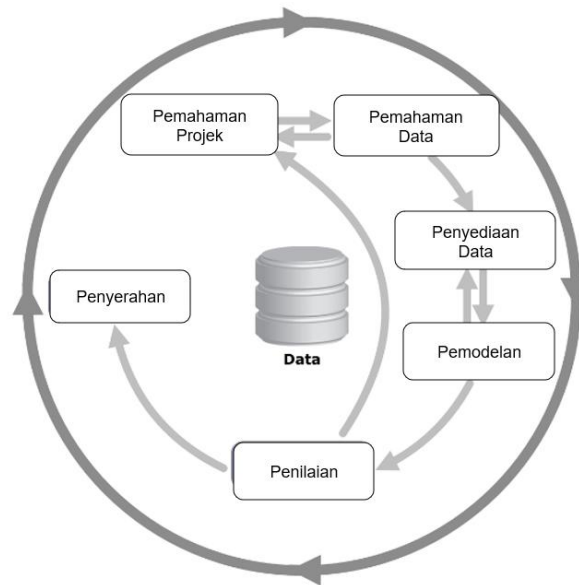
Berdasarkan pemerhatian di atas, didapati hanya sebilangan kecil sahaja kajian yang memfokuskan kepada pembangunan sistem/aplikasi peramalan prestasi akademik pelajar. Sebelum membangunkan aplikasi ini, satu model akan dibina untuk meramal prestasi akademik pelajar menggunakan algoritma MLR. Kajian ini menggunakan set data yang diekstrak daripada *Google Form* di mana semua pelajar tahun 3 jurusan Sains Komputer sidang akademik 2022/2023 perlu mengisi selain daripada data akademik pelajar yang diperolehi daripada Bahagian Pengurusan Akademik (BPA), UPM. Berdasarkan model ramalan, satu aplikasi yang berasaskan web, MyCGPA akan dibangunkan untuk meramal skor CGPA seseorang pelajar dengan lebih cepat dan efisien. Aplikasi ini juga menyediakan pelbagai kemudahan kepada pengguna khususnya pihak fakulti atau penasihat akademik untuk menganalisis data pelajar dengan lebih cekap dan menghasilkan keputusan ramalan skor CGPA pelajar dalam bentuk yang mudah difahami oleh pengguna

Metodologi Penyelidikan

Kajian ini dilaksanakan mengikut metodologi Cross Industry Standard Process for Data Mining (CRISP-DM). Menurut Shafique & Qaiser (2014), CRISP-DM merupakan model yang paling banyak digunakan oleh penyelidik perlombongan data kerana ia menawarkan rangka kerja dan garis panduan lengkap untuk menyelesaikan masalah sedia ada. CRISP-DM menyediakan 6 fasa utama: pemahaman kajian, pemahaman data, penyediaan data, pemodelan, penilaian dan penyerahan (rujuk Rajah 1).

Fasa 1: Fasa Pemahaman Kajian

Fasa 1 adalah memahami objektif dan merancang keperluan kajian. Objektif utama kajian ini adalah membangunkan satu aplikasi web yang dapat mengaplikasikan model ramalan MLR bagi membantu penasihat akademik untuk meramal keputusan akademik pelajar. Pelajar yang berprestasi lemah dalam bidang akademik akan dikenal pasti supaya bantuan awal dapat diberikan bagi mengelakkan keciciran pelajar.



Rajah 1: Aliran Kerja Metodologi CRISP-DM (Wirth & Hipp, 2000)

Fasa 2: Fasa Memahami Data

Fasa ini tertumpu kepada memahami set data yang dikumpulkan. Kajian ini adalah berdasarkan kepada 2 set data utama iaitu set data soalan kaji selidik dan set data akademik pelajar. Penyediaan set soalan kaji selidik ini adalah berdasarkan beberapa kajian terdahulu seperti El Aissaoui et al. (2020), Alyahyan and Düşteğör et al. (2020), Hamoud et al (2018) dan Fadilah et al. (2021). Berdasarkan kajian mereka, terdapat pelbagai faktor (atribut) digunakan untuk meramal prestasi akademik pelajar seperti GPA (PNGS), CGPA(PNGK), data demografi seperti status keluarga, pekerjaan ibu dan bapa serta banyak lagi. Jadual 2 menunjukkan senarai atribut yang digunakan dalam kajian ini. Seramai 97 responden yang terdiri daripada pelajar semester 5 jurusan Sains Komputer FSTP, sidang 2022/2023 berjaya dikumpulkan. Walaupun atribut GPA dan CGPA dikumpulkan melalui kaji selidik, penyelidik perlu membuat semakan data yang diinput oleh pelajar. Set data akademik pelajar ini diperolehi daripada Bahagian Pengurusan Akademik (BPA), UPNM. Ini adalah untuk memastikan data yang diberikan oleh pelajar adalah tepat. Set data mentah yang dikumpulkan ini akan disimpan dalam bentuk fail teks atau dalam format fail Ms Excel iaitu csv. Set data yang dikumpulkan ini mengandungi 21 atribut dan 97 pelajar. Rajah 2 menunjukkan contoh atribut set data yang akan digunakan dalam fasa seterusnya

	Matric	Gender	Race	TypeStudent	NumSiblings	ParentsAlive	ParentsStatus	FQualification	MQualification
0	2200979	Female	Malay	Civilian	between 2-4	Yes	Married	Graduate	Secondary
1	2200249	Male	Malay	Cadet	between 2-4	Yes	Married	Elementary	Graduate
2	2200226	Male	Malay	Civilian	> 5	Yes	Married	Secondary	Secondary
3	2200252	Male	Malay	Cadet	> 5	Yes	Married	Secondary	Secondary
4	2200255	Male	Malay	Cadet	between 2-4	Yes	Married	Graduate	Secondary

Rajah 2: Contoh Set Data Mentah

Fasa 3: Fasa Penyediaan Data Kajian

Set data mentah yang dikumpulkan di Fasa 2 akan melalui fasa prapemprosesan data. Prapemprosesan merupakan tugas yang penting kerana ia menghasilkan data berkualiti tinggi (Alwarthan et al., 2022; Zainol et al., 2018) dan analisis bebas ralat (Navlani et al., 2021). Selain itu, ia juga untuk memastikan set data yang dipilih adalah bersih dan sedia digunakan di fasa permodelan. Rajah 6 menunjukkan tiga langkah utama dalam prapemprosesan data iaitu penapisan data, *label encoding* dan pemilihan fitur telah digunakan pada set data.

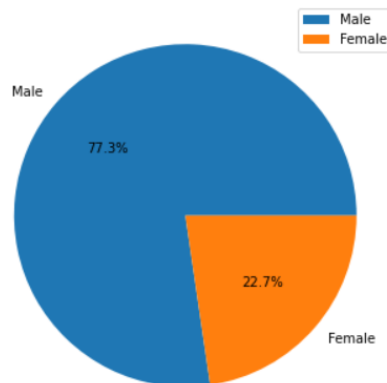
**Rajah 6: Proses Prapemprosesan Data**

Daripada jumlah 21 atribut, hanya 20 atribut yang dipilih (selepas mengalih keluar atribut *Matric*). Jadual 2 menunjukkan senarai atribut dengan penerangannya.

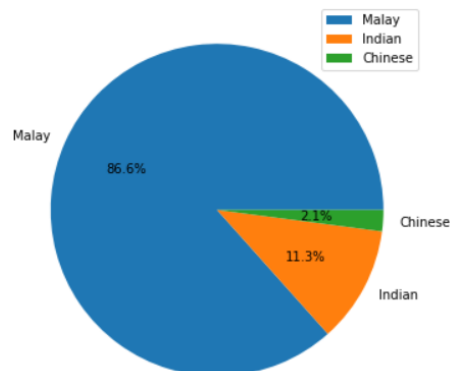
Jadual 2: Senarai Atribut Kajian

Bil	Atribut	Perihal Data
1	<i>Gender</i>	<i>Male, Female</i>
2	<i>Race</i>	<i>Malay, Chinese, Indian</i>
3	<i>TypeStudent</i>	<i>Cadet, Civilian, PALAPES</i>
4	<i>NumSiblings</i>	<i>1, between 2-5, more than 5</i>
5	<i>ParentsAlive</i>	<i>Yes, No</i>
6	<i>ParentsStatus</i>	<i>Married, Divorce, Widowed</i>
7	<i>FQualification</i>	<i>Graduate, Secondary, Elementary, No education</i>
8	<i>MQualification</i>	<i>Graduate, Secondary education, Elementary, No education</i>
9	<i>FJobStatus</i>	<i>Currently on service, Housewife, Retired, None</i>
10	<i>MJobStatus</i>	<i>Currently on service, Retired, None</i>
11	<i>FamilyStatus</i>	<i>B40, M40, T20</i>
12	<i>Scholarship</i>	<i>Yes, No</i>
13	<i>SAccommodation</i>	<i>In campus, rent a room, stay with family</i>
14	<i>Activity</i>	<i>Sports, Student Representation Council (MPP), clubs</i>
15	<i>AssignmentTime</i>	<i>Less than 1 hour, between 2-4 hours, more than 4 hours</i>
16	<i>SocmedTime</i>	<i>Less than 1 hour, between 2-4 hours, more than 4 hours</i>
17	<i>Absence</i>	<i>Less than 3 days, between 4-7 day, more than 7 days</i>
18	<i>Part time</i>	<i>Yes, No</i>
19	<i>GPA (PNGS)</i>	<i>Grade Point Average (from 0 – 4.00)</i>
20	<i>CGPA(PNGK)</i>	<i>Cumulative Grade Point Average (from 0 – 4.00)</i>

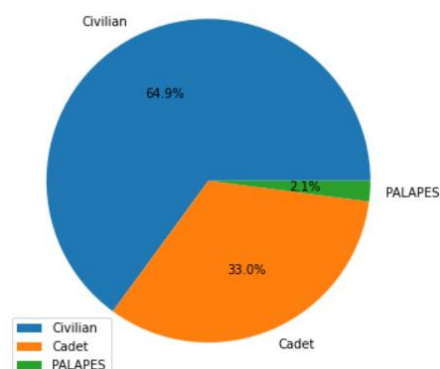
Rajah 3 sehingga 5 menunjukkan analisis deskriptif dalam bentuk carta pai. Responden pelajar lelaki mewakili 78.5% (n=75) daripada keseluruhan responden manakala responden pelajar perempuan adalah seramai 21.5% (n=22). Data juga menunjukkan sebahagian besar responden 84.1% (n=88) berbangsa Melayu, diikuti dengan kaum India, 13.1% (n=7) dan Cina 2.8 (n=2). Seterusnya, 64.9% (n=62) responden adalah pelajar awam, diikuti 33.0% (n=33) responden bagi kadet dan 2.1% (n=2) adalah PALAPES.



Rajah 3: Taburan Data Demografi Responden – Jantina



Rajah 4: Taburan Data Demografi Responden - Kaum



Rajah 5: Taburan Data demografi Responden – Jenis Pelajar

Seterusnya, teknik *label encoder* (pengekodan label) digunakan untuk menukarkan atribut jenis kategori kepada berangka. Dalam pembelajaran berselia, algoritma pembelajaran mesin menunjukkan prestasi yang lebih baik apabila menggunakan data berangka. Oleh itu, 18 atribut jenis kategori seperti *Gender*, *Race*, *TypeStudent*, *NumSiblings*, *ParentAlive*, *ParentStatus*, *FQualification*, *MQualification*, *FJobStatus*, *MJobStatus*, *FamilyStatus*, *Scholarship*, *SAccommodation*, *Activity*, *AssignmentTime*, *SocmedTime*, *Absence* dan *Part time* perlu diwakilkan dalam bentuk berangka sebelum digunakan dalam pembangunan model. Pengekodan label merupakan salah satu teknik prapemprosesan yang digunakan untuk mengendalikan data kategori. Dalam kajian ini, semua data kategori telah ditukar kepada data berangka menggunakan kod arahan Python. Sebagai contoh, atribut *Gender* mengandungi dua kategori utama iaitu *Male* dan *Female*. Kedua-dua atribut ini telah dikodkan dengan nilai integer '0' untuk *Female* dan '1' untuk *Male*. Sampel set data selepas menggunakan pengekodan label ditunjukkan di Rajah 7.

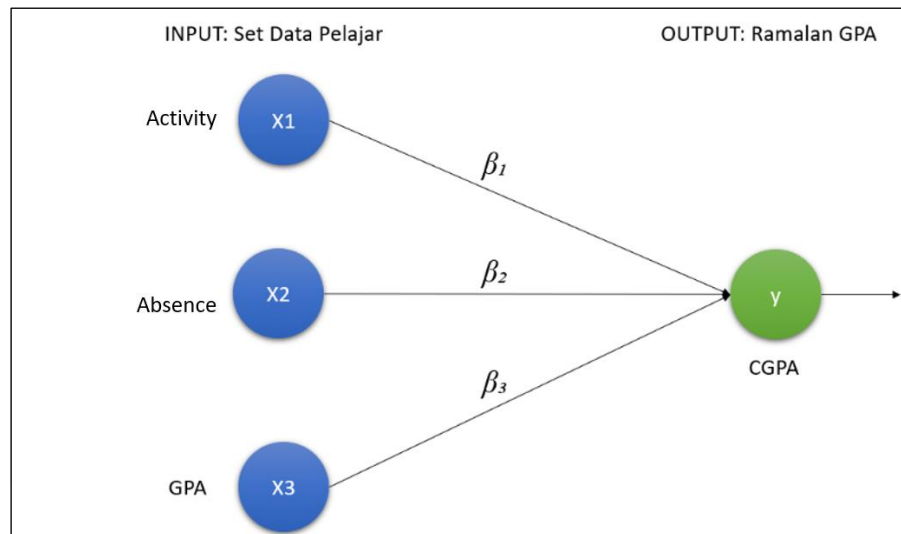
	Gender	Race	TypeStudent	NumSiblings	ParentsAlive	ParentsStatus	FQualification	MQualification
0	0	2	1	2	1	1	1	3
1	1	2	0	2	1	1	0	1
2	1	2	1	1	1	1	2	3
3	1	2	0	1	1	1	2	3
4	1	2	0	2	1	1	1	3

Rajah 7: Contoh Set Data Yang Diwakilkan Proses Pengekodan Label

Langkah seterusnya ialah memilih satu set atribut (pemboleh ubah) tidak bersandar yang paling relevan dengan pemboleh ubah bersandar (sasaran). Set data yang besar dengan atribut yang banyak boleh menyebabkan model mengambil masa yang lama untuk dilaksanakan. Untuk mengatasi masalah ini, teknik *Ordinary Least Squared Regression (OLS)* digunakan untuk mendapatkan atribut yang terbaik bagi membantu meramal prestasi pelajar tersebut. OLS merupakan kaedah statistik yang digunakan untuk membuat anggaran persamaan linear regresi di mana ia menerangkan tentang hubungan di antara satu atau lebih pemboleh ubah tidak bersandar. Hasil keputusan pengiraan ujian OLS menunjukkan *R-squared* bagi pemboleh ubah *Activity*, *Absence* dan GPA mempunyai nilai 0.999 (menghampiri nilai 1). Ini menunjukkan keputusan hubungan yang paling tinggi. Dalam kajian ini, pemboleh ubah *Activity*, *Absence* dan GPA telah dipilih sebagai pemboleh ubah tidak bersandar manakala CGPA dipilih sebagai pemboleh ubah bersandar. Kesemua pemboleh ubah ini telah sedia untuk digunakan dalam fasa pembangunan model.

Fasa 4: Fasa Pemodelan

Kajian ini menggunakan algoritma MLR untuk membina model ramalan. Seperti yang dinyatakan dalam Fasa 3, set data yang telah dipilih akan dibahagikan kepada 2 bahagian utama: set latihan dan set ujian mengikut nisbah 60:40. Pada peringkat ini, algoritma MLR akan melatih set data latihan untuk menghasilkan model ramalan. Model MLR yang terlatih ini akan menjadi pengelas dan data ujian akan digunakan untuk menilai model tersebut. Model MLR dapat digambarkan seperti di Rajah 8. Input model MLR ialah *Activity*, *Absence* dan GPA telah diberikan kepada pemboleh ubah 'X'. Manakala pemboleh ubah 'CGPA' diberikan kepada pemboleh ubah 'Y'.



Rajah 8: Model MLR

Jadual 3: Rumus Model MLR

	Pemboleh ubah tidak bersandar	Rumus model MLR yang dibangunkan dalam kajian ini
Model	<i>a. Activity</i>	- 0.3283 - 0.0034 * Activity - 0.0595 *
MLR	<i>b. Absence</i>	<i>Absence</i> + 1.0950 * GPA
	<i>c. GPA</i>	

Berdasarkan eksperimen, formula untuk model MLR ditunjukkan dalam Jadual 3. MLR menghasilkan pintasan sebagai - 0.3283 dan pekali regresi ialah - 0.0034 (*Activity*), -0.0595 (*Absence*) dan 1.0950 (*GPA*). Berdasarkan jadual di atas, didapati kesemua pemboleh ubah mempunyai korelasi negatif kecuali *GPA*.

Fasa 5: Fasa Evaluasi

Fasa evaluasi merupakan satu fasa yang penting untuk menilai model yang telah dibina. Dalam fasa ini, penulis akan merujuk kembali kepada ke fasa penyediaan data dan fasa permodelan untuk menilai sama ada terdapat sebarang penambahbaikan atau sebarang butiran yang perlu ditambah pada kedua-dua fasa. Tujuan fasa ini adalah untuk memastikan data dan model yang dibina disediakan dengan baik untuk menghasilkan keputusan yang baik. Dengan pemilihan pemboleh ubah yang tepat, model ramalan prestasi akademik pelajar yang cekap dapat dibina. Bagi memastikan model ramalan berfungsi dengan cekap, ujian ketepatan ramalan telah dijalankan. Rajah 9 menunjukkan peratus ketepatan ramalan bagi model MLR sebanyak 85.12%.

```
In [85]: Accuracy=r2_score(y_test,y_predict)*100
print(" Accuracy of the model is %.2f" %Accuracy)

Accuracy of the model is 85.12
```

Rajah 9: Keputusan Ujian Ketepatan Model Ramalan

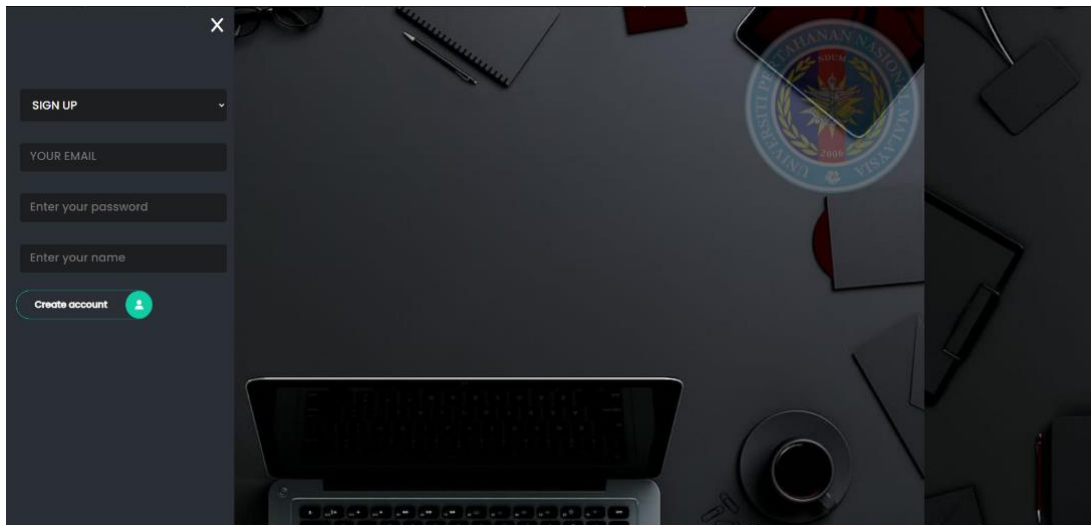
Fasa 6: Fasa Penyerahan

Fasa ini menghasilkan laporan dan perlu dipersembahkan kepada pengguna akhir. Model MLR yang dibina di Fasa 5 akan digunakan dalam pembangunan aplikasi MyCGPA. Aplikasi ini direka bentuk untuk meramal skor CGPA pelajar berdasarkan input tiga pemboleh ubah utama iaitu 'Activity', 'Absence' dan GPA.

Aplikasi Ramalan Prestasi Akademik Pelajar (MyCGPA)

Model ramalan MLR yang dibangunkan di Fasa 3 akan diintergrasikan bersama dengan aplikasi MyCGPA. Aplikasi ini dibangunkan menggunakan *Jupyter Notebook*, *Python Web Framework* dan *Visual Studio Code* sebagai editor kod. *Visual Studio Code* menggunakan rangka kerja web mikro *Flask* sebagai *front-end* dan bahasa pengaturcaraan *Python* sebagai *back-end*. MyCGPA mengandungi 2 peringkat capaian iaitu pentadbir dan pengguna aplikasi. Pentadbir bertanggungjawab menguruskan semua capaian dan kawalan dalam aplikasi semasa proses pendaftaran keahlian.

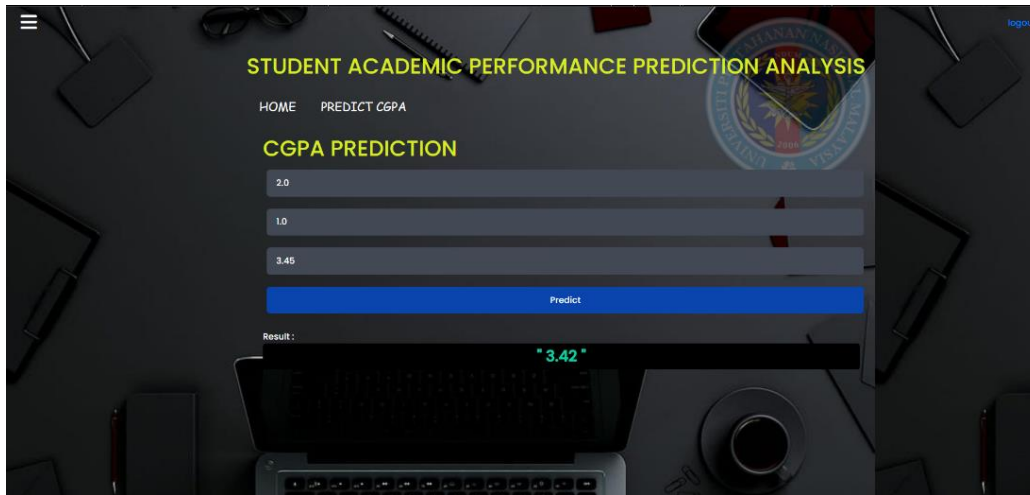
Rajah 10 menunjukkan contoh antara muka untuk halaman log masuk aplikasi MyCGPA. Pengguna yang berdaftar boleh melog masuk ke dalam aplikasi dengan memasukkan e-mel dan kata laluan yang tepat. Jika maklumat yang dimasukkan tidak tepat, pengguna tidak dapat melog masuk ke dalam aplikasi dan satu mesej ralat akan dipaparkan kepada pengguna. Semua kata laluan yang disimpan dalam pangkalan data.



Rajah 10: Paparan Halaman Log Masuk Aplikasi MyCGPA

Selepas log masuk ke aplikasi, pengguna akan dialihkan ke papan pemuka aplikasi. Pengguna perlu membaca terlebih dahulu ke semua arahan dan panduan yang telah tertera sebelum beralih ke halaman peramalan. Seterusnya, pengguna boleh memulakan proses ramalan pada bahagian paparan ramalan. Pengguna perlu memasukkan data yang diperlukan iaitu *Activity*, *Absence* dan GPA. Semua data yang diinput mestilah dalam bentuk berangka. Model ramalan yang dilatih menggunakan algoritma MLR akan meramal keputusan ramalan menggunakan data input. Pengguna perlu menekan butang *Predict* setelah selesai memasukkan ke semua input. Sebagai contoh, aplikasi MyCGPA ini berjaya meramal skor CGPA iaitu 3.42

menggunakan data input seperti *Activity (2.0)*, *Absence (1.0)* dan *GPA (3.45)*. Rajah 11 menunjukkan contoh hasil keputusan ramalan CGPA pelajar bagi seorang pengguna.



Rajah 11: Antara Muka Paparan Keputusan Ramalan Aplikasi MyCGPA

Berdasarkan kepada fasa pengujian dan implementasi yang telah dilakukan, boleh disimpulkan bahawa aplikasi MyCGPA berjaya berfungsi dengan baik dan memenuhi objektif aplikasi. Walau bagaimanapun, penambahbaikan perlu dilakukan dari masa ke masa supaya aplikasi ini menjadi lebih efisien.

Jadual 4 menunjukkan keputusan perbezaan ramalan model berdasarkan nilai sebenar dan nilai ramalan. Sebagai contoh, skor CCPA sebenar ialah 3.12 dan aplikasi ini meramalkan skor CGPA sebanyak 3.158 (perbezaan sebanyak -0.038).

Jadual 4: Perbezaan Nilai Sebenar dan Nilai Ramalan

Nilai Sebenar	Nilai Ramalan	Perbezaan Nilai
3.67	3.660	0.010
3.65	3.589	0.061
3.52	3.383	0.137
3.12	3.158	-0.038
3.33	3.372	-0.042
3.58	3.425	0.155
3.67	3.671	-0.001
3.67	3.685	-0.015
3.55	3.479	0.071
3.87	3.781	0.089
3.28	3.275	0.005
3.02	3.072	-0.052
3.94	3.888	0.052
3.18	3.414	-0.234
3.43	3.425	0.005
3.33	3.457	-0.127
3.72	3.824	-0.104
3.33	3.340	-0.010

3.67	3.653	0.017
3.42	3.393	0.027
3.69	3.696	-0.006
3.32	3.361	-0.041
3.63	3.653	-0.023
3.58	3.457	0.123
3.54	3.393	0.147
3.64	3.802	-0.162
2.92	3.072	-0.152
3.26	3.286	-0.026
3.45	3.415	0.035
3.46	3.521	-0.061
3.33	3.500	-0.170
3.44	3.415	0.025
3.68	3.585	0.095
3.46	3.457	0.003
3.65	3.653	-0.003
3.23	3.286	-0.056
3.42	3.393	0.027
3.76	3.778	-0.018
3.63	3.585	0.045

Kesimpulan

Kajian ini mencadangkan model regresi untuk meramal prestasi pelajar menggunakan algoritma MLR. Set data yang digunakan dalam kajian ini terdiri daripada 97 pelajar dengan 21 atribut daripada pelajar Sarjana Muda Sains Komputer UPNM semester 5 sesi 2022/2023. Dalam kajian ini, aplikasi MyCGPA berjaya meramal skor CGPA menggunakan tiga pemboleh ubah utama iaitu *Activity*, *Absence* dan *GPA*. Penggunaan algoritma MLR berpotensi membantu penasihat akademik dan universiti untuk meramal skor CGPA khususnya kepada pelajar yang berprestasi lemah. Kesimpulannya, aplikasi MyCGPA boleh dimanfaatkan pelbagai pihak di universiti mahupun pelajar itu sendiri dengan mudah. Aplikasi ini juga menyediakan pelbagai kemudahan kepada pengguna khususnya pihak fakulti atau penilai akademik untuk menganalisis data pelajar dengan lebih cekap dan menghasilkan keputusan ramalan nilai CGPA pelajar dalam bentuk yang mudah difahami oleh pengguna.

Penghargaan

Penulis ingin mengucapkan terima kasih kepada Universiti Pertahanan Nasional Malaysia (UPNM) kerana menyokong kerja-kerja penyelidikan ini.

Rujukan

- Abdo, A. M., Rasid, N. M. A., Badli, N. A. H. M., Sulaiman, S. N. A., Wani, S., & Zainol, Z. (2021). Student's Performance Based on E-Learning Platform Behaviour using Clustering Techniques. *International Journal on Perceptive and Cognitive Computing*, 7(1), 72-78.
- Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a web-based prediction system for students' academic performance. *Data*, 7(2), 21.

- Als Salman, Y. S., Halemah, N. K. A., AlNagi, E. S., & Salameh, W. (2019). *Using decision tree and artificial neural network to predict students academic performance*. Paper presented at the 2019 10th international conference on information and communication systems (ICICS).
- Alsariera, Y. A., Baashar, Y., Alkawsy, G., Mustafa, A., Alkahtani, A. A., & Ali, N. a. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*, 2022.
- Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review. *Journal od Applied Computational Intelligence & Soft Computing*, 2022.
- Alyahyan, E., & Düşteğör, 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.
- Baig, M. A., Shaikh, S. A., Khatri, K. K., Shaikh, M. A., Khan, M. Z., & Rauf, M. A. (2023). Prediction of Students Performance Level Using Integrated Approach of ML Algorithms. *International Journal of Emerging Technologies in Learning*, 18(1).
- Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*: Pearson Education India.
- El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., & El Alliou, Y. (2020). A multiple linear regression-based approach to predict student performance. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)* (Vol. 1102, pp. 9-23): Springer.
- Fadilah, K. I. M., Zainol, Z., Ebrahim, M., & Lee, A. S. H. (2021). *Covid-19 Effect On Undergraduate Computing Students' Performance At Higher Education: Pilot Study*. Paper presented at the 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE).
- Fida, S., Masood, N., Tariq, N., & Qayyum, F. (2022). A Novel Hybrid Ensemble Clustering Technique for Student Performance Prediction. *Journal of Universal Computer Science*, 28(8), 777-798.
- Hamdan, A. R., Abu Bakar, A., & Ahamd Nazri, M. Z. (2018). Pengenalan Sains Data. In A. R. Hamdan, A. Abu Bakar, & M. Z. Ahamd Nazri (Eds.), *Sains Data Penerokaan Pengetahuan Dari Data Raya* (pp. 26-51): Penerbit Universiti Kebangsaan Malaysia.
- Hamoud, A., 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, 26-31.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.
- Iatrellis, O., Savvas, I. K., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and information technologies*, 26, 69-88.
- Jauhari, F., & Supianto, A. A. (2019). Building student's performance decision tree classifier using boosting algorithm. *Indones. J. Electr. Eng. Comput. Sci*, 14(3), 1298-1304.
- Moharekar, T. T., & Pol, U. R. (2019). ACADEMIC PERFORMANCE PREDICTION APPLICATION (APPA).
- Navlani, A., Fandango, A., & Idris, I. (2021). *Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python*: Packt Publishing Ltd.

- Nohuddin, P., Zainol, Z., Lee, A. S. H., Nordin, I., & Yusoff, Z. (2018a). A case study in knowledge acquisition for logistic cargo distribution data mining framework. *International Journal of Advanced and Applied Sciences*, 5(1), 8-14.
- Nohuddin, P. N. E., Zainol, Z., & Hijazi, M. H. A. (2021). Study of B40 Schoolchildren Lifestyles and Academic Performance using Association Rule Mining. *Annals of Emerging Technologies in Computing (AETiC)*, 5(5).
- Nohuddin, P. N. E., Zainol, Z., & Nordin, A. (2018b). Monitoring students performance using self organizing map trend clustering. *International Journal of Defence Science, Engineering & Technology*, 1(1), 50-56.
- Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*, 80, 3782-3785.
- Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., . . . Chen, P.-k. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, 12(1), 453.
- Radwan, R. M., Mustapha, A., & Abdullah, B. (2020). Ramalan Prestasi Akhir Pelajar Melalui Kaedah Perlombongan Data [Student Final Performance Forecasting Through Data Mining Methods]. *J Asian Journal of Civilizational Studies*, 2(4), 1-16.
- Rahayu, S. B., Kamarudin, N. D., & Zainol, Z. (2018). Case Study of UPNM Students Performance Classification Algorithms. *Journal of Engineering & Technology*, 7(4.31), 5. doi:10.14419/ijet.v7i4.31.23382
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *J IEEE Transactions on Systems, Man, Cybernetics, Part C*, 40(6), 601-618.
- Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation Scientific Research*, 12(1), 217-222.
- Sokkhey, P., & Okazaki, T. (2020). Developing web-based support systems for predicting poor-performing students using educational data mining techniques. *International Journal of Advanced Computer Science Applications (IJACSA)*, 11(7).
- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. Paper presented at the Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.
- Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 11.
- Zainol, Z., Jaymes, M. T. H., & Nohuddin, P. N. E. (2018). VisualUrText: A Text Analytics Tool for Unstructured Textual Data. *Journal of Physics: Conference Series*, 1018(1), 012011. doi:10.1088/1742-6596/1018/1/012011