KLASIFIKASI PRESTASI PELAJAR B40 DI INSTITUSI PENGAJIAN TINGGI AWAM MALAYSIA BERASASKAN TEKNIK PENGELOMPOKAN

AHMAD FIKRI BIN MOHAMED NAFURI

UNIVERSITI KEBANGSAAN MALAYSIA

KLASIFIKASI PRESTASI PELAJAR B40 DI INSTITUSI PENGAJIAN TINGGI AWAM MALAYSIA MENGGUNAKAN TEKNIK PENGELOMPOKAN

AHMAD FIKRI BIN MOHAMED NAFURI

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT UNTUK MEMPEROLEHI IJAZAH SARJANA SAINS KOMPUTER (KECERDASAN BUATAN)

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2022

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

04 APRIL 2022

AHMAD FIKRI BIN MOHAMED NAFURI
P102926

PENGHARGAAN

Terlebih dahulu saya ingin mengucapkan syukur Alhamdulillah ke hadrat Allah S.W.T kerana dengan dengan limpah kurnia-Nya, maka tesis ini berjaya disiapkan walaupun penuh dengan dugaan dan rintangan.

Di kesempatan ini, saya ingin mengucapkan jutaan terima kasih dan penghargaan kepada Dr. Nor Samsiah binti Sani, selaku penyelia utama, dan Dr. Abdul Hadi bin Abd. Rahman, selaku penyelia bersama, di atas ilmu, khidmat nasihat dan bimbingan yang diberikan kepada saya telah membantu penghasilan tesis ini dengan jayanya. Segala bantuan, strategi dan kebijaksanaan beliau telah mengajar saya untuk menjadi seorang penyelidik dan juga pendidik yang baik, insyaAllah.

Terima kasih juga diucapkan kepada para pensyarah di Fakulti Teknologi dan Sains Maklumat (FTSM), Universiti Kebangsaan Malaysia, Bangi yang telah mencurahkan ilmu secara langsung ataupun tidak langsung sepanjang saya mengikuti pengajian kerja kursus dan penyelidikan.

Sekalung penghargaan saya tujukan kepada pihak Bahagian Biasiswa dan Pembiayaan, Kementerian Pendidikan Malaysia kerana telah menganugerahkan biasiswa Cuti Belajar Bergaji Penuh Dengan Biasiswa (CBBPDB) bagi melanjutkan pengajian ke peringkat sarjana ini.

Buat keluarga tercinta terutama isteri, Siti Hajar binti Yusof, kedua ibu bapa dan mentua, serta anak-anak, Ahmad Izz Shafiy, Adeena Shafiyyah dan Ayra Shafiyyah, terima kasih kerana memahami kesibukan dan mendoakan kejayaan saya sepanjang pengajian ini.

Tidak dilupakan ucapan terima kasih kepada rakan-rakan sepengajian di atas kesudian kalian berkongsi idea dan pendapat dalam penghasilan projek ini. Semoga kita semua berjaya dengan cemerlang dalam pengajian ini.

Saya doakan semoga Allah merahmati kita semua, memberikan nikmat kesihatan yang baik dan dijauhi daripada wabak Covid-19.

ABSTRAK

Ledakan kuantiti data dalam sektor pendidikan telah mendorong pihak Institusi Pengajian Tinggi Awam (IPTA) untuk membuat keputusan dan tindakan berasaskan pengetahuan corak dari data bagi meningkatkan tahap kualiti pendidikan di IPT. Peningkatan kuantiti data juga memberi cabaran kepada golongan penyelidik untuk memastikan prestasi algoritma pembelajaran mesin yang dibangunkan adalah terbaik dan tepat dalam menyelesaikan sesuatu masalah. Banyak kajian yang telah diterbitkan dalam skop peramalan prestasi pelajar, namun begitu, penggunaan teknik pembelajaran mesin tanpa selia untuk mengelas pelajar B40 berdasarkan prestasi dan pencapaian mereka di IPTA pada umumnya masih kurang. B40 merupakan kumpulan isi rumah terendah yang berpendapatan purata di bawah RM4850 sebulan. Tambahan pula, bagi meningkatkan taraf isi rumah B40 ke arah yang lebih baik, pencapaian dan prestasi pelajar B40 di IPTA perlu diberikan perhatian khusus. Sehubungan dengan itu, ia telah mendorong kajian ini untuk dijalankan iaitu pembangunan model pembelajaran mesin tanpa selia menggunakan teknik pengelompokan untuk mengklasifikasi prestasi pelajar B40 di IPTA Malaysia. Objektif utama kajian ini adalah untuk membangunkan model pengelompokan yang terbaik untuk mengelas pelajar berdasarkan prestasi mereka di IPTA berdasarkan set data pelajar daripada Kementerian Pengajian Tinggi (KPT). Pelaksanaan projek ini melibatkan empat fasa iaitu pengumpulan dan penyediaan data; pembangunan model pengelompokan berasaskan teknik k-min, BIRCH dan DBSCAN untuk mengelas prestasi pelajar; pembangunan teknik pengekstrakan ciri untuk mengenal pasti atribut penting dalam mengelas pelajar berdasarkan prestasi mereka di IPT dan yang terakhir adalah analisa serta penilaian model. Kajian ini mendapati model pengelompokan terbaik adalah k-minB yang dibangunkan berasaskan algoritma k-min dengan 10 atribut penting dan telah dinormalisasi menggunakan teknik MinMax. k-minB telah menghasilkan lima kelompok prestasi pelajar B40 iaitu prestasi paling tinggi (kelompok 0), prestasi tinggi (kelompok 1), prestasi sederhana (kelompok 3), prestasi rendah (kelompok 4) dan prestasi paling rendah (kelompok 2). Model klasifikasi prestasi pelajar B40 di IPTA ini boleh menyumbang kepada pengurangan kadar keciciran di kalangan pelajar IPTA kerana berupaya mengenalpasti tahap prestasi pelajar dalam pengajian.

PERFORMANCE CLASSIFICATION OF B40 STUDENTS IN MALAYSIAN HIGHER EDUCATION USING CLUSTERING TECHNIQUE

ABSTRACT

The explosive growth of educational data in recent years has foster data driven actions regarding education quality improvement in Public Higher Education Institutions (HEIs) based on discovering patterns and knowledge. The increase in the quantity of data also poses a challenge to researchers to ensure that the performance of the machine learning algorithm developed is the best and has an ability to solve any problems accurately. Many studies have been published in the scope of student's academic performance prediction. However, application of unsupervised machine learning technique to classify B40 students based on their performance and achievement in Malaysian Public HEI is generally still lacking. B40 represents the lowest group of household with an average income of below than RM4850 per month. Furthermore, in order to improve the standard of living among this group, the achievements and performance of B40 students in HEIs need to be given special attention. In this regard, this has prompted us to propose this study which is a development of unsupervised machine learning models using clustering algorithms for the classification of B40 students performance in Malaysian Public HEIs. Main objective in this study is to develop a clustering model to classify students based on their performance in university using the dataset from Ministry of Higher Education (MOHE). This project development is divided into four phases: data collection and data preparation; k-means, BIRCH and DBSCAN model development to classify the student's performance; use of feature extraction to identify main attributes in classifying students based on their performance and model evaluation. The implementation of this project involves four phases namely data collection and preparation; development of k-min, BIRCH and DBSCAN clustering models to classify student performance; development of feature extraction techniques to identify important attributes in classifying students based on their performance in HEIs; and the last one is model analysis and evaluation. This study found that the best clustering model was k-meansB which has been developed based on a k-means algorithm with 10 attributes and been normalized using MinMax technique. The clustering model have produced five clusters of students based on their performance which are highest performance (cluster 0), high performance (cluster 1), medium performance (cluster 3), low performance (cluster 4) and lowest performance (cluster 2). The classification model developed on B40 student's performance in higher education can contribute in minimizing dropout rate among the students in Malaysian Public HEIs. This is due to the model capabilities in determining the student's academic performance during their studies.

KANDUNGAN

Halaman

Halaman

Halaman

LAMPIRAN

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

ANN	Rangkaian Neural Buatan (*Artificial Neural Network)*
BIRCH	*Balanced Iterative Reducing and Clustering using Hierarchies*
BPPD	Bahagian Perancangan dan Penyelidikan Dasar
CGPA	Purata Nilai Gred Kumulatif (*Cumulative Grade Point Average*)
CH	*Calinski Harabasz*
DB	*Davies-Bouldin*
DBSCAN	*Density-Based Spatial Clustering of Application with Noise*
EDM	Perlombongan Data Pendidikan (*Educational Data Mining*)
HEI	Higher Education Institutions
KPT	Kementerian Pengajian Tinggi
HEI	*Higher Education Institutions*
IG	*Information gain*
IPTA	Institusi Pengajian Tinggi Awam
MOHE	*Ministry of Higher Education*
PCA	Analisis Komponen Principal (*Principal Component Analysis*)
Std	Sisihan Piawai (*Standard Deviation*)
STEM	Sains, teknologi, kejuruteraan dan matematik (*Science, technology, engineering and mathematics*)
UA	Universiti Awam

BAB I

PENDAHULUAN

1.1 PENGENALAN

Setiap tahun, jumlah kuantiti data pelajar dalam sektor pendidikan menunjukkan peningkatan yang tinggi. Data-data ini disimpan dalam kuantiti yang sangat besar dan kaya dengan maklumat-maklumat penting dan agak mustahil untuk dianalisis secara manual. Jadi, peralatan dan pendekatan yang boleh menganalisis data-data ini secara automatik amatlah diperlukan bagi mengeluarkan corak dan pengetahuan yang tersembunyi. Seterusnya, maklumat ini boleh dimanfaatkan oleh para pendidik untuk membantu meningkatkan prestasi pelajar.

Romero dan Ventura (2020) dalam penulisannya menerangkan bahawa terdapat dua komuniti yang menggunakan data pendidikan bagi meraih kepentingan dan mendapatkan faedah iaitu komuniti yang menggunakan perlombongan data pendidikan (*educational data mining*) dan juga analitik pembelajaran (*learning analytics*). Menurut Avella et al. (2016), perlombongan data pendidikan, analitik pembelajaran dan analitik akademik adalah konsep yang hampir sama dan saling berkaitan antara satu-sama lain. Perlombongan data pendidikan adalah satu kaedah untuk mengeluarkan pengetahuan daripada data pendidikan di mana hal ini memberi tumpuan kepada membangunkan teknik-teknik analitik untuk meneroka data. Manakala, analitik, mengikut kamus dewan membawa maksud berkenaan dengan penyelidikan, pendekatan dan sebagainya yang bersifat analisis ataupun kemampuan atau kebolehan untuk menganalisis sesuatu.

Terkini, perlombongan data pendidikan digunakan sebagai satu alat yang sangat bermanfaat dalam menganalisis dan meramal tingkah laku dan prestasi pelajar pada masa hadapan (Ahuja et al., 2019). Selain itu, penggunaan analitik pembelajaran pula telah

berkembang sejak kebelakangan ini disebabkan oleh faktor peningkatan yang besar kuantiti data, format data bertambah baik, kemajuan perkomputeran dan ketersediaan alatan yang canggih bagi membuat analisis (Okewu et al., 2021).

Bab ini membincangkan latar belakang kajian berkaitan bidang klasifikasi prestasi pelajar, permasalahan kajian, persoalan kajian, hipotesis kajian, objektif kajian, skop kajian, metodologi kajian dan organisasi tesis.

1.2 LATAR BELAKANG KAJIAN

Pembelajaran mesin tanpa selia terutama teknik pengelompokan telah digunakan dalam bidang pendidikan bagi membuat analisis tentang prestasi pelajar dan mengelas pelajar mengikut faktor dan ciri tertentu. Dalam proses perlombongan data, peningkatan amaun data yang sangat tinggi akan menyebabkan penghasilan kelompok menjadi lebih kompleks. Hal ini kerana algoritma terpaksa memilih titik data yang serupa untuk diletakkan ke dalam kelompok yang sama berdasarkan struktur data.

Sejak kebelakangan ini, teknik pembelajaran mesin banyak diaplikasikan dalam kajian mengenai peramalan prestasi pelajar di pelbagai peringkat pengajian. Namun, majoriti kajian hanya menggunakan algoritma pengelasan dalam membuat ramalan. Penambahbaikan seperti penggunaan teknik pembelajaran mesin tanpa selia dengan menggunakan algoritma pengelompokan untuk mengelas pelajar berdasarkan prestasi, tingkah laku dan pencapaian dijangka dapat meningkatkan tahap ketepatan model yang dibina.

Garg et al. (2020) dalam kajiannya telah menyenaraikan kelebihan-kelebihan algoritma pembelajaran mesin tanpa selia dalam proses perlombongan data. Algoritma k-min yang memerlukan data jenis numerikal merupakan satu teknik yang ringkas dan mudah untuk dijalankan. Selain itu, algoritma k-min juga merupakan teknik yang popular dan sering digunakan oleh para pengkaji dalam pelbagai bidang terutama pendidikan. Algoritma DBSCAN pula berupaya menapis hingar daripada sampel data dan juga mencari bentuk kelompok yang arbitrari. Jadi, teknik ini boleh digunakan terhadap set data dengan taburan yang tidak diketahui. Seterusnya algoritma BIRCH pula digunakan secara efisien dalam

menganalisis set data yang besar. Algoritma ini juga sesuai digunakan untuk membuat analisis terhadap data yang mengandungi struktur hierarki secara tersembunyi.

Meramal prestasi pelajar dan juga keciciran dalam pelajaran adalah amat penting di peringkat pengajian tinggi terutama di universiti awam (UA). Bergelut dengan situasi ekonomi dunia yang tidak menentu, pentadbir IPT awam mahupun swasta terpaksa mengurangkan kos dan meningkatkan kecekapan institusi masing-masing semenjak dunia belum dilanda wabak Covid-19 lagi. Justeru, mengekalkan bilangan pelajar semaksimum yang mungkin pada setiap semester sehingga ke akhir pengajian menjadi agenda utama bagi menjamin kemapanan sesebuah IPT.

Berdasarkan laporan "Statistik Pendidikan Tinggi Universiti Awam Tahun 2020" yang dikeluarkan oleh KPT, bilangan keluaran pelajar yang berjaya menamatkan pengajian adalah lebih rendah pada tahun 2020 (68,606 orang) berbanding keluaran tahun 2019 (78,485 orang). Selain itu, perbandingan antara jumlah keluaran dengan jumlah kemasukan atau enrolmen juga menunjukkan penurunan yang sangat ketara dan situasi di IPT awam seluruh negara ini disifatkan sebagai amat membimbangkan.

1.3 PENYATAAN MASALAH

Masalah yang perlu diambil perhatian dalam kajian ini adalah kekurangan penggunaan kaedah pembelajaran mesin tanpa selia iaitu teknik pengelompokan terhadap data pendidikan oleh para pengkaji. Dapat dilihat kajian-kajian sebelum ini tertumpu kepada pembangunan model pembelajaran mesin tanpa selia dengan menggunakan set data pelajar daripada perisian pembelajaran atas talian bagi meramal prestasi pelajar (DeFreitas & Bernard, 2015; Hooshyar et al., 2019; Navarro & Moreno-Ger, 2018; Šarić-Grgić et al., 2020). Walaupun teknik pengelompokan telah digunakan untuk meramal prestasi pelajar, namun set data yang digunakan hanya merangkumi set data prestasi pelajar daripada pengajian atas talian. Oleh itu, kajian ini akan menggunakan set data pelajar sebenar yang mengikuti pengajian di IPTA untuk mengelas mereka berdasarkan prestasi.

Beberapa algoritma pengelompokan yang dicadang oleh para pengkaji bagi membina model pengelompokan data pendidikan adalah seperti algorima k-min, BIRCH dan DBSCAN

(Li et al., 2021; Valarmathy & Krishnaveni, 2019). Algoritma-algoritma ini mempunyai kelebihan dan kekurangan tersendiri namun dapat menghasilkan kelompok dengan prestasi yang memuaskan. Antara kekangan yang digariskan oleh pengkaji di atas adalah kesukaran untuk mencari hasil pengelompokan yang baik kerana algoritma ini sensitif terhadap tetapan parameter algoritma pengelompokan. Justeru bagi mengatasi masalah kedua dalam kajian ini, algoritma k-min, BIRCH dan DBSCAN dibangunkan untuk mengenalpasti kelompok prestasi pelajar B40 di IPTA.

Faktor tingkah laku pelajar sangat kurang diambil perhatian dalam meramal keciciran pelajar dalam kalangan isi rumah B40 di Universiti Awam. Dimensi baru iaitu tingkah laku pelajar ini akan melihat kepada faktor bilangan aktiviti kurikulum yang disertai oleh pelajar sewaktu pengajian di IPTA. Teori Fredricks (Fredricks et al., 2004) menjelaskan bahawa penglibatan tingkah laku pelajar mempunyai hubungan yang positif dengan pencapaian akademik dan kadar keciciran. Menurutnya lagi, salah satu takrifan penglibatan tingkah laku adalah keterlibatan pelajar dalam aktiviti kurikulum (Yang et al., 2021). Sehubungan dengan itu, selain dimensi demografi, prestasi dan kerjaya pelajar, kajian ini juga akan melihat kepada dimensi penglibatan tingkah laku dalam mengelompok atau mengelas prestasi pelajar B40 di IPTA.

Justeru, bagi menyelesaikan isu-isu di atas, kajian ini akan membangunkan model pengelompokan pelajar B40 di IPT menggunakan algoritma k-min, BIRCH dan DBSCAN. Set data yang digunakan adalah data pelajar B40 sebenar di IPTA.

1.4 PERSOALAN KAJIAN

Berdasarkan permasalahan kajian yang telah dibincangkan di atas, berikut merupakan beberapa persoalan kajian yang perlu dijawab iaitu:

i. Apakah model pembelajaran mesin tanpa selia terbaik untuk mengelompok pelajar B40 mengikut prestasi mereka di IPTA berdasarkan atribut-atribut yang digunakan?
ii. Apakah atribut-atribut penting bagi mengelas setiap kelompok pelajar B40?
iii. Apakah model pengelasan yang terbaik untuk mengelas pelajar B40 berdasarkan kelas prestasi yang dihasilkan dari model pengelompokan?

1.5 HIPOTESIS KAJIAN

Berdasarkan persoalan kajian yang telah dinyatakan di atas, maka hipotesis kajian adalah seperti berikut:

i. Model pembelajaran mesin tanpa selia terbaik untuk mengelompok pelajar B40 dapat dihasilkan melalui teknik pengelompokan k-min, BIRCH dan DBSCAN.
ii. Teknik pengelompokan yang dibangunkan dapat mengelompok pelajar B40 mengikut prestasi pelajar berdasarkan atribut-atribut yang digunakan menerusi pendekatan pengekstrakan ciri untuk menganalisis data.
iii. Teknik pengelasan yang dibangunkan dapat mengelas pelajar B40 mengikut prestasi pelajar berdasarkan kelas prestasi yang dihasilkan dari model pengelompokan yang terbaik.

1.6 OBJEKTIF KAJIAN

Bagi membuktikan hipotesis penyelidikan yang telah dinyatakan di atas, maka objektif kajian adalah seperti berikut:

i. Membangunkan model pengelompokan k-min, BIRCH dan DBSCAN untuk mengelas prestasi pelajar B40 di IPTA .
ii. Mengenalpasti kelompok prestasi pelajar dan atribut penting dalam setiap kelompok untuk mengelas pelajar B40 berdasarkan prestasi mereka di IPTA.
iii. Membangunkan model pengelasan untuk mengelas pelajar B40 berdasarkan kelompok prestasi yang telah dikenal pasti.

1.7 SKOP KAJIAN

Skop kajian memfokuskan kepada prestasi pelajar dalam kalangan isi rumah B40 di universiti awam, Kementerian Pendidikan Malaysia. Data yang digunakan adalah data pelajar daripada 20 UA di peringkat ijazah sarjana muda bagi sesi tamat pengajian tahun 2015 hingga 2019 sama ada mengalami keciciran atau berjaya menamatkan pengajian. Bilangan data keseluruhan adalah sebanyak 248,568 data pelajar sepenuh masa dan mempunyai 53 atribut. Kajian menumpukan kepada para pelajar dalam golongan B40 kerana kemampuan mereka menamatkan pengajian akan memberi impak secara langsung kepada taraf hidup golongan B40

dan kadar kemiskinan di Malaysia. Metodologi kajian adalah menggunakan teknik pembelajaran mesin tanpa selia iaitu algoritma pengelompokan k-min, BIRCH dan DBSCAN. Selain itu, teknik pembelajaran mesin berselia juga telah digunakan iaitu algoritma pengelasan pohon keputusan, hutan rawak dan ANN. Setelah itu, penilaian prestasi pengelompokan yang diterapkan adalah berdasarkan pengesahsahihan dalaman *Davies Bouldin*, pekali *silhouette* dan *Calinski-Harabasz*. Jabatan Pengajian Tinggi, Kementerian Pendidikan Malaysia menjadi sumber rujukan dan tempat untuk mendapatkan data berkaitan keciciran pelajar.

1.8 METODOLOGI PENYELIDIKAN

Bagi memastikan persoalan kajian dijawab, hipotesis kajian dibuktikan dan objektif kajian tercapai, maka kaedah kuantitatif digunakan dalam menjalankan kajian ini. Set data yang digunakan terdiri daripada atribut-atribut yang diukur dan dinilai dengan kaedah pengukuran matematik dan juga statistik. Algoritma-algoritma pengelompokan yang digunakan juga berasaskan kepada formula matematik dan memerlukan input numerikal untuk beroperasi.

Secara umumnya, pelaksanaan kajian ini melibatkan empat fasa iaitu:

Fasa 1: Pengumpulan dan penyediaan data yang mengandungi prestasi dan tingkah laku pelajar B40 di IPTA.

Fasa 2: Pembangunan model pembelajaran mesin tanpa selia berasaskan algoritma pengelompokan iaitu k-min, BIRCH dan DBSCAN untuk mengelas prestasi pelajar B40 di IPTA.

Fasa 3: Penggunaan teknik pengekstrakan ciri semasa analisis hasil pengelompokan untuk mengenal pasti atribut penting dalam mengelas pelajar B40 berdasarkan prestasi mereka di IPTA.

Fasa 4: Penilaian model.

1.9 ORGANISASI TESIS

Secara keseluruhan, tesis ini merangkumi lima bab dan disusun mengikut struktur yang berikut:

Bab 1: Menerangkan secara ringkas dan padat mengenai kajian yang akan dilaksanakan. Penerangan dalam bab ini meliputi latar belakang, permasalahan, persoalan, hipotesis, objektif, skop dan metodologi kajian.

Bab 2: Membincangkan kajian kepustakaan iaitu sorotan kajian-kajian lalu mengenai prestasi pelajar, kumpulan isi rumah B40 dan teknik-teknik pengelompokan yang akan digunakan dalam kajian ini.

Bab 3: Menerangkan dengan terperinci metodologi kajian yang akan dilaksanakan termasuk kaedah pra-pemprosesan data, analisis deskriptif data serta kaedah pelaksanaan eksperimen.

Bab 4: Membincangkan hasil eksperimen pembangunan model pengelompokan meliputi penetapan nilai *k* terbaik dan keputusan penilaian pengelompokan bagi penentuan model dan algoritma terbaik.

Bab 5: Memperincikan analisis hasil pengelompokan bagi mengenalpasti atribut-atribut yang mempengaruhi setiap kelompok yang dihasilkan. Model pengelasan prestasi pelajar B40 menggunakan algoritma pengelasan juga akan dibangunkan dan dibincangkan di dalam bab ini.

Bab 6: Membuat rumusan keseluruhan tentang hasil kajian termasuk kekangan, sumbangan dan cadangan perluasan kajian pada masa hadapan.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Bab ini akan menerangkan dengan lebih lanjut tentang kajian-kajian terdahulu yang telah dibuat mengenai klasifikasi prestasi pelajar menggunakan teknik pembelajaran mesin berselia dan tanpa selia bagi membincangkan permasalahan yang telah diketengahkan. Tinjauan yang dibuat juga adalah bagi menyiasat keupayaan teknik pengelompokan untuk mengelas pelajar berdasarkan prestasi mereka di IPT. Kajian-kajian yang menyiasat faktor penyumbang penting kepada prestasi pelajar IPT akan dikupas serta dianalisis bagi mencari titik persamaan dengan atribut yang ada di dalam set data kajian. Rujukan melalui kajian kepustakaan ini adalah dari kajian-kajian yang telah dilakukan di luar negara dan juga tempatan.

2.2 PEMBELAJARAN MESIN DAN PERLOMBONGAN DATA

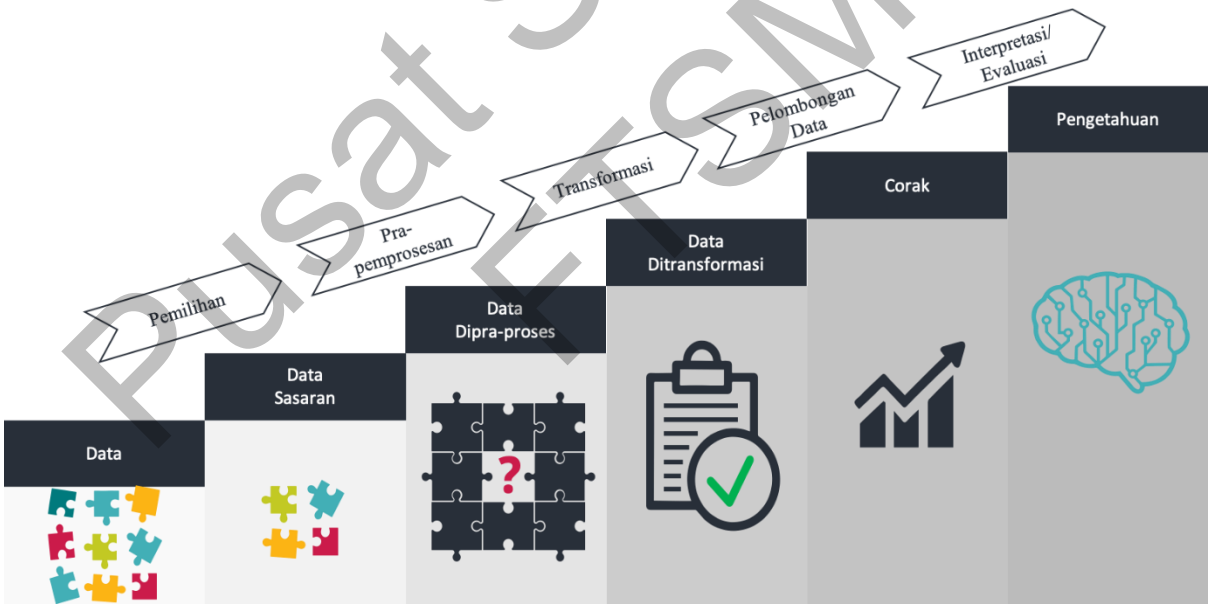
Kecerdasan buatan ialah satu sistem yang direka untuk menyesuaikan diri dan belajar. Definisi awal kecerdasan buatan datangnya daripada ujian Turing yang dilaksanakan untuk menguji kebolehan mesin dalam mendemonstrasikan kecerdasan. Hal ini banyak diaplikasikan dalam menyelesaikan masalah berkaitan pengelasan, peramalan dan pengoptimuman termasuklah dalam bidang pendidikan.

Pembelajaran mesin adalah subbidang daripada kecerdasan buatan. Untuk memiliki kecerdasan, sesebuah sistem perlu mempunyai kebolehan untuk belajar dalam persekitaran yang sentiasa berubah-ubah. Setelah sistem itu boleh belajar dan menyesuaikan diri terhadap sebarang perubahan, sistem boleh berdikari dan menghasilkan penyelesaian bagi semua masalah yang mungkin terjadi. Dengan cara ini, mendapatkan pengetahuan daripada data tidak

lagi bergantung kepada manusia tetapi boleh dibuat oleh komputer dan lain-lain peralatan elektronik (Sani et al., 2018).

Perlombongan data ialah aplikasi dan penggunaan algoritma yang spesifik dalam mengeluarkan corak daripada data. Istilah ini sering digunakan oleh ahli statistik, penganalisis data dan komuniti pengurusan sistem informasi. Teknik perlombongan data kebanyakannya berdasarkan kaedah statistik dan juga pembelajaran mesin, di mana teknik ini tergolong dalam bidang kecerdasan buatan (Fayyad et al., 1996).

Perlombongan data adalah langkah penting dalam proses penerokaan pengetahuan. Proses penerokaan pengetahuan melibatkan beberapa fasa seperti berikut : fasa pertama ialah integrasi dan penyediaan data; fasa kedua ialah pra-pemprosesan data; fasa ketiga ialah transformasi data dengan cara pengurangan data iaitu mengenal pasti atribut penting yang menerangkan sesuatu data; fasa keempat ialah perlombongan data; fasa kelima ialah persembahan dan interpretasi pengetahuan di mana corak keputusan ditafsir dalam bentuk yang bermakna dan mudah difahami.



Rajah 2.1 : Langkah-Langkah Dalam Proses Penerokaan Pengetahuan

Sumber: Fayyad et al. (1996).

Keupayaan untuk meramal pencapaian akademik pelajar adalah sangat penting dalam lapangan pendidikan. Kesan ke atas pencapaian akademik ini datang dari pelbagai punca seperti faktor peribadi, sosial, psikologi dan persekitaran (Al-Hagery et al., 2020; Mayra &

Mauricio, 2018). Oleh itu, penggunaan kaedah perlombongan data dalam mencapai objektif kajian dilihat amat bersesuaian kerana kebolehannya mengendalikan jumlah data yang besar, berupaya mencari corak dan perhubungan yang tersembunyi dan keputusannya dapat digunakan oleh para pembuat dasar.

Hellas et al. (2018) menyatakan dalam kajian meramal prestasi pelajar, beberapa kaedah yang selalu digunakan oleh para pengkaji boleh dikategorikan kepada beberapa kumpulan iaitu pengelasan (pembelajaran berselia), pengelompokan (pembelajaran tanpa selia), perlombongan (mencari corak yang kerap dan/atau pengekstrakan ciri) dan statistik (korelasi, regresi, ujian t dan lain-lain).

Perlombongan data pendidikan (EDM) boleh menjawab beberapa persoalan yang timbul dari penemuan pengetahuan berdasarkan corak tersembunyi yang didapati daripada data pelajar seperti (Kabra & Bichkar, 2011):

i. Siapakah pelajar yang berisiko?
ii. Apakah peluang penempatan latihan industri pelajar?
iii. Siapakah pelajar yang berisiko tercicir?
iv. Apakah kualiti pelajar?
v. Apakah kursus pengajian yang patut ditawarkan kepada pelajar?

Semua pihak dalam bidang pendidikan boleh mengambil manfaat dengan mengaplikasikan perlombongan data ke atas data pendidikan tinggi. Hasil yang diperoleh seperti sampel yang menarik, trend dan juga maklumat tersembunyi dapat membantu pihak-pihak berkepentingan dalam meningkatkan proses pengajaran, penerokaan dan penerangan terhadap fenomena yang berlaku dalam bidang pendidikan (Agić et al., 2014).

2.3 PRESTASI PELAJAR

Prestasi pelajar merupakan satu tema yang popular yang telah dijumpai ketika kajian kepustakaan dijalankan. Beberapa artikel kajian menggariskan terma “pelajar yang berisiko” dalam kajian mereka yang merujuk kepada golongan pelajar yang dalam keadaan tidak bernasib baik, tidak mempunyai kelebihan dan lemah dalam pelajaran (Nik Nurul Hafzan et

al., 2019; Ortiz-Lozano et al., 2018; Soobramoney & Singh, 2019). Beberapa kajian lain yang menggunakan terma "prestasi pelajar" memberi fokus kepada pencapaian pelajar yang berisiko sepanjang pengajian (Ekubo & Esiefarienrhe, 2019; Tomasevic et al., 2020; Widyaningsih et al., 2019). Justeru, dapat disimpulkan dalam kajian ini bahawa prestasi pelajar adalah elemen penting di IPT dan boleh didefinisikan sebagai prestasi dalam peperiksaan atau pentaksiran, prestasi dalam kos atau program dan keciciran daripada pengajian.

Keciciran pelajar daripada pengajian adalah satu masalah serius dalam dunia pendidikan kerana ia melibatkan ramai pelajar di sekolah mahupun IPT di seluruh dunia. Kesannya melibatkan masalah kewangan, kadar bergraduat yang rendah dan kejatuhan reputasi institusi yang terlibat dari pandangan masyarakat (Márquez-Vera et al., 2016). Justeru, banyak kajian berimpak yang telah dilakukan oleh para penyelidik dalam meramal masalah keciciran pelajar ini. Kajian-kajian ini telah menggunakan pelbagai jenis teknik peramalan, mengambil kira pelbagai faktor serta atribut dan menggunakan pelbagai pendekatan semasa pra-pemprosesan data.

Chen et al. (2018) telah menjalankan kajian terhadap keciciran pelajar kolej dalam bidang sains, teknologi, kejuruteraan dan matematik (STEM) terutamanya di permulaan pengajian mereka. Pengkaji telah melaksanakan peramalan keciciran pelajar dengan kaedah pembelajaran mesin seperti regresi logistik, pohon keputusan, hutan rawak, Naive Bayes dan AdaBoost. Model ramalan AdaBoost mencatatkan ketepatan tertinggi dan diikuti dengan Naive Bayes. Atribut seperti GPA, masa enrolmen dan umur kemasukan ke universiti adalah signifikan dalam mempengaruhi keciciran pelajar.

Liang et al. (2016) dalam kajiannya menggunakan rekod aktiviti pelajar di dalam log server bagi meramal keciciran pelajar dari platform pembelajaran atas talian. Teknik pengelasan pembelajaran berselia yang telah diaplikasikan ialah mesin sokongan vektor, regresi logistik, hutan rawak dan pohon keputusan dengan *gradient boosting*. Model peramalan pohon keputusan dengan *gradient boosting* telah menunjukkan ketepatan 88% iaitu lebih tinggi berbanding model pengelas yang lain.

Kabra dan Bichkar (2011) dalam kajiannya memberikan hujah bahawa kebanyakan atribut yang dipilih untuk dibuat model peramalan adalah berdasarkan pencapaian lepas kerana bertindak sebagai penunjuk bagi pencapaian akan datang seseorang pelajar itu. Beberapa kajian

terdahulu (Shahiri et al., 2015) telah mendapati CGPA adalah atribut yang paling berpengaruh dalam menentukan survival pelajar sama ada berupaya menghabiskan pengajian atau sebaliknya. CGPA juga mempunyai nilai atau perhubungan yang ketara dengan tahap pendidikan dan mobiliti karier seseorang pada masa akan datang. Selain itu, atribut data persendirian pelajar seperti pendapatan dan pendidikan ibu bapa adalah susah untuk diperolehi kerana keengganan untuk berkongsi maklumat ataupun pemalsuan maklumat itu sendiri.

Pelbagai atribut telah menjadi penyumbang kepada keciciran pelajar dalam pelajaran. Para penyelidik telah menggunakan atribut-atribut ini dalam meramal pencapaian pelajar menggunakan teknik perlombongan data pendidikan. Berdasarkan tinjauan yang telah dijalankan oleh (Kumar et al., 2017) terhadap beberapa kertas penyelidikan yang telah diterbitkan, beliau dapat menyenaraikan 10 atribut penting dalam peramalan keciciran pelajar. Atribut-atribut tersebut ialah gred peperiksaan, jantina, struktur keluarga, kelayakan ibu bapa, pekerjaan ibu bapa, keperluan menjalankan kerja di rumah, ketagihan (alkohol, rokok, dadah dan sebagainya), kemudahan asas di institusi pendidikan, kelemahan teknik pengajaran dan status berkahwin.

Analitik ke atas data pelajar yang mengandungi prestasi akademik dan tingkah laku adalah sangat penting dalam memahami mengapa keciciran berlaku dengan kerap terutama di IPT awam. Banyak kajian yang telah dilakukan dan mengambil kira faktor-faktor persekitaran pelajar yang mempengaruhi kadar bergraduat dan keciciran seperti masalah akademik, kewangan, motivasi dan masalah dalaman institusi. Namun, keciciran wujud bukannya disebabkan oleh satu faktor yang spesifik, tetapi dipengaruhi oleh pelbagai faktor dan permasalahan yang berlainan dari setiap universiti (Nik Nurul Hafzan et al., 2019; Perez et al., 2018).

Penglibatan murid dalam pembelajaran mempunyai pelbagai definisi. Ia boleh dikategorikan kepada penglibatan akademik dan penglibatan sosial. Penglibatan merupakan peramal yang kuat dalam pencapaian akademik (Dehyadegary et al., 2012) dan mempunyai hubungan dengan pencapaian pelajar (Sbrocco, 2009). Teori Fredricks, Blumenfeld dan Paris (2004) menjelaskan dapatan bahawa penglibatan murid mempunyai hubungan signifikan yang positif terhadap pencapaian akademik pelajar. Beliau memfokuskan penglibatan berdasarkan tiga dimensi iaitu, tingkah laku, emosi dan kognitif. Berdasarkan kajian oleh Sbrocco (2009), penglibatan tingkah laku didefinisikan sebagai penglibatan aktif dalam akademik dan juga

bukan akademik, dan ia mempengaruhi pencapaian murid. Manakala, menurut Hughes et al.(2008) penglibatan tingkah laku ialah penglibatan akademik dan sosial atau aktiviti tambahan. Terdapat tiga tahap dalam penglibatan tingkah laku iaitu; i) tingkah laku yang berkaitan dengan pembelajaran seperti ketekunan usaha, tumpuan, perhatian, bertanya soalan dan sumbangan dalam perbincangan kelas; ii) mematuhi peraturan sekolah; iii) keterlibatan dalam aktiviti kurikulum. Oleh itu, dapat disimpulkan bahawa penglibatan tingkah laku pelajar dan pencapaian saling berkaitan antara satu sama lain.

Contoh-contoh kajian yang telah dijalankan dalam skop peramalan prestasi pelajar dengan menyenaraikan CGPA, keciciran dan gred serta markah peperiksaan sebagai ciri-ciri yang diramal adalah seperti ditunjukkan di dalam Jadual 2.1.

Jadual 2.1. Kajian-Kajian Dalam Skop Peramalan Prestasi Pelajar.

Aspek Prestasi Pelajar Yang Diramal	**Penulis dan Tahun**	**Model Terbaik Yang Digunakan**	**Ketepatan (%)**
CGPA	Xu et al (2017)	*Ensemble*	-
Keciciran	Gil et al (2020)	Pohon keputusan	98.95
	Lee & Chung (2019)	*Ensemble*	-
	Chen et al (2018)	*Ensemble*	-
	Dharmawan et al (2018)	Pohon Keputusan	66.00
	Mayra & Mauricio (2018)	Pohon Keputusan	98.00
	Perez et al (2018)	Pohon Keputusan	94.00
	Samsiah Sani et al (2018)	Pohon Rawak	95.93
Gred Peperiksaan Akhir	Govindasamy & Velmurugan (2018)	*Fuzzy C-Means*	-
	Predic et al (2018)	*Ensemble*	90.43
Gred	PanduRanga et al (2019)	*Hierarchical clustering*	-
	Rana & Garg (2017)	Regresi	-
Markah Peperiksaan	Krizanic (2020)	Pohon Keputusan	-

Oleh itu, selain faktor maklumat pelajar dan prestasi pelajar, kajian ini akan membangunkan model pengelompokan pelajar B40 berdasarkan faktor penglibatan tingkah laku pelajar iaitu memfokus kepada keterlibatan pelajar dalam aktiviti kurikulum sepanjang pengajian di IPTA. Kajian korelasi juga akan dijalankan untuk melihat perkaitan antara faktor-

faktor tersebut dengan kumpulan pelajar B40 yang akan dihasilkan berdasarkan teknik pengelompokan tanpa selia.

2.4 KUMPULAN B40

Kemiskinan adalah masalah sejagat yang berlaku di seluruh pelusuk dunia. Justeru Pertubuhan Bangsa-Bangsa Bersatu (PBB) melalui Pelan Pembangunan Mapan menggariskan misi melenyapkan kemiskinan ekstrim di dunia menjelang tahun 2030 sebagai matlamat nombor satu mereka.

Di Malaysia, B40, M40 dan T20 merupakan kumpulan isi rumah yang wujud dalam struktur taburan pendapatan isi rumah. B40 dikasifikasikan sebagai isi rumah yang berpendapatan di bawah RM 4,360.00 sebulan. Berdasarkan laporan "Pendapatan dan Perbelanjaan Isi Rumah M40 dan B40 mengikut Negeri" yang dikeluarkan oleh Jabatan Perangkaan Malaysia pada tahun 2020, sebanyak 2.78 juta isi rumah tergolong dalam kumpulan B40 dan tiga negeri mencatatkan bilangan tertinggi iaitu Perak (12.5%), Selangor (11.4%) dan Sarawak (11.0%).

Laporan Statistik Utama Tenaga Buruh di Malaysia oleh Jabatan Perangkaan Malaysia menyatakan bilangan penganggur adalah seramai 517 ribu orang pada tahun berakhir Disember 2019. Angka tersebut menunjukkan terdapat peningkatan 0.5% berbanding bulan Disember pada tahun sebelumnya. Angka ini adalah sangat besar dan apa yang menjadi kerisauan adalah ada dalam kalangan graduan dan siswazah B40 tergolong dalam golongan yang menganggur. Situasi ini menunjukkan bahawa peluang melanjutkan pelajaran dan bergraduasi di menara gading pada ketika ini bukan lagi menjadi jaminan bagi anak-anak golongan B40 untuk mengubah nasib diri dan keluarga untuk keluar dari belengu kemiskinan.

Kadar kemiskinan akan berkurangan sekiranya semakin ramai golongan itu mendapat akses kepada pendidikan yang berkualiti. Justeru Kementerian Pendidikan Malaysia, menerusi Menteri Pendidikan telah meningkatkan kuota kepada anak-anak keluarga miskin B40 bagi kemasukan ke sekolah berasrama penuh dan IPT. Kelebihan ini akan membuka ruang kepada anak-anak B40 untuk mendapatkan pendidikan berkualiti sama ada di peringkat menengah mahupun pengajian tinggi. Hal ini telah menjadi fokus utama pendidikan negara sejak

kebelakangan ini iaitu memberi manfaat untuk pelajar B40 sebagaimana terkandung di dalam 9 Teras Pencapaian Setahun KPM (2019).

Hubungan antara kemiskinan dengan pencapaian pendidikan dapat dikaitkan dengan kesan tingkah laku pelajar yang bermasalah. Shong et al. (2019) dalam kajiannya mendapati faktor penyumbang utama kepada keterlibatan pesalah juvana dalam jenayah adalah kerana faktor kemiskinan. Faktor kemiskinan ini membawa kepada beberapa keadaan seperti institusi keluarga menjadi porak-peranda, kegagalan dalam pelajaran dan keciciran dari persekolahan serta terpengaruh dengan rakan-rakan yang tidak bermoral. Jika tidak ditangani secara efisien, masalah ini boleh menyebabkan gejala sosial meningkat dan tahap pendidikan golongan B40 menjadi semakin rendah.

Kajian oleh Aina et al. (2021) terhadap penerbitan sosio-ekonomi dan keciciran pelajar di pengajian tinggi mendapati keciciran berlaku disebabkan faktor-faktor seperti individu, institusi dan ekonomi serta ditambah dengan kebolehan pelajar menyesuaikan diri dalam sistem akademik itu sendiri. Tambahan pula, pelajar dengan latar belakang kemiskinan dan datang daripada keluarga berpendapatan rendah adalah sangat berpotensi untuk tercicir daripada pelajaran kerana sangat terkesan dengan yuran pengajian yang tinggi. Umum mengetahui bahawa kemiskinan merupakan satu faktor yang utama dalam menyebabkan keciciran manakala godaan untuk memasuki bidang pekerjaan pula merupakan faktor penyumbang yang kerap kali berlaku. Pelajar-pelajar miskin ini terpaksa bekerja mencari sumber kewangan bagi membantu menambah pendapatan keluarga mereka. Perkara besar yang terpaksa dikorbankan oleh mereka ialah masa untuk belajar dan membuat kerja rumah (Rashid & Samat, 2018). Keadaan di atas menunjukkan bahawa faktor-faktor yang mempengaruhi prestasi pelajar miskin terutamanya golongan isi rumah B40 perlu dikaji dan dikenal pasti. Hal ini bersesuaian dengan situasi belia negara kita pada masa kini yang sedang bergelut dengan isu pendidikan, pengangguran dan juga miskin bandar.

2.5 TEKNIK PENGELOMPOKAN

Sejak kebelakangan ini, keberkesanan penggunaan teknik pengelompokan dalam kajian peramalan prestasi pelajar telah menarik minat ramai pengkaji (Abu Saa et al., 2019; Alyahyan & Düştegör, 2020; Avella et al., 2016a; Hellas et al., 2018). Teknik pengelompokan merujuk

kepada satu kaedah mengumpulkan beberapa objek yang serupa ke dalam satu kelompok manakala objek yang berbeza ke dalam kelompok yang lain (Singh et al., 2016). Pengiraan jarak kebiasaannya dilakukan dalam menentukan sejauh mana persamaan sesuatu objek tersebut. Tujuan pemilihan jarak dilakukan adalah kerana dapat mendedahkan corak atau formasi kelompok yang terbentuk daripada set data yang besar. Berdasarkan kelompok data yang telah terbina, contoh-contoh data yang baru boleh dikelaskan berdasarkan jarak terdekat dengan kelompok yang ada.

Teknik pengelompokan akan menjadi sangat berguna sekiranya ciri dan kategori dalam sesebuah kumpulan atau set data itu tidak diketahui (Avella et al., 2016a). Di samping itu, pembahagian set data yang besar kepada beberapa kelompok kecil yang logik akan memudahkan para pengkaji untuk memeriksa dan menerangkan maksud sesuatu data itu.

Jadual 2.2. Algoritma-Algoritma Bagi Teknik Pengelompokan Mengikut Kategori.

Kategori Pengelompokan	**Contoh Algoritma**
Pembahagian	k-min, k-medoids, fuzzy c-min, PAM, CLARA
Hierarki	BIRCH, CURE, DIANA, ROCK, AGNES
Ketumpatan	DBSCAN, OPTICS, DENCLUE
Grid	WaveCluster, STING
Model	expectation maximization (EM), Gaussian mixture model

Teknik pengelompokan sering digunakan ketika pra-pemprosesan data kerana dapat menyingkirkan data yang tidak diperlukan, melonggokkannya dalam satu kelompok dan seterusnya memudahkan analisis (Ahuja et al., 2019). Secara umumnya, teknik pengelompokan terbahagi kepada lima kategori iaitu algoritma pembahagian (partitioning), hierarki, berasaskan ketumpatan/kepadatan, berasaskan grid dan berasaskan model (Govindasamy & Velmurugan, 2018). Antara pendekatan algoritma yang mewakili kelima-lima kategori ini ditunjukkan di dalam Jadual 2.2 (Hancer et al., 2020; Mittal et al., 2019).

a. Algoritma k-min

Algoritma k-min merupakan salah satu teknik pengelompokan yang popular dan menjadi pilihan para pengkaji untuk digunakan bagi tujuan mengelompok data. Penggunaannya adalah sangat meluas dan digunakan dalam pelbagai domain seperti analisis data secara statistikal, perlombongan data dan lain-lain solusi industri (Ali et al., 2020). Teknik ini popular dan menjadi pilihan ramai kerana cara implementasinya adalah sangat mudah dan keputusannya pula senang untuk difahami (Mohd Ariff et al., 2020).

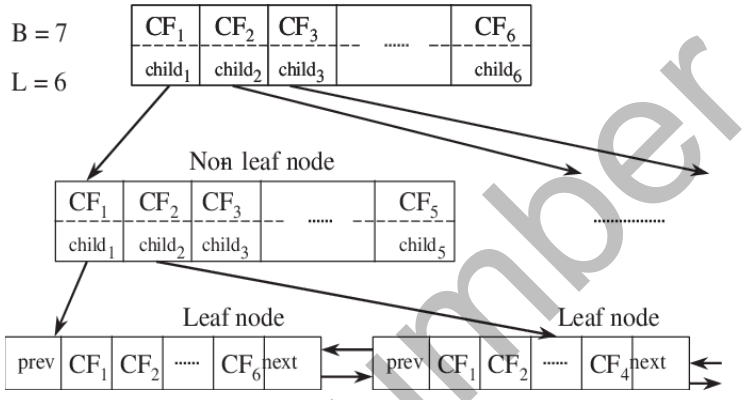
Algoritma k-min bagi pengelompokan adalah sejenis pembelajaran tanpa selia di mana kaedah ini bertujuan untuk mengelompokkan objek-objek berhampiran ke dalam *k* bilangan sentroid di mana koordinat setiap sentroid adalah min koordinat setiap objek di dalam setiap *k* kelompok (PanduRanga Vital et al., 2019).

Menurut Patel dan Thakral (2016), algoritma k-min mempunyai beberapa kelemahan seperti terlalu bergantung kepada pemulaan, sensitif kepada *outliers*, hanya boleh dijalankan terhadap kelompok dengan taburan data berbentuk sfera yang simetri dan penentuan nilai *k*. Oleh kerana pengelompokan k-min sangat sensitif kepada *outliers*, maka teknik baru iaitu k-medoids dihasilkan dengan beberapa ciri penambahbaikan untuk memperbaiki kelemahan K-min.

b. Algoritma BIRCH

Algoritma pengelompokan BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) terdiri daripada dua langkah iaitu membina pohon *Clustering Feature* (CF) dan kemudiannya pengelompokan menyeluruh. Pohon CF yang digunakan oleh algoritma BIRCH ditunjukkan dalam Rajah 2.2 di bawah. Pada langkah pertama, BIRCH merangkumkan set data besar ke kawasan yang lebih kecil dan padat yang dipanggil pemasukan (*entry*) CF. Bagi setiap kelompok, BIRCH menyimpan hanya tiga nilai, iaitu (N, LS,SS) di mana ‘N’ ialah bilangan titik data di dalam kelompok, ‘LS’ ialah hasil tambah linear bagi titik-titik data dan ‘SS’ ialah hasil tambah kuasa dua bagi titik-titik data di dalam kelompok. Tiga nilai ini dipanggil CF dan disimpan di dalam sebuah pohon yang dipanggil pohon CF.

Langkah kedua pula melibatkan penerapan algoritma pengelompokan ke atas nod akhir atau daun di pohon CF. Setiap nod akhir atau daun mengandungi atau menyimpan sub kelompok. Setiap pemasukan di pohon CF mengandungi penunjuk ke nod anak dan satu pemasukan CF terdiri daripada sejumlah pemasukan CF di dalam nod anak. BIRCH juga dikenali sebagai pengelompokan dua langkah (*Two Step Clustering*) merujuk kepada dua langkah yang diterangkan di atas.



Rajah 2.2. Pohon CF Yang Digunakan Oleh Algoritma BIRCH. (Andritsos, 2007)

Terdapat tiga parameter untuk menjalankan algoritma ini iaitu *branching factor*, *threshold* dan *n_clusters*. Parameter *branching factor* ialah bilangan maksimum sub kelompok CF di setiap nod, *threshold* pula ialah bilangan maksimum titik data yang boleh diterima oleh nod daun di pohon CF dan *n_clusters* ialah bilangan kelompok yang dikembalikan selepas seluruh algoritma BIRCH selesai dijalankan.

c. Algoritma DBSCAN

Salah satu algoritma paling popular dalam kategori pengelompokan ketumpatan ialah DBSCAN (*Density-Based Spatial Clustering of Application with Noise*) dan telah diperkenalkan oleh Martin Ester, Hans-Peter Kriegel, Jorg Sander dan Xiaowei Xu pada tahun 1996 (Ester et al., 1996). Algoritma ini memisahkan titik-titik data kepada tiga bahagian. Bahagian pertama ialah titik utama iaitu titik-titik yang berada di dalam kelompok. Bahagian kedua ialah titik sempadan iaitu titik-titik yang termasuk ke dalam kawasan kejiranan titik utama. Bahagian terakhir ialah titik hingar iaitu titik-titik yang bukan termasuk dalam titik utama dan titik sempadan.

DBSCAN sangat sensitif kepada penetapan parameter *epsilon* di mana perubahan kepada nilai yang kecil akan menyebabkan kelompok-kelompok yang terhasil akan dikategorikan sebagai hingar. Manakala perubahan kepada nilai yang besar pula akan menyebabkan kelompok-kelompok dicantum menjadi padat (DeFreitas & Bernard, 2015). DBSCAN tidak memerlukan penetapan bilangan kelompok pada fasa pemulaan algoritma.

2.5.1 Kajian Lepas Teknik Pengelompokan

PanduRanga Vital et al. (2019) dalam kajiannya telah membuat analisis terhadap prestasi pelajar menggunakan pembelajaran mesin tidak berselia iaitu algoritma pengelompokan hierarki dan k-min. Teknik yang digunakan didapati memberikan keputusan yang baik dalam meramal prestasi gagal dan lulus pelajar. Pengelompokan hierarki telah menunjukkan faktor penyumbang utama yang mempengaruhi keputusan pelajar melalui pertalian dalam dendogram seperti aktiviti ekstra kurikulum, kehadiran dan bilangan kelas yang gagal.

Rana & Garg (2016) juga telah menjalankan kajian yang hampir sama dengan menggunakan teknik pengelompokan hierarki dan k-min terhadap 58 pelajar semester ketiga bidang informasi teknologi. Ketepatan algoritma dijelaskan oleh contoh yang berjaya dikelaskan dengan tepat dalam kajian ini. Bilangan atribut diubah-ubah daripada lapan ke enam dan pengelasan hierarki menunjukkan peningkatan ketepatan namun k-min adalah sebaliknya. Masa yang dicatat untuk membina model juga adalah singkat iaitu kosong saat.

Križanić (2020) telah menjalankan penyiasatan elemen kelakuan pelajar yang direkodkan di dalam sistem e-pembelajaran yang mana boleh menyumbang kepada kejayaan dalam peperiksaan. Teknik perlombongan data yang digunakan adalah k-min dan pohon keputusan. Satu model oleh analisis kelompok menunjukkan tiga kumpulan pelajar berdasarkan kelakuan mereka dalam e-pembelajaran dan tiga model pohon keputusan dibina berdasarkan analisis kelompok itu tadi. Ciri yang memberikan *information gain* tertinggi ialah markah pelajar dalam peperiksaan pertengahan penggal. Didapati kekerapan yang rendah dalam mengakses bahan kuliah dan pembelajaran dalam talian akan menyebabkan pencapaian skor yang rendah dalam peperiksaan.

Selain daripada itu, Govindasamy & Velmurugan (2018) telah menjalankan satu analisis perbandingan melibatkan empat algoritma pengelompokan iaitu k-min, k-medoids,

fuzzy c-min dan e*xpectation maximization* (EM). Data pelajar kolej seramai 1,531 orang telah digunakan bagi meramal prestasi pelajar dalam peperiksaan akhir semester. Keputusan kajian mendapati *fuzzy c-means dan* EM adalah lebih baik kualiti pengelompokannya dari segi *purity* dan *normalized mutual information* (NMI) namun masa pelaksanaannya yang lebih lama.

Navarro dan Moreno-Ger (2018) dalam kajiannya telah menggunakan tujuh algoritma pengelompokan terhadap set data pendidikan yang mengandungi beberapa set gred pencapaian pelajar dalam kos-kos pengajian yang berbeza. Dapatan kajian mendapati teknik k-min dan PAM adalah yang terbaik dalam kategori pembahagian manakala DIANA dan hierarki pula adalah yang terbaik bagi kategori hierarki. Selain itu, mereka juga mendapati gred pencapaian pelajar sangat mudah untuk dikelompokkan dan boleh diaplikasikan pada set data pendidikan yang lain.

Kajian-kajian lain juga telah menunjukkan persamaan dalam menggunakan pembelajaran mesin tanpa selia bagi membuat ramalan prestasi pelajar (Prabha & Shanmuga Priyaa, 2017; Tabrez Nafis & Taha Owais, 2017). Teknik pengelompokan yang digunakan menunjukkan prestasi yang baik dalam membuat ramalan dan menghasilkan corak yang menarik bagi penggunaan set data pelajar yang bersaiz besar. Jadual 2.3 di bawah menunjukkan senarai kajian lepas mengikut jenis algoritma pengelompokan dalam domain pendidikan.

Berdasarkan kajian-kajian lepas yang telah dibuat sorotan, algoritma k-min, BIRCH dan DBSCAN telah dipilih untuk digunakan pada fasa pembangunan model pengelompokan bagi kajian ini. Ketiga-tiga algoritma ini telah menunjukkan prestasi yang baik dalam membuat pengelompokan pelajar dan sesuai dalam mengelas kumpulan pelajar mengikut prestasi dan tingkah laku. Selain itu, algoritma k-min merupakan satu teknik pengelompokan yang paling banyak digunakan dalam kajian-kajian lepas dan secara konsisten menghasilkan prestasi pengelompokan yang lebih baik berbanding algoritma yang lain dengan menggunakan data pelajar. Pemilihan algoritma BIRCH dan DBSCAN dalam kajian ini juga dapat mempelbagaikan penggunaan algoritma daripada pelbagai jenis kategori pengelompokan.

Jadual 2.3. Senarai Kajian Dan Pemilihan Jenis Algoritma Pengelompokan Dalam Domain Pendidikan

Penulis dan Tahun	**Algoritma Pengelompokan**
Palani et al (2021)	fuzzy c-min, hierarchical, gaussian mixture, k-prototype
Li et al (2021)	k-min, DBSCAN, BIRCH, CLIQUE, EM
Krizanic (2020)	k-min
Al-Hagery et al (2020)	x-min, k-min
Saric-Grgic(2020)	mean shift
Mallik et al (2019)	mean shift, k-min
Francis & Babu (2019)	k-min
PanduRanga et al (2019)	k-min, hierarchical
Macedo et al (2019)	fuzzy c-min
Valarmathy et al (2019)	EM, CLOPE, DBSCAN, k-min, CLARA, filtered cluster, farthest first
Govindasamy et al (2018)	k-medoid, EM
Prabha et al (2017)	k-min, k-medoid
Tabrez Nafis et al (2017)	k-min, k-medoid, x-min
Rana & Garg (2016)	k-min, hierarchical
Li et al (2016)	fuzzy c-means

2.5.2 Kelebihan Dan Kekurangan Algoritma Pengelompokan

Bhagat et al. (2016) dan Garg et al. (2020) telah membuat kajian perbandingan terhadap pelbagai teknik pengelompokan dalam perlombongan data dan berupaya menyenaraikan kelebihan dan kekurangannya bagi setiap algoritma. Jadual 2.4 di bawah menyenaraikan dan menerangkan setiap kelebihan dan kekurangan tersebut secara ringkas mengikut teknik pengelompokan yang dipilih untuk digunakan dalam kajian ini.

Jadual 2.4. Kelebihan Dan Kekurangan Algoritma Pengelompokan Yang Digunakan Dalam Kajian Ini.

Algoritma	**Jenis**	**Metrik**	**Kelebihan**	**Kekurangan**
k-min	Pembahagian	Jarak Euclidean antara titik	● Mencatatkan prestasi yang baik dengan banyak jenis set data ● Berfungsi dengan baik terhadap set data yang besar dan berdimensi tinggi	● Memerlukan pelbagai nilai k untuk menentukan kelompok terbaik ● Prestasi bergantung kepada tetapan pemulaan ● Hanya berfungsi dengan baik terhadap data berbentuk cembung dan elips
BIRCH	Hierarki	Jarak Euclidean antara titik	● Berguna untuk analisis data yang mengandungi struktur hierarki ● Efisien dalam menganalisis set data yang besar	● Tidak memberikan skala yang baik dengan data berdimensi tinggi
DBSCAN	Ketumpatan	Jarak Euclidean antara titik terdekat	● Berupaya mengenalpasti kelompok yang berbentuk arbitrari ● Tidak memerlukan tetapan bilangan kelompok sebagai pemulaan	● Pengendalian *outlier* yang lemah mempengaruhi ketepatan kelompok

2.6 TEKNIK PENGELASAN

Pengelasan adalah satu teknik pembelajaran mesin secara berselia di mana sesuatu model itu telah dilatih dengan satu set data berlabel iaitu setiap atribut di dalam set data dilabelkan sebagai kelas. Pengelasan adalah satu masalah kajian yang besar dalam perlombongan data pendidikan. Secara umumnya, pengelasan boleh didefinisi sebagai pembangunan model peramalan berdasarkan aspek pemboleh ubah tak bersandar seterusnya merumuskan nilai

pemboleh ubah bersandar (Predić et al., 2018). Terdapat dua tujuan utama yang boleh dihuraikan apabila peramalan diaplikasikan dalam perlombongan data pendidikan. Pertama ialah meramal serta membekalkan maklumat hasil dapatan daripada proses pengajaran dan pembelajaran tanpa meramal faktor dalaman. Yang kedua, meramal pemboleh ubah output dalam konteks label yang dihasilkan seperti aplikasi algoritma pengelasan dalam membina model ramalan bagi pelbagai permasalahan pendidikan. Predic juga dalam kajiannya menyatakan ketepatan model ramalan berdasarkan pengelasan bergantung kepada faktor-faktor seperti sifat data, saiz data dan kewujudan nilai yang tidak tepat serta nilai yang hilang dalam dataset yang digunakan.

a. Algoritma Pohon Keputusan

Pohon keputusan ialah struktur data berbentuk hierarki yang mengimplementasikan strategi pecah dan perintah. Setiap pohon keputusan terdiri daripada nod yang mewakili atribut, dahan atau cabang mewakili nilai dari atribut manakala daun pula mewakili label kelas. Nod teratas dalam pohon digelar sebagai akar. ID3, C4.5, CART dan J48 adalah antara pilihan algoritma pengelasan pohon keputusan yang menggunakan kriteria berbeza semasa pembahagian nod dalam dan kerap digunakan oleh para pengkaji dalam kajian-kajian lepas yang berasaskan pendekatan pembelajaran berselia (Predić et al., 2018).

Beberapa teknik pengelasan telah dikaji dalam kajian kepustakaan dan pohon keputusan telah dikenal pasti sebagai pengelas yang sesuai dengan kajian menggunakan data pendidikan. Pohon keputusan merupakan satu kaedah pengelasan yang banyak digunakan dalam perlombongan data pendidikan kerana keputusan yang dihasilkan dalam bentuk infografik dan visual adalah amat mudah untuk difahami (Križanić, 2020).

b. Algoritma Hutan Rawak

Model pengelasan menggunakan algoritma hutan rawak adalah satu model berasaskan pembelajaran bergabung (*ensemble learning*) di mana terdapat beberapa pengelas individu iaitu satu set pohon keputusan yang digabungkan bagi meningkatkan prestasi peramalan. Algoritma ini dilaporkan sangat mudah untuk digunakan dan sangat stabil serta mempunyai banyak ciri-ciri yang menarik (Beaulac & Rosenthal, 2019). Salah satu ciri tersebut ialah

kebolehannya untuk membenarkan pengiraan kepentingan atribut dengan menilai kepentingan peramal individu sepanjang keseluruhan proses ramalan.

Tidak memfokuskan kepada proses mencari keputusan terbaik daripada keseluruhan pohon keputusan, algoritma hutan rawak sebaliknya akan memilih subset peramal secara rawak dan seterusnya akan mencari keputusan terbaik di antaranya. Pengubahsuaian ini menjadi kelebihan hutan rawak kerana dapat mengurangkan masalah *overfitting* di dalam pohon keputusan dan seterusnya dapat membantu meningkatkan prestasi pengelasan.

c. Algoritma ANN

Rangkaian neural buatan atau *artificial neural network* (ANN) adalah antara model popular yang digunakan dalam perlombongan data pendidikan. Operasi ANN terdiri daripada fungsi pengiraan matematik yang diimplementasikan pada sistem pengkomputeran. Model ANN digambarkan sebagai satu set perhubungan antara input kepada output di mana rangkaian nod-nod yang menghubungkannya mempunyai pemberat. Untuk memproses data dengan jumlah yang besar, ANN memerlukan lebih banyak neuron dan kuasa perkomputeran (Ogutcu, 2020). *Perceptron* pelbagai lapisan ialah satu kelas ANN kaedah suapan ke hadapan. Kaedah ini mempunyai tiga lapisan yang terdiri daripada lapisan input, satu atau lebih lapisan tersembunyi dan lapisan output.

Kelebihan rangkaian neural ialah kebolehannya untuk mengesan semua interaksi yang berkemungkinan berlaku antara pemboleh ubah peramal. Selain daripada itu, rangkaian neural juga boleh membuat pengesanan lengkap tanpa sebarang kemusykilan di dalam hubungan tak linear yang kompleks antara pemboleh ubah bersandar dan tak bersandar (Shahiri et al., 2015).

2.6.1 Kajian Lepas Teknik Pengelasan

Kajian dan pembangunan model peramalan prestasi pelajar dalam domain pendidikan di IPT telah dilaksanakan secara meluas semenjak akhir-akhir ini. Sekiranya fokus diarahkan ke Malaysia, terdapat beberapa kajian dalam beberapa tahun kebelakangan (2015 – 2020) yang mengaplikasikan teknik pengelasan dalam bidang pendidikan.

Shahiri et al. (2015) dalam kajiannya telah meninjau 30 kertas kajian sejak dari tahun 2002 hingga 2015 dan mendapati kaedah paling popular dalam meramal prestasi dan pencapaian pelajar ialah kaedah pengelasan melibatkan algoritma pohon keputusan, rangkaian neural buatan (ANN), *naive bayes*, k-jiran terdekat dan mesin sokongan vektor. Keputusan tinjauan beliau mendapati algoritma yang menghasilkan ketepatan ramalan yang tertinggi ialah ANN diikuti oleh pohon keputusan.

Terbaru, Sani et al. (2020) dalam kajian mereka telah menggunakan teknik pembelajaran mesin bagi meramal keciciran dalam kalangan pelajar B40 yang mengambil ijazah sarjana muda di UA seluruh Malaysia. Tiga model berasaskan algoritma pohon keputusan, hutan rawak dan ANN telah dibangunkan menggunakan data sebenar pelajar UA yang diperolehi daripada KPT. Hasil kajian mendapati prestasi algoritma hutan rawak mengatasi algoritma-algoritma yang lain serta menunjukkan perbezaan yang signifikan secara statistik.

Nik Nurul Hafzan et al. (2019) pula telah melaporkan tinjauannya terhadap lebih 50 kertas kajian mengenai model peramalan bagi mengenal pasti pelajar berisiko di IPTA. Didapati hampir semua teknik pembelajaran mesin telah diuji dan yang paling popular adalah pohon keputusan, hutan rawak, naive bayes dan regresi. Teknik penyatuan telah menghasilkan prestasi yang baik mengatasi teknik yang lain kerana menggabungkan beberapa teknik yang berketepatan tinggi.

Wan Yaacob et al. (2020) dalam kajiannya telah mereka bentuk satu model untuk mengenal pasti faktor penyumbang utama terhadap kadar keciciran yang tinggi dalam program sains komputer. Mereka telah mendapatkan set data berkaitan demografi dan rekod transkrip yang menyenaraikan keputusan lengkap dalam kos-kos teras sains komputer dan memberikan impak secara langsung terhadap keciciran. Empat teknik pengelasan berbeza telah digunakan terhadap set data iaitu k-jiran terdekat, pohon keputusan, rangkaian neural dan regresi logistik. Keputusan kajian menunjukkan pengelas regresi logistik adalah yang terbaik berbanding yang lain dengan ketepatan sebanyak 91%. Selain itu, hasil kajian turut mendedahkan bahawa terdapat lima kos teras penting yang perlu pelajar tingkatkan pencapaian bagi meminimakan peluang untuk tercicir.

Selain kajian tersebut, beberapa pengkaji lain juga telah menerbitkan kajian menggunakan teknik pengelasan dalam domain pendidikan di Malaysia antaranya ialah (Makhtar et al., 2017; Nawang et al., 2018; Sangodiah & Balakrishnan, 2014). Mereka telah memberi fokus dan tumpuan terhadap peramalan prestasi pelajar serta peramalan keciciran pelajar di IPT dan sekolah di serata Malaysia.

Bedregal-Alpaca et al. (2020) telah mencadangkan pembinaan model pengelasan yang memanfaatkan maklumat akademik yang diperolehi daripada pihak universiti bagi mengenal pasti pelajar yang berisiko tercicir dalam pelajaran. Maklumat demografi, prestasi akademik, ujian kemasukan dan maklumat kursus pelajar telah digunakan untuk penilaian. Model yang dihasilkan berupaya menunjukkan bahawa ciri penting mempengaruhi prestasi ialah subjek yang ditinggalkan oleh pelajar.

Gil et al. (2020) telah mengadaptasi pohon keputusan dan naïve bayes bagi mengenal pasti faktor tersirat yang mempengaruhi keciciran pelajar di sebuah sekolah awam di Filipina. Pengkaji menggunakan perisian Weka bagi memanfaatkan algoritma pengelas terhadap set data pelajar. Hasil kajian yang dihasilkan mengandungi perbandingan antara setiap algoritma dari segi *recall*, ketepatan dan kepersisan. Penunjuk utama kadar keciciran pelajar ialah faktor prestasi akademik seperti yang ditunjukkan oleh algoritma pohon keputusan.

Manakala Mardolkar dan Kumaran (2020) hanya memfokuskan algoritma k-jiran terdekat dalam membuat penilaian dan meramal keciciran pelajar di peringkat awal pengajian. Teknik k-jiran terdekat adalah serba boleh, ringkas dan boleh mengendalikan pelbagai jenis data. Keputusan kajian boleh membantu ahli akademik dalam mengenal pasti pelajar-pelajar berisiko tercicir dan membantu mereka untuk memperbaiki prestasi dan kebajikan diri.

Márquez-Vera et al. (2016) dalam kajiannya telah mengadaptasikan teknik perlombongan data bagi mencari model peramalan yang komprehensif dalam meramal keciciran pelajar seawal yang mungkin. Model yang didapati mempunyai ketepatan yang tinggi akan digunakan sebagai sistem amaran awal yang juga bertindak sebagai satu mekanisme menentukan pelajar yang berisiko tercicir seawal yang mungkin. Mereka juga mendapati faktor akademik iaitu prestasi di universiti dan sekolah terdahulu, sosiodemografi, kelakuan dan penglibatan dalam aktiviti ekstra kurikulum mungkin mempengaruhi kadar keciciran. Daripada

set data besar yang diperiksa, hanya sebahagian kecil atribut yang menunjukkan sumbangan yang tinggi dalam membuat peramalan.

Tomasevic et al. (2020) telah melaporkan dalam penulisannya satu kajian tentang analisis komprehensif dan perbandingan antara teknik-teknik pembelajaran mesin berselia yang digunakan dalam bidang peramalan pelajar berisiko tercicir daripada pengajian. Bagi menjayakan kajiannya, beberapa pengelas telah digunakan antaranya KNN, mesin sokongan vektor, ANN, pohon keputusan, naïve bayes dan regresi logistik sebagai alat pengelasan. Secara keseluruhan, algoritma ANN memberikan ketepatan tertinggi setelah menganalisis data penglibatan pelajar di dalam pembelajaran atas talian serta data prestasi lampau. Keputusan kajian juga telah melaporkan yang atribut demografi tidak menunjukkan pengaruh yang signifikan kepada prestasi pelajar.

Penggunaan data pelajar kejuruteraan di India bagi menjalankan kajian meramal keciciran oleh (Viloria et al., 2019) telah menunjukkan keberkesanan. Mereka telah memanfaatkan algoritma ANN, pohon keputusan dan rangkaian bayesian berbanding algoritma lain. Hasil kajian mendedahkan keputusan akademik dan keadaan sosioekonomi mempengaruhi prestasi pelajar. Sekiranya pengurusan faktor-faktor ini dilakukan dengan baik, kadar keciciran dapat dikurangkan dalam kalangan pelajar tersebut.

Chen et al. (2018) telah menjalankan kajian bagi membangunkan model ramalan awal pelajar yang akan tercicir bagi dua tahun pertama pengajian di kolej melibatkan bidang STEM. Model ramalan dibangunkan dengan mengambil kira data keputusan akademik di kolej, maklumat kursus, GPA sekolah tinggi serta demografi pelajar. Seterusnya, mereka mengimplementasi kaedah pembelajaran mesin termasuk regresi logistik, pohon keputusan, hutan rawak, naive bayes dan *adaboost*. *Adaboost* telah menunjukkan prestasi yang baik berbanding algoritma yang lain. Teknik penyatuan pembelajaran mesin telah diakui sebagai sebuah teknik yang berpengaruh dalam bidang perlombongan data dan pembelajaran mesin dalam tempoh sedekad yang lalu. Teknik ini menggabungkan beberapa model individu menjadi satu dan kebiasaannya menghasilkan model akhir dengan ketepatan yang lebih tinggi dan teguh (Satyanarayana & Nuckowski, 2016). Hal ini kerana, teknik penyatuan akan meningkatkan kelemahan-kelemahan pengelas tunggal supaya boleh membuat ramalan yang lebih tepat.

Manakala (Liang et al., 2016) pula dalam penulisan kajian mereka telah menggunakan algoritma pohon keputusan *gradient boosting* beserta mesin sokongan vektor, regresi logistik dan hutan rawak bagi mengkaji ramalan keciciran pelajar dalam kursus pembelajaran atas talian. Log data yang mengandungi rekod aktiviti dan tingkah laku pelajar semasa menjalani 39 jenis kursus telah diambil. Kemudian, ramalan telah dibuat terhadap pelajar yang akan tercicir dalam masa 10 hari setelah kursus pembelajaran tersebut bermula. Teknik penyatuan telah berjaya menghasilkan 88% ketepatan iaitu yang paling tinggi berbanding teknik yang lain.

Berdasarkan kajian-kajian lepas yang telah dikupas dan dibuat sorotan, algoritma pohon keputusan, ANN dan hutan rawak telah dipilih untuk digunakan pada fasa output bagi membangunkan model pengelasan terbaik untuk mengelas pelajar B40 berdasarkan kelas prestasi pelajar di IPTA. Ketiga-tiga algoritma ini telah menunjukkan rekod yang baik disamping menghasilkan ketepatan yang tinggi bagi setiap model yang dibangunkan dengan menggunakan data-data pelajar di IPT seluruh dunia.

2.7 PEMILIHAN ATRIBUT

Pemilihan atribut adalah satu langkah dalam peringkat pra-pemprosesan di mana dapat mengurangkan dimensi sesuatu data ataupun membuang data yang tidak relevan. Proses ini mengambil kira hanya ciri subset yang berkenaan dan penting sahaja daripada keseluruhan ciri-ciri asal. Tujuannya adalah untuk mengeluarkan data yang bertindih dan juga hingar. Bagi sesetengah aplikasi, atribut yang tidak mempunyai korelasi dengan kelas bertindak sebagai hingar lalu menyebabkan ralat dalam peramal dan mengurangkan ketepatan ramalan. Apabila sebahagian data ini telah dikeluarkan, keupayaan perlombongan data dapat ditingkatkan dengan menaikkan ketepatan peramalan dan melajukan algoritma perlombongan data dalam menghasilkan keputusan yang menyeluruh (Makhtar et al., 2017; Shahiri et al., 2015). Di dalam kajian ini, teknik pemilihan atribut secara berselia akan menggunakan algoritma hutan rawak, pohon ekstra, *information gain* (IG) dan chi kuasa dua.

Antara ciri menarik yang ada pada hutan rawak adalah boleh digunakan untuk menentukan kepentingan setiap peramal individu yang terdiri daripada pemboleh ubah atau atribut (Beaulac & Rosenthal, 2019). Analisis kepentingan atribut ini bertujuan untuk

memahami kesan yang dibawa oleh setiap atribut di akhir eksperimen iaitu pada output pengelasan.

Pohon ekstra merupakan satu jenis teknik pembelajaran bergabung yang mengumpulkan beberapa keputusan daripada pohon keputusan bagi mengeluarkan output pengelasan. Secara konsep, pohon ekstra adalah hampir sama dengan hutan rawak namun berbeza dari segi pembinaan pepohon keputusannya. Sharma et al. (2019) dalam penulisannya menyatakan kriteria pemisahan pohon ekstra adalah secara rawak dan keseluruhan set latihan diambil kira. Pohon yang dihasilkan mengandungi lebih banyak nod daun seterusnya lebih efisien dari segi pengiraan atau perkomputeran. Kelebihan menggunakan pengelas berasaskan pohon (*tree-based*) dalam melakukan pemilihan atribut adalah memerlukan sedikit memori perkomputeran, laju dan berupaya menunjukkan hampir keseluruhan atribut yang penting di peringkat awal implementasi. Setelah setiap kali pemisahan (*split*) berlaku, setiap atribut akan diberi skor dan kemudiannya disusun mengikut kedudukan. Ciri-ciri yang menarik ini akan digunakan semasa pemilihan atribut dalam kajian ini.

IG merupakan satu kaedah penapisan dalam pemilihan atribut dan menggunakan pengukuran berasaskan statistik dan entropi (*entropy-based*) (Punlumjeak & Rachburee, 2015). Nilai maklumat sesuatu atribut dikira oleh IG berdasarkan nilai entropi berkaitan dengan kelas di mana atribut itu berada. Nilai entropi boleh berubah daripada 0 ke 1 di mana nilai 0 menunjukkan "tiada maklumat" dan nilai 1 pula menunjukkan "maklumat maksimum". Seterusnya fungsi *gain* diterapkan pada atribut-atribut supaya atribut dengan nilai maklumat yang lebih tinggi dapat dibezakan dengan atribut dengan nilai maklumat yang lebih rendah (Sani et al., 2018). Kaedah ini biasa digunakan dalam mengukur perkaitan ciri atau atribut dalam strategi penapisan yang menilai ciri secara individu (Pereira et al., 2015). Kelebihan kaedah IG dalam membuat pemilihan atribut adalah kerana kelajuannya dan berkebolehan

Ujian chi kuasa dua adalah satu kaedah dalam pemilihan atribut yang boleh dilakukan ke atas atribut data jenis kategori atau nominal (Nadiyah, 2019). Ujian ini mengukur perkaitan dan kebersandaran antara dua pemboleh ubah iaitu atribut dan kelas label. Atribut yang telah dinilai akan disusun mengikut kedudukan di mana semakin tinggi kedudukan menunjukkan semakin penting sesuatu atribut di dalam kajian.

Manakala pemilihan atribut tanpa selia dalam kajian ini pula telah menggunakan teknik nilai ambang varians (*variance threshold*). Chandrashekar dan Sahin (2014) menyatakan yang algoritma pemilihan atribut melalui tapisan adalah satu kaedah yang tidak melibatkan algoritma pembelajaran dalam menilai subset sesuatu atribut. Kaedah tapisan secara prinsipnya memilih sesuatu atribut dengan menyusunnya mengikut susunan keutamaan berdasarkan kebergantungan atau korelasi yang kuat dengan label kelas. Jelas disini bahawa terdapat satu nilai ambang yang ditetapkan dan jika skor sesuatu atribut itu berada di bawah nilai tersebut, atribut akan diabaikan kerana tidak berpengaruh ke atas label kelas. Pendekatan kaedah ini menggunakan pengukuran yang mudah seterusnya menghasilkan keputusan dengan cepat dan ringkas (Shahiri et al., 2017).

2.8 RUMUSAN

Penggunaan teknik pembelajaran mesin dalam domain pendidikan untuk membuat ramalan, pengelasan dan pengelompokan pelajar berdasarkan prestasi pengajian di IPT adalah sangat besar impaknya. Perbincangan mengenainya telah dibuat melalui kajian-kajian oleh pengkaji di luar mahupun di dalam negara. Walaubagaimanapun, melalui kajian kepustakaan yang telah dilakukan mendapati kajian seumpama ini yang menggunakan set data sebenar pelajar di IPT dan memfokuskan pelajar-pelajar miskin masih kurang menonjol untuk dikaji di Malaysia.

Hasil tinjauan berjaya mengenal pasti kaedah-kaedah pembelajaran mesin yang digunakan bagi menganalisis pencapaian akademik pelajar berdasarkan teknik pengelompokan di mana teknik k-min merupakan teknik yang sering digunakan oleh para pengkaji. Hal ini kerana teknik k-min mempunyai kelebihan dari segi implementasi yang sangat mudah dan keputusannya pula senang untuk difahami. Kemudian, teknik-teknik pengelasan yang sering digunakan untuk meramal pencapaian pelajar dalam kajian-kajian terdahulu meliputi teknik pohon keputusan, mesin sokongan vektor, ANN dan *naive bayes*. Teknik-teknik ini boleh menjadi asas dan rujukan dalam membina model pengelasan peramalan prestasi pelajar dengan ketepatan yang tinggi.

BAB III

METODOLOGI KAJIAN

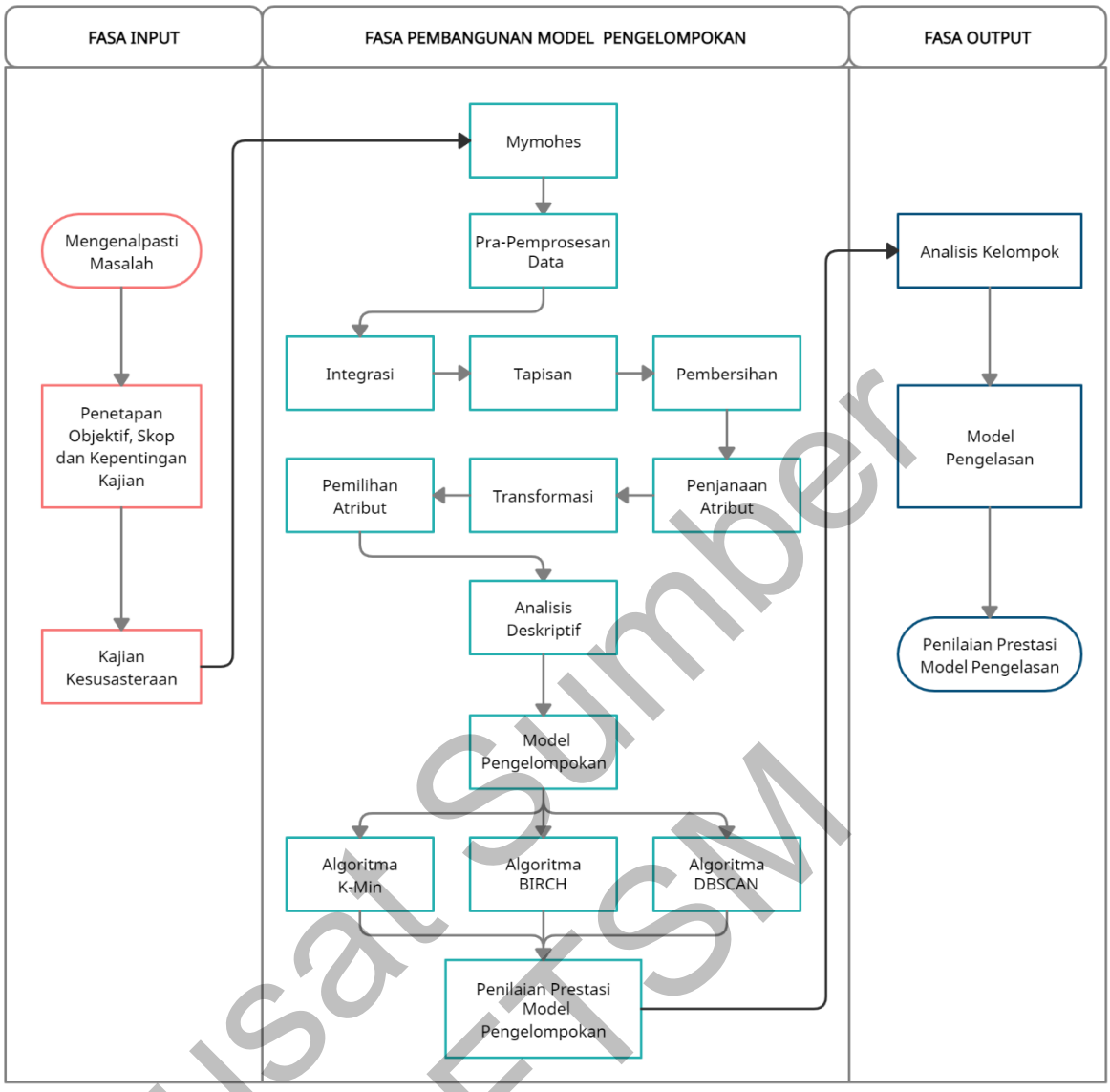
3.1 PENGENALAN

Bab ini membincangkan dengan lebih lanjut tentang metodologi yang digunakan dalam melaksanakan kajian ini. Ia bertujuan memberi penjelasan komprehensif mengenai rangka kerja pembangunan model pengelompokan bermula dengan fasa pengumpulan dan penyediaan data, penggunaan perisian dalam membangunkan model, penggunaan teknik pengekstrakan ciri untuk mengenalpasti atribut penting dan analisa serta penilaian model.

3.2 RANGKA KERJA

Rangka kerja bagi kajian ini merangkumi tiga fasa iaitu fasa input, fasa pembangunan model dan fasa output di mana visualisasi bagi setiap fasa ditunjukkan dalam Rajah 3.1 di bawah.

Fasa input bagi kajian ini adalah di mana proses mengenalpasti masalah kajian setelah beberapa perkara dipertimbangkan. Seterusnya, penetapan objektif, skop dan kepentingan kajian akan dilakukan berdasarkan permasalahan kajian. Kajian kesusasteraan yang mengambil kira kajian-kajian yang telah diterbitkan dalam bidang pembelajaran mesin dan pendidikan akan dilakukan secara mendalam.



Rajah 3.1. Rangka Kerja Model Pengelompokan.

Fasa pembangunan model pengelompokan pula mengandungi langkah-langkah pra-pemprosesan data dan diikuti dengan pemilihan atribut. Kemudian, analisis deskriptif secara statistik akan dilakukan untuk melihat ciri dan nilai setiap atribut yang telah dipilih. Seterusnya, model pengelompokan akan dibangunkan dan penilaian prestasi model pengelompokan akan dilakukan. Fasa terakhir iaitu output mengandungi proses analisis kelompok menggunakan teknik pengekstrakan ciri dan diikuti dengan pembangunan model pengelasan untuk mengelas pelajar berdasarkan prestasi yang telah dihasilkan. Prestasi model pengelasan ini juga akan dinilai bagi mengenalpasti teknik yang terbaik.

3.3.1 FASA PEMBANGUNAN MODEL PENGELOMPOKAN

a. Data

MyMoheS merupakan satu sistem yang dibangunkan oleh KPT dalam usaha mengumpul data dari semua IPTA merangkumi data berkaitan staf, R&D, pelajar serta maklumat institusi dan disimpan dalam satu gudang data. Tujuannya adalah bagi memudahkan pihak pengurusan KPT membuat capaian berkaitan data Universiti Awam dalam satu tempat seterusnya memudahkan pelaporan kepada pihak berkepentingan. Aktiviti pengumpulan dan penghantaran data dibuat sebanyak 2 kali setahun oleh staf teknikal UA ke server pangkalan data MyMoheS. Manakala aktiviti pembetulan data yang tidak lengkap dan bermasalah dijalankan secara berkala sepanjang tahun bagi menjamin kualiti dan kebolehpercayaan data dalam gudang data MyMoheS (Pekeliling KPT 2016).

Data yang digunakan bagi kajian ini telah dimohon dan diperolehi daripada BPPD, KPT. Pihak UA sentiasa menghantar data terkini pelajar yang mendaftar, mengikuti dan yang menamatkan pengajian di institusi mereka kepada BPPD. Data-data ini terdiri daripada maklumat umum seperti demografi, kursus pengajian, latihan industri, pencapaian akademik, aktiviti, perwakilan pelajar, anugerah dan status pekerjaan bagi graduan IPTA peringkat pengajian ijazah sarjana muda bagi sesi tamat pengajian tahun 2015 hingga 2019. Data mentah yang diterima adalah dalam format *comma separated values* (CSV) dan mengandungi 248,568 data serta sebanyak 53 atribut.

b. Pra-Pemprosesan Data

a. Integrasi Data

Setiap atribut telah dibahagikan kepada 4 dimensi yang merangkumi atribut berkaitan demografi, prestasi, penglibatan tingkah laku dan kerjaya pelajar B40 di IPTA. Set data pelajar merupakan set data utama yang mengandungi maklumat setiap pelajar UA termasuk maklumat demografi, kewarganegaraan, pendaftaran universiti, kursus yang diambil dan status pengajian.

Jadual 3.1. Senarai Set Data Berserta Maklumat.

Dimensi	Tema	Atribut	Jumlah Atribut	Jumlah Data
Demografi	Pelajar	ID pelajar, tarikh lahir, jantina, taraf perkahwinan, negeri lahir, poskod, dun, parlimen, negara asal, kediaman penginapan, kumpulan pendapatan, kelas pendapatan, warganegara, jenis sekolah menengah, universiti, kod institusi, tarikh daftar masuk, sesi daftar masuk, kelayakan masuk, no semester semasa, program, kod program, peringkat pengajian, mod pengajian, tarikh tamat pengajian, sesi tamat pengajian, tarikh berhenti pengajian, tajaan	29	248,568
Prestasi		CGPA Mymohes, status pengajian		
Penglibatan Tingkah Laku	Aktiviti	ID pelajar, aktiviti pelajar, kod aktiviti, jawatan aktiviti, jenis aktiviti, keperluan bergraduat, keusahawanan	7	660,319
	Anugerah	ID pelajar, anugerah, peringkat anugerah, jenis anugerah, tahun anugerah	5	5,945
	Latihan Industri	ID pelajar, keputusan li, poskod li, nama institusi li	4	95,752
	Majlis Perwakilan Pelajar	ID pelajar, sesi mpp, jawatan mpp	3	1,348
Kerjaya	Pekerjaan	ID pelajar, CGPA SKPG, tahun konvo, status pekerjaan, taraf pekerjaan	5	224,174
	Jumlah		53	1,231,864

Set data aktiviti pula mengandungi maklumat aktiviti-aktiviti yang disertai pelajar sepanjang tempoh pengajian merangkumi badan beruniform, kelab dan persatuan, keusahawanan, sukan dan aktiviti sukarelawan. Set data anugerah menyenaraikan anugerah-anugerah yang diterima oleh pelajar di peringkat universiti, negeri, kebangsaan dan antarabangsa berdasarkan kejayaan dalam aktiviti yang disertai. Set data latihan industri mengandungi maklumat penglibatan pelajar dalam latihan industri, latihan amali atau praktikal di institusi-institusi yang terlibat berserta dengan keputusan yang diperolehi. Set data majlis perwakilan pelajar (MPP) menyenaraikan penglibatan pelajar yang dilantik menjadi ahli MPP, sesi lantikan dan jawatan yang disandang. Set data pekerjaan mengandungi maklumat pasca

pengajian seperti tahun konvokesyen, status pekerjaan selepas tamat pengajian dan taraf pekerjaan sekiranya telah bekerja. Jadual 3.1 menyenaraikan enam set data berserta atribut, jumlah atribut dan jumlah data secara keseluruhannya.

Keenam-enam set data yang berbeza seperti yang diterangkan di atas digabungkan menjadi satu set data yang mengandungi 248,568 data yang dihubungkan melalui ID pelajar. Manakala set data aktiviti, anugerah dan MPP mempunyai ID pelajar duplikat kerana seorang pelajar ada yang mempunyai rekod penyertaan dalam beberapa aktiviti, menerima beberapa anugerah serta dilantik menjadi MPP bagi beberapa sesi yang berbeza. Set data ini akan dipra-proses pada bahagian penjanaan atribut bagi memastikan tiada ID duplikat pada set data akhir.

b. Tapisan Data

Atribut seperti warganegara ditapis dengan hanya memilih warganegara Malaysia dan mengeluarkan bukan warganegara serta penduduk tetap. Kemudian atribut negara asal ditapis dengan hanya memilih negara Malaysia manakala atribut julat pendapatan keluarga hanya mengambil kira RM 1 hingga RM 4000 iaitu termasuk dalam kumpulan B40 sahaja. Seterusnya atribut peringkat pengajian hanya mengambil kira pelajar peringkat sarjana muda dan ijazah pertama. Atribut mod pengajian menapis keluar mod fleksibel dan separuh masa meninggalkan hanya pelajar mod sepenuh masa sahaja bagi tujuan kajian. Jadual 3.2 di bawah menyenaraikan atribut-atribut dan nilai tapisan yang dipilih.

Jadual 3.2. Senarai Atribut Dan Nilai Tapisan Yang Dipilih

Bil.	Atribut	Nilai Dipilih
1	Warganegara	Warganegara Malaysia
2	Negara asal	Malaysia
3	Kumpulan pendapatan	RM 1 – RM 4000
4	Peringkat pengajian	Sarjana Muda/Ijazah Pertama
5	Mod pengajian	Sepenuh Masa

c. Pembersihan Data

i. Atribut tiada nilai

Setelah integrasi set-set data dilakukan, semakan statistik data mendapati wujudnya beberapa rekod yang tidak mempunyai nilai (missing values). Jadual 3.3 di bawah menyenaraikan atribut-atribut yang mempunyai data tiada nilai.

Jadual 3.3. Senarai Atribut Yang Mengandungi Data Tiada Nilai.

Bil.	Atribut	Bilangan Data Tiada Nilai
1	Perkahwinan	216
2	Poskod	104
3	Dun	39,425
4	Parlimen	37,486
5	Sekolah menengah	28,820
6	Kelayakan masuk	308
7	Tarikh daftar masuk	1
8	No. semester semasa	39

Atribut dun, parlimen dan poskod dibuang daripada set data kerana bilangan data yang tidak mempunyai nilai atau tidak lengkap adalah terlalu tinggi. Manakala atribut seperti perkahwinan, jenis sekolah menengah, tarikh daftar masuk dan nombor semester semasa yang mempunyai beberapa data tiada nilai atau tidak lengkap telah ditapis keluar. Atribut-atribut yang mempunyai data tidak lengkap seperti kelayakan masuk, tajaan dan CGPA Mymohes juga ditapis keluar. Rekod-rekod ini telah ditapis keluar dan boleh tidak diambil peduli memandangkan bilangannya yang dianggap kecil.

ii. Atribut nilai tunggal

Atribut yang mempunyai satu nilai sahaja akan dikeluarkan daripada data set seperti negara asal iaitu Malaysia, warganegara iaitu semua warganegara Malaysia, kumpulan pendapatan iaitu semua kumpulan B40, peringkat pengajian iaitu hanya pelajar di peringkat sarjana muda dan ijazah pertama dan mod pengajian iaitu hanya pelajar sepenuh masa. Jadual 3.4 menyenaraikan atribut yang mempunyai hanya satu nilai sahaja.

Jadual 3.4. Senarai Atribut Yang Mempunyai Hanya Satu Nilai.

Bil.	Atribut	Nilai tunggal
1	Warganegara	Warganegara Malaysia
2	Negara asal	Malaysia
3	Kelas pendapatan	Kumpulan B40
4	Peringkat pengajian	Sarjana Muda/ Ijazah pertama
5	Mod pengajian	Sepenuh masa

iii. Atribut yang berulang

Atribut yang berulang dan menerangkan nilai yang hampir sama telah dibuang daripada set data. Antara atribut-atribut yang terlibat adalah universiti, program pengajian, tarikh daftar masuk pengajian, tarikh tamat pengajian dan CGPA SKPG. Jadual 3.5 di bawah menyenaraikan atribut-atribut yang berulang.

Jadual 3.5. Senarai Atribut Yang Berulang.

Bil.	Atribut	Atribut Berulang
1	Universiti	Kod institusi
2	Program	Kod program
3	Tarikh daftar masuk	Sesi daftar masuk
4	Tarikh tamat pengajian	Sesi tamat pengajian Tarikh berhenti pengajian Tahun konvo
5	CGPA SKPG	CGPA Mymohes
6	Status pekerjaan	Taraf pekerjaan

iv. Pengebinan

Tujuan pengebinan adalah bagi mengendalikan data hingar, meningkatkan kecekapan pemprosesan data, menghasilkan perwakilan data yang lebih mudah diterangkan dan meningkatkan kefahaman ke atas keputusan perlombongan data kelak. Pengebinan ke atas set data dalam kajian ini diaplikasikan ke atas data numerikal bagi menukarkannya kepada jenis ketegorikal. Atribut CGPA telah dibin mengikut julat kategori ijazah seperti kelas pertama, kelas kedua tinggi, kelas kedua rendah, kelas ketiga dan gagal. Jadual 3.6 di bawah menyenaraikan bin CGPA mengikut julat kategori ijazah..

Jadual 3.6. Senarai Bin Atribut CGPA.

Bil.	Bin	Julat CGPA
1	Bawah 1.99	0 hingga 1.99
2	2.00-2.99	2.00 hingga 2.99
3	3.00-3.49	3.00 hingga 3.49
4	3.50-4.00	3.50 hingga 4.00

Bagi atribut umur daftar dan bil aktiviti pula, pengebinan kepada jenis kategorikal telah dibuat dengan membahagikannya berdasarkan bilangan frekuensi yang sama rata. Jadual 3.7 di bawah menyenaraikan bin umur dan bilangan aktiviti mengikut frekuensi.

Jadual 3.7. Senarai Bin Atribut Umur Dan Bilangan Aktiviti.

Bil.	Atribut	Bin	Frekuensi
1	Umur daftar	Bawah 19 tahun	38,902
		20 tahun	50,039
		Atas 21 tahun	28,128
2	Bilangan aktiviti	Tiada	15,618
		1	27,083
		2	16,991
		3&4	19,040
		5,6,7,8&9	18,806
		Atas 20	19,531

Manakala data kategorikal yang lain juga telah melalui proses pengebinan dan diwakilkan dengan kumpulan yang lebih umum. Atribut-atribut yang terlibat dan perwakilan data adalah seperti di Lampiran A. Perwakilan data bagi atribut kumpulan pendapatan dibahagikan kepada tiga julat had pendapatan berdasarkan pengelasan baharu isi rumah dari Jabatan Perangkaan Negara Malaysia. Manakala atribut universiti dibahagikan kepada tiga kumpulan UA, iaitu Universiti Penyelidikan, Universiti Komprehensif dan Universiti Berfokus (teknikal, pendidikan, pengurusan dan pertahanan). Set data dalam kajian ini terdiri daripada lima Universiti Penyelidikan, empat Universiti Komprehensif dan sebelas Universiti Berfokus. Universiti Penyelidikan memberikan tumpuan kepada bidang penyelidikan, Universiti Komprehensif menawarkan pelbagai kursus dan bidang pengajian manakala Universiti Berfokus pula memberikan tumpuan kepada bidang khusus berkaitan dengan penubuhannya.

Atribut bidang pengajian pula diwakilkan berdasarkan lapan bidang besar di dalam Kod Pendidikan Nasional yang dikeluarkan oleh Kementerian Pendidikan Malaysia (Pengajian Tinggi). Kod Pendidikan Nasional ini menetapkan satu set kod yang standard bagi mengklasifikasi dan mendefinisikan semua program dan latihan dalam bidang pendidikan untuk memenuhi keperluan standard maklumat statistik negara kita. Perwakilan data cicir dalam atribut status pengajian pula mengandungi rekod-rekod pelajar dengan nilai data berhenti dan diberhentikan yang membawa maksud tercicir daripada pengajian.

d. Penjanaan Atribut

Bagi memanfaatkan atribut tarikh yang ada di dalam set data, pengiraan dan penjanaan atribut baru telah dilakukan. Antaranya ialah atribut umur daftar dijana menggunakan atribut tarikh lahir dan tarikh daftar masuk. Manakala atribut tempoh pengajian pula dijana menggunakan tarikh daftar masuk dan tarikh tamat pengajian. Selain daripada itu, atribut bilangan aktiviti dijana dengan mengira bilangan penglibatan aktiviti ko-kurikulum mengikut ID pelajar yang disenaraikan. Teknik ini dapat mengelakkan berlakunya duplikat pada ID Pelajar memandangkan seorang pelajar mempunyai penglibatan dalam beberapa kegiatan ko-kurikulum di dalam set data aktiviti. Jadual 3.8 di bawah menyenaraikan atribut baru yang telah dijana.

Jadual 3.8. Senarai Atribut Yang Telah Dijana.

Bil.	Atribut Dijana	Atribut Asal
1	Umur daftar	Tarikh lahir Tarikh daftar masuk
2	Bil aktiviti	Aktiviti pelajar

e. Transformasi Data

i. Perwakilan jenis data nominal dan ordinal kepada numerikal

Secara umumnya, kebanyakan algoritma pembelajaran mesin memerlukan semua pembolehubah input dan output berada dalam bentuk numerik. Hal ini merupakan satu kekangan yang perlu dihadapi bagi pelaksanaan dan pembentukan model pembelajaran mesin. Ini bermakna, semua atribut yang mengandungi pembolehubah kategorikal atau nominal perlu

menjalani proses pemetaan atau pengekodan kepada nombor sebelum dimuatkan ke dalam model pengelompokan dan pengelasan. Bagi kajian ini, pengekodan integer terhadap set data pelajar akan dilakukan di mana setiap nilai kategori yang unik akan ditukar kepada nilai integernya yang tersendiri. Jadual 3.9 di bawah menyenaraikan kesemua 16 atribut berserta jenis data.

Jadual 3.9. Senarai Atribut Dengan Perwakilan Dan Jenis Data.

Bil.	Atribut	Nilai	Perwakilan	Jenis Data
1	Jantina	Lelaki	1	Nominal
		Perempuan	2	
2	Umur Daftar	Bawah 19 tahun	1	Nominal
		20	2	
		Atas 21 tahun	3	
3	Perkahwinan	Bujang	1	Nominal
		Berkahwin	2	
4	Negeri Lahir	Utara	1	Nominal
		Tengah	2	
		Selatan	3	
		Pantai Timur	4	
		Malaysia Timur dan lain-lain	5	
5	Kumpulan Pendapatan	Bawah RM2000	1	Ordinal
		RM2001-3000	2	
		RM3001-4000	3	
6	Sekolah Menengah	SMK	1	Nominal
		SBP dan MRSM	2	
		SM Agama	3	
		SM Teknik	4	
		Lain-lain	5	
7	Kelayakan	SPM, STPM dan lain-lain	1	Nominal
		Matrikulasi dan Asasi	2	
		Diploma	3	
8	Tajaan	Biasiswa	1	Nominal
		Persendirian	2	
		Pinjaman	3	
9	Kediaman	Residen	1	Nominal
		Bukan residen	2	

bersambung...

... sambungan

10	Universiti	Universiti Penyelidikan	1	Nominal
		Universiti Komprehensif	2	
		Universiti Berfokus	3	
11	Bidang Pengajian	Pendidikan	1	Nominal
		Sastera dan Kemanusiaan	2	
		Sains Sosial, Perniagaan dan Perundangan	3	
		Sains, Matematik dan Komputer	4	
		Kejuruteraan, Pembuatan dan Pembinaan	5	
		Pertanian dan Veterinar	6	
		Kesihatan dan Kebajikan	7	
		Perkhidmatan	8	
12	CGPA	Bawah 2.00	1	Ordinal
		2.01 – 2.99	2	
		3.00 – 3.49	3	
		3.50 – 4.00	4	
13	Keputusan Latihan Industri	Tiada	0	Nominal
		Lulus	1	
		Gagal	2	
14	Bil Aktiviti	Tiada aktiviti	0	Ordinal
		1	1	
		2	2	
		3 dan 4	3	
		5 hingga 9	4	
		10 ke atas	5	
15	Status Pekerjaan	Bekerja	1	Nominal
		Belum bekerja	2	
		Melanjutkan pengajian	3	
16	Status Pengajian	Cicir	0	Nominal
		Tamat pengajian	1	

ii. Normalisasi

Seterusnya, data di atas akan melalui dua proses normalisasi menggunakan kaedah *StandardScaler* dan *MinMaxScaler* bagi membentuk set data bagi model yang berbeza. Kaedah yang pertama, *StandardScaler* atau dikenali sebagai *z-transformation* membolehkan nilai-nilai setiap atribut berada dalam julat yang sesuai dan boleh dibuat perbandingan. Formula pengiraan *StandardScaler* diberikan pada persamaan 3.1 di bawah di mana hasil perbezaan

nilai sampel *x* dengan nilai min sampel dibahagikan dengan nilai sisihan piawai sampel. Hasilnya data yang ditransformasi atau skor-z akan membentuk taburan sekitar nilai min '0' dengan sisihan piawai '1'.

$$z = \frac{x - min(x)}{sisihan\ piawai(x)} \tag{3.1}$$

Manakala Kaedah kedua, *MinMaxScaler* akan menukarkan semua atribut ke dalam julat [0,1] yang bermaksud nilai minimum atribut ialah kosong dan nilai maksimum atribut pula ialah satu. Formula matematik bagi *MinMaxScaler* adalah didefinisikan seperti berikut (3.2):

$$x = \frac{x - min(x)}{max(x) - min\,(x)} \tag{3.2}$$

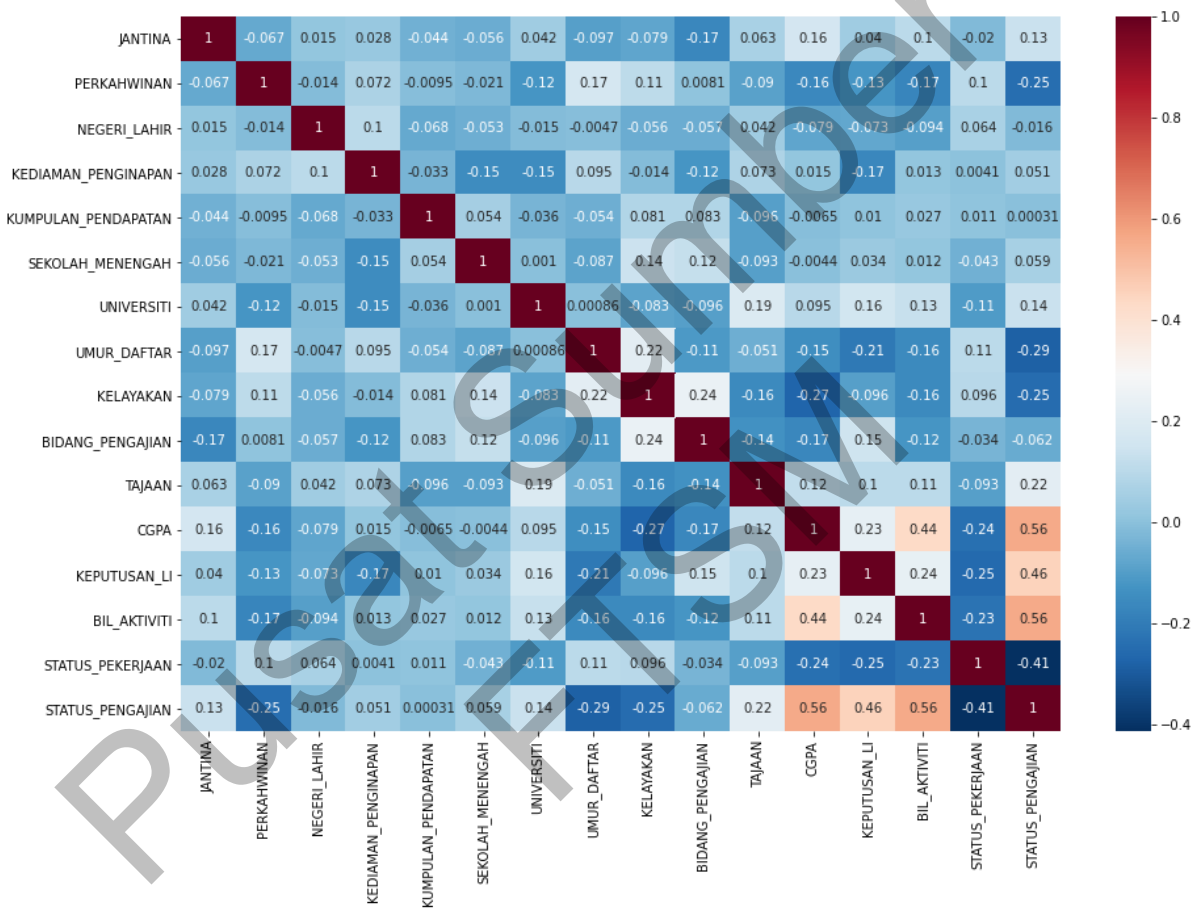
f. Pemilihan Atribut

Terdapat terlalu banyak atribut dalam set data pendidikan yang boleh menyebabkan masalah dan kesukaran ketika memproses dan menganalisa data (curse of *dimensionality*) jika semuanya digunakan dalam teknik pengelompokan. Ini kerana pengiraan jarak oleh algoritma untuk mengukur kemiripan antara dua sampel mungkin tidak berkesan bagi data berdimensi tinggi. Bagi menyelesaikan masalah ini, teknik pemilihan atribut akan diambil bagi menganalisis dan mencari atribut-atribut yang terbaik sahaja untuk dibuat kajian.

i. Ujian Korelasi Spearman - Berselia

Sebelum membangunkan model pengelompokan, satu pengukuran statistik menggunakan korelasi *Spearman* dijalankan bagi menentukan hubungan antara atribut. Sesuatu ciri atribut berkemungkinan bertindih antara satu sama lain sekiranya ia diterbitkan berdasarkan daripada ciri-ciri atribut yang lain (X. Li et al., 2021). Pemilihan korelasi *Spearman* dalam kajian ini adalah berdasarkan kesesuaiannya terhadap set data pelajar yang mengandungi nilai yang diskret (Hussain et al., 2018). Setiap atribut yang dikaji akan menerima satu nilai pekali korelasi (*r*) yang mewakili kekuatan dan arah bagi hubungan linear antara pasangan atribut. Kekuatan bagi setiap hubungan boleh dikategorikan sebagai sangat lemah (0.00-0.19), lemah (0.2-0.39), sederhana (0.4-0.59), kuat (0.6-0.79) dan sangat kuat (0.80-1.00) (Gough et al., 2021).

Rajah 3.2 di bawah menunjukkan tiga atribut mencatatkan korelasi dengan status pengajian melebihi 0.45 iaitu CGPA, keputusan LI dan bilangan aktiviti. Atribut-atribut lain yang menunjukkan korelasi positif yang rendah ialah jantina, kediaman penginapan, sekolah menengah, universiti dan tajaan. Manakala atribut-atribut yang mencatatkan korelasi negatif dengan status pengajian ialah perkahwinan, negeri lahir, kumpulan pendapatan, umur daftar, kelayakan, bidang pengajian dan status pekerjaan.



Rajah 3.2. Plot *Heatmap* Korelasi Antara Semua Atribut.

Setelah melakukan analisis korelasi *Spearman*, setiap pemboleh ubah tidak bersandar dalam set data akan menerima pekali korelasi (*r*) dan keputusan analisis korelasi ditunjukkan dalam Jadual 3.10 di bawah. Keputusan ujian statistik menunjukkan lima belas atribut adalah signifikan dan mempunyai korelasi dengan status pengajian pelajar di mana nilai-*P* adalah kurang daripada 0.05. Manakala hanya atribut kumpulan pendapatan yang menunjukkan tiada perkaitan dengan status pengajian pelajar. Walaupun begitu, kesemua atribut akan digunakan

dalam analisis seterusnya kerana mungkin masih berguna dalam aspek-aspek yang lain (Luan & Zhao, 2006).

Jadual 3.10. Analisis Korelasi Bagi Setiap Atribut Dengan Status Pengajian.

Bil.	Atribut	Nilai *r*	Nilai *p*
1	CGPA	0.5633	<0.01
2	Bil aktiviti	0.5593	<0.01
3	Keputusan LI	0.4573	<0.01
4	Tajaan	0.2171	<0.01
5	Universiti	0.1384	<0.01
6	Jantina	0.1345	<0.01
7	Sekolah menengah	0.0592	<0.01
8	Kediaman penginapan	0.0506	<0.01
9	Kumpulan pendapatan	0.0003	0.9159
10	Negeri lahir	-0.0156	<0.01
11	Bidang pengajian	-0.0621	<0.01
12	Perkahwinan	-0.2458	<0.01
13	Kelayakan	-0.2488	<0.01
16	Umur daftar	-0.2855	<0.01
15	Status pekerjaan	-0.4129	<0.01

ii. Statistik Kendall's W - Berselia

Seterusnya, atribut-atribut dalam set data akan melalui proses pemilihan atribut dengan menggunakan teknik hutan rawak (HR), pohon ekstra (PE), *Info Gain* (IG) dan *chi square.*

Statistik *Kendall's W* adalah statistik *non-parametric* yang digunakan untuk menilai persetujuan antara beberapa penilai (*rater*) yang berbeza. Nilai statistik *Kendall's W* adalah sentiasa antara 0 dan 1. Jika nilai kosong, tiada persetujuan antara penilai, manakala nilai satu menunjukkan persetujuan yang sangat baik. Formula matematik bagi *Kendall's W* adalah didefinisikan seperti berikut (3.3):

$$W = \frac{12S}{m^2(n^3-n)} \tag{3.3}$$

di mana *S* ialah hasil tambah *deviations* kuasa dua, *m* ialah bilangan penilai dan *n* ialah jumlah bilangan objek yang dinilai.

Setelah fasa pemilihan atribut dijalankan, atribut-atribut telah diberi nilai pemberat mengikut kepentingan. Atribut yang mempunyai nilai pemberat boleh disusun mengikut kedudukan berdasarkan penilai masing-masing seperti yang ditunjukkan dalam Jadual 3.11 di bawah. Atribut-atibut yang dipilih hanya pada kedudukan 9 teratas termasuk kelas label status pengajian manakala atribut 6 terbawah kesemuanya adalah atribut jenis demografi dan didapati tidak signifikan.

Jadual 3.11. Atribut Yang Mempengaruhi Pencapaian Pelajar Mengikut Kedudukan.

Atribut	HR	PE	IG	Chi2	Nilai Purata	Kedudukan
Bil aktiviti	0	1	0	0	0.25	1
CGPA	1	0	1	1	0.75	2
Status pekerjaan	2	2	2	3	2.25	3
Keputusan LI	3	3	3	2	2.75	4
Umur daftar	5	4	4	4	4.25	5
Kelayakan	4	5	5	5	4.75	6
Tajaan	8	8	6	6	7	7
Universiti	9	7	8	7	7.75	8
Bidang pengajian	6	6	11	9	8	9
Perkahwinan	13	11	7	11	10.5	10
Negeri lahir	7	9	14	13	10.75	11
Kediaman penginapan	10	10	12	12	11	12
Jantina	14	14	9	8	11.25	13
Sekolah menengah	12	13	10	10	11.25	13
Kumpulan pendapatan	11	12	13	14	12.5	14

Keputusan ujian statistik *Kendall's W* ditunjukkan di dalam Jadual 3.12 di bawah. Ujian statistik *Kendall's W* menghasilkan nilai 0.8862 dan menunjukkan persetujuan yang baik dan kuat antara semua penilai yang digunakan. Nilai *p* yang rendah di bawah tahap signifikan 0.05 memberikan bukti yang kuat untuk menolak hipotesis *null*.

H_0: Kedudukan yang dihasilkan oleh penilai-penilai adalah tidak sejajar dan bersetuju antara satu sama lain.

H_1: Terdapat sekurang-kurangnya satu penilai yang sejajar dan bersetuju dengan satu penilai lain, atau dengan beberapa penilai lain.

Jadual 3.12. Nilai Statistik Kendall's W.

Statistik Kendall's W	
Kendall W	0.8862
ChiSq	49.625
df	14
p	<0.0001

Di akhir proses pemilihan atribut secara berselia, hanya 10 atribut sahaja yang tinggal dan akan digunakan pada fasa yang seterusnya.

iii. Nilai ambang varians – Tanpa Selia

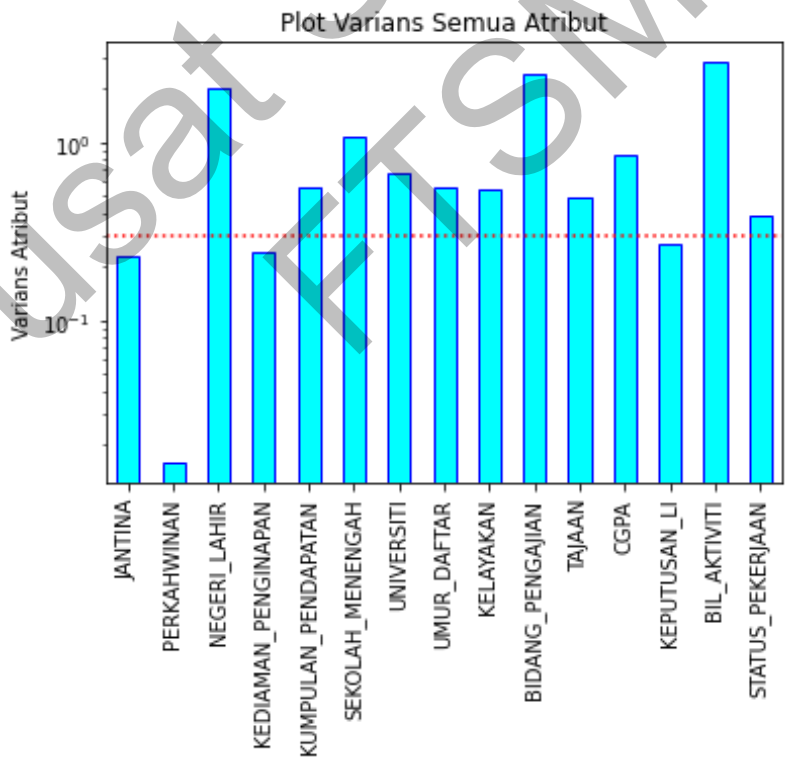
Pemilihan atribut tanpa selia berdasarkan pendekatan tapisan terbahagi kepada dua iaitu *univariate* dan *multivariate*. Pendekatan ini akan menggunakan beberapa kriteria *intrinsic* bagi menilai setiap atribut dan atribut-atribut ini kemudiannya akan diletakkan dalam sebuah senarai mengikut kedudukan (*ranking*). Kaedah ini membolehkan kita mengenalpasti atribut tidak relevan dan mengeluarkannya daripada senarai secara efektif. Tetapi kelemahannya ialah ia tidak berupaya mengenalpasti atribut bertindih kerana tidak mengambilkira kebergantungan antara atribut semasa proses tapisan dijalankan. Kaedah tapisan ini adalah sangat mudah, tidak memerlukan algoritma pengelompokan bagi mencari atribut yang relevan dan sangat laju untuk diimplementasikan.

Varians adalah satu pengukuran bagi melihat sejauh mana penyelerakan titik-titik pada taburan data sesuatu set data. Model pengelompokan yang akan dihasilkan memerlukan atribut yang mempunyai variasi agar tidak menghasilkan model yang berat sebelah dan condong terhadap sesuatu atribut. Oleh sebab itu pemilihan atribut berdasarkan varians adalah penting sebelum membina model pembelajaran mesin tanpa selia.

Atribut dengan varians yang tinggi bermaksud nilai yang dipunyai oleh atribut adalah berlainan atau mempunyai kekardinalan (*cardinality*) yang tinggi. Atribut dengan varians yang rendah mempunyai nilai yang hampir serupa manakala varians kosong menunjukkan atribut

dengan nilai yang sama. Selain itu, atribut dengan varians yang rendah cenderung untuk berada dekat dengan nilai min, sekaligus memberikan maklumat pengelompokan yang minima (X. Li et al., 2021). Teknik nilai ambang varians sesuai untuk pemodelan tanpa selia kerana hanya memeriksa atribut input (X) tanpa mengambilkira maklumat daripada atribut bersandar (y).

Varians semua atribut pelajar ditunjukkan dalam plot Rajah 3.3 dibawah. Nilai ambang dalam kajian ini telah ditetapkan pada 0.3 bagi pemilihan atribut. Berdasarkan pemerhatian daripada rajah, terdapat sepuluh atribut dengan nilai varians melebihi 0.3 iaitu negeri lahir, kumpulan pendapatan, sekolah menengah, universiti, umur daftar, kelayakan masuk, bidang pengajian, tajaan, CGPA, bil aktiviti dan status pekerjaan. Manakala terdapat empat atribut dengan nilai varians kurang daripada 0.3 iaitu jantina, perkahwinan, kediaman penginapan dan keputusan li. Atribut negeri lahir, bidang pengajian dan bil aktiviti menunjukkan nilai varians tertinggi dan boleh digunakan untuk menunjukkan pola atau corak prestasi pelajar. Selain itu, atribut jantina, perkahwinan dan kediaman penginapan mencatatkan varians terendah kerana ketiga-tiga atribut hanya mempunyai dua nilai bagi menerangkan demografi pelajar.



Rajah 3.3. Plot Varians Bagi Semua Atribut Pelajar.

Di akhir proses pemilihan atribut secara tanpa selia, hanya 12 atribut sahaja yang tinggal dan akan digunakan pada fasa yang seterusnya.

iv. Set Data Akhir

Jadual 3.13, 3.14 dan 3.15 menunjukkan senarai set atribut yang dipilih selepas proses pemilihan atribut tanpa selia dan berselia menggunakan dua kaedah normalisasi *StandardScaler* dan *MinMaxScaler*. Set-set atribut ini dinamakan sebagai Model A, Model B dan Model C yang kemudiannya akan menjadi input kepada model pengelompokan di peringkat seterusnya.

Jadual 3.13. Model A (*StandardScaler* dan pemilihan atribut berselia dengan 10 atribut).

Bil.	Atribut	Bil	Atribut
1.	Bil aktiviti	6.	Kelayakan
2.	CGPA	7.	Tajaan
3.	Status pekerjaan	8.	Universiti
4.	Keputusan LI	9.	Bidang pengajian
5.	Umur daftar	10.	Status pengajian

Jadual 3.14. Model B (*MinMaxScaler* dan pemilihan atribut berselia dengan 10 atribut).

Bil.	Atribut	Bil	Atribut
1.	Bil aktiviti	6.	Kelayakan
2.	CGPA	7.	Tajaan
3.	Status pekerjaan	8.	Universiti
4.	Keputusan LI	9.	Bidang pengajian
5.	Umur daftar	10.	Status pengajian

Jadual 3.15. Model C (*MinMaxScaler* dan pemilihan atribut tanpa selia dengan 12 atribut).

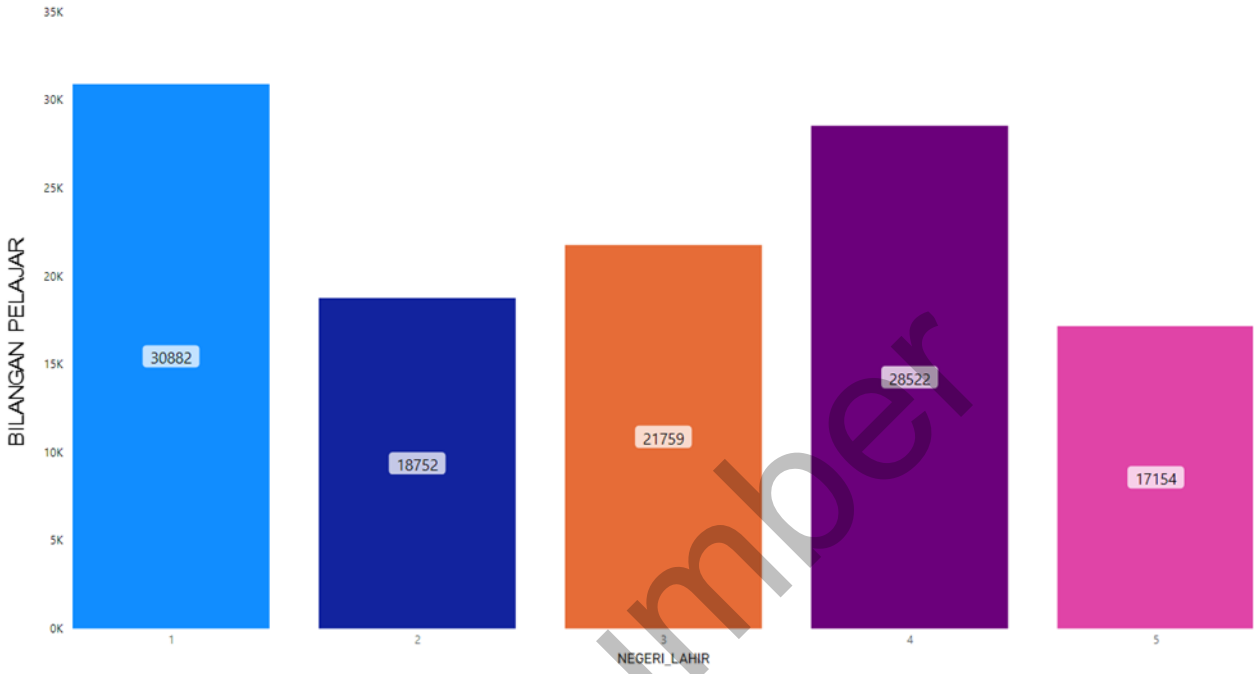
Bil.	Atribut	Bil	Atribut
1.	Negeri lahir	7.	Umur daftar
2.	Kumpulan pendapatan	8.	Kelayakan
3.	Sekolah menengah	9.	Tajaan
4.	Bil aktiviti	10.	Universiti
5.	CGPA	11.	Bidang pengajian
6.	Status pekerjaan	12.	Status pengajian

Model A ialah set data yang dinormalisasi dengan *StandardScaler* dan mempunyai 10 atribut selepas pemilihan atribut secara berselia. Model B pula ialah set data yang dinormalisasi dengan *MinMaxScaler* dan mempunyai 10 atribut selepas pemilihan atribut secara berselia. Manakala Model C ialah set data yang dinormalisasi dengan *MinMaxScaler* dan mempunyai 12 atribut selepas pemilihan atribut tanpa selia.

c. Analisis Deskriptif

Setelah proses pra-pemprosesan dan pemilihan atribut telah selesai, data-data tersebut seterusnya dianalisis secara visual bagi memeriksa taburan nilainya. Terdapat sebanyak sebanyak 117,069 rekod data pelajar bagi kumpulan B40 yang akan menjadi input kepada model pengelompokan di fasa seterusnya. Rajah-rajah graf dibawah telah dihasilkan menggunakan perisian Microsoft Power BI.

a. Negeri Lahir



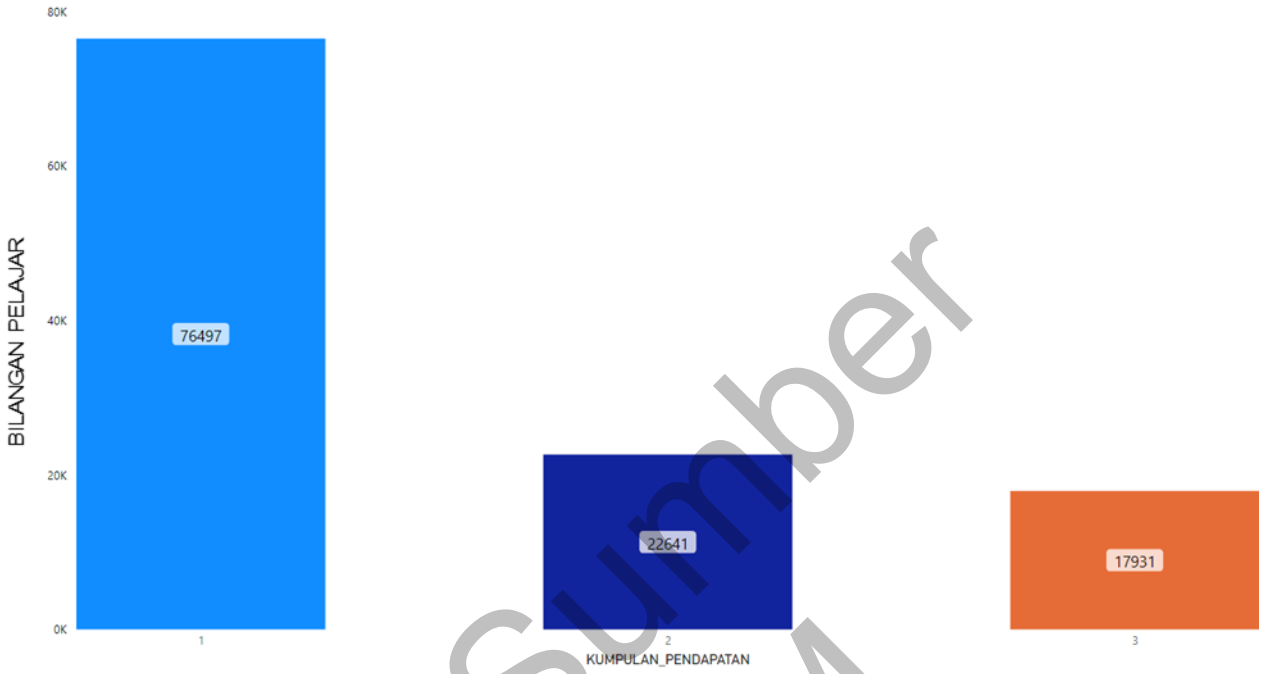
Rajah 3.4. Statistik Pelajar B40 Mengikut Negeri Lahir.

Jadual 3.16. Jadual Statistik Pelajar B40 Mengikut Negeri Lahir.

Numerik	Negeri Lahir	Bilangan	Peratus
1	Utara	30882	26.38
2	Tengah	18752	16.02
3	Selatan	21759	18.59
4	Pantai Timur	28522	24.36
5	Malaysia Timur dan lain-lain	17154	14.65

Rajah 3.4 menunjukkan taburan negeri kelahiran pelajar B40 yang telah dibahagikan kepada lima kumpulan. Peratusan tertinggi pelajar berasal dari negeri bahagian utara (Perlis, Kedah, P. Pinang, Perak) sebanyak 26.38% manakala peratusan terkecil pelajar berasal dari Sabah, Sarawak dan lain-lain negara kelahiran sebanyak 14.65%.

b. Kumpulan Pendapatan



Rajah 3.5. Statistik Pelajar B40 Mengikut Kumpulan Pendapatan.

Jadual 3.17. Jadual Statistik Pelajar B40 Mengikut Kumpulan Pendapatan.

Numerik	Kumpulan Pendapatan	Bilangan	Peratus
1	Kurang RM2000	76497	65.34
2	RM2001-3000	22641	19.34
3	RM3001-4000	17931	15.32

Rajah 3.5 menunjukkan majoriti pelajar B40 tergolong dalam kumpulan pendapatan isi rumah paling rendah iaitu kurang daripada RM2000 sebulan. Mereka ini adalah seramai 76,497 orang pelajar, diikuti dengan kumpulan pendapatan RM2001-RM3000 seramai 22,641 orang dan kumpulan pendapatan RM3001-RM4000 seramai 17,931 orang pelajar.

c. Sekolah Menengah

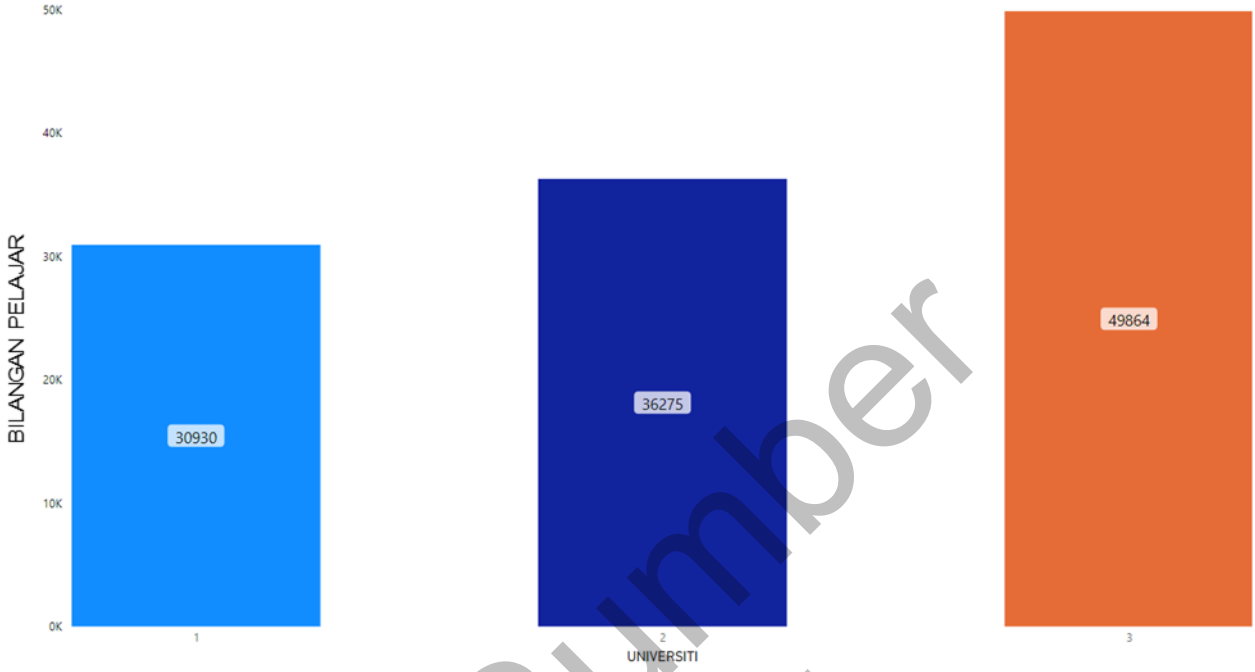
Rajah 3.6. Statistik Pelajar B40 Mengikut Peringkat Pendidikan Sekolah Menengah.

Jadual 3.18. Jadual Statistik Pelajar B40 Mengikut Peringkat Pendidikan Sekolah Menengah.

Numerik	Sekolah Menengah	Bilangan	Peratus
1	SMK	96304	82.26
2	SBP dan MRSM	6876	5.87
3	SM Agama	4443	3.80
4	SM Teknik	4234	3.62
5	Lain-lain	5212	4.45

Statistik pada rajah 3.6 menunjukkan seramai 82.26% pelajar B40 mendapat pendidikan menengah di sekolah menengah kebangsaan iaitu sekolah harian milik Kerajaan Malaysia. Bilangan ini jauh meninggalkan para pelajar yang mendapat pendidikan di sekolah menengah berasrama SBP dan MRSM (5.87%), sekolah menengah agama (3.80%), sekolah menengah teknik (3.62%) dan lain-lain jenis sekolah (4.45%).

d. Universiti



Rajah 3.7. Statistik Pelajar B40 Mengikut Universiti.

Jadual 3.19. Jadual Statistik Pelajar B40 Mengikut Universiti.

Numerik	Universiti	Bilangan	Peratus
1	Penyelidikan	30930	26.42
2	Komprehensif	36275	30.99
3	Berfokus	49864	42.59

Rajah 3.7 menunjukkan statistik pelajar B40 mengikut universiti pengajian dimana bilangan paling ramai adalah daripada universiti berfokus seramai 42.59% diikuti penuntut daripada universiti komprehensif seramai 30.99% dan yang terakhir penuntut daripada universiti penyelidikan seramai 26.42%. Taburan pelajar adalah lebih tinggi di universiti berfokus kerana bilangan universiti di dalam kumpulan tersebut juga adalah tinggi iaitu sebanyak sebelas buah universiti berbanding kumpulan universiti komprehensif dan penyelidikan iaitu masing-masing sebanyak lima dan empat buah universiti sahaja.

e. Umur Daftar

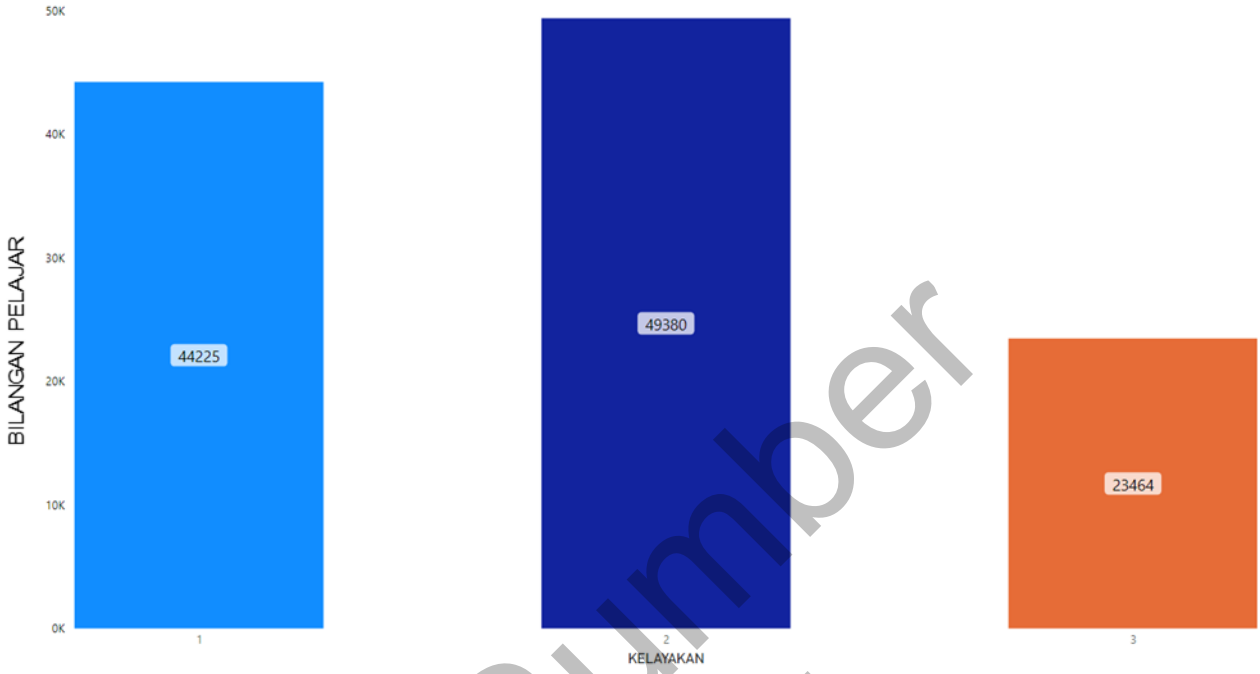
Rajah 3.8. Statistik Pelajar B40 Mengikut Umur Daftar.

Jadual 3.20. Jadual Statistik Pelajar B40 Mengikut Umur Daftar.

Numerik	Umur Daftar	Bilangan	Peratus
1	19 tahun ke bawah	38902	33.23
2	20 tahun	50039	42.74
3	21 tahun ke atas	28128	24.03

Umur daftar merujuk kepada umur pelajar ketika mendaftar memasuki program ijazah sarjana muda di universiti. Atribut ini dikira berdasarkan perbezaan antara tarikh daftar masuk dan tarikh lahir seseorang pelajar itu. Umur daftar dibahagikan kepada tiga kumpulan iaitu pelajar berumur 19 tahun ke bawah, 20 tahun dan 21 tahun ke atas. Kumpulan pelajar paling ramai adalah yang yang berumur 20 tahun sebanyak 42.74%, diikuti kumpulan umur daftar 19 tahun ke bawah sebanyak 33.23% dan yang terakhir kumpulan umur daftar 21 tahun ke atas sebanyak 24.03%.

f. Kelayakan Masuk



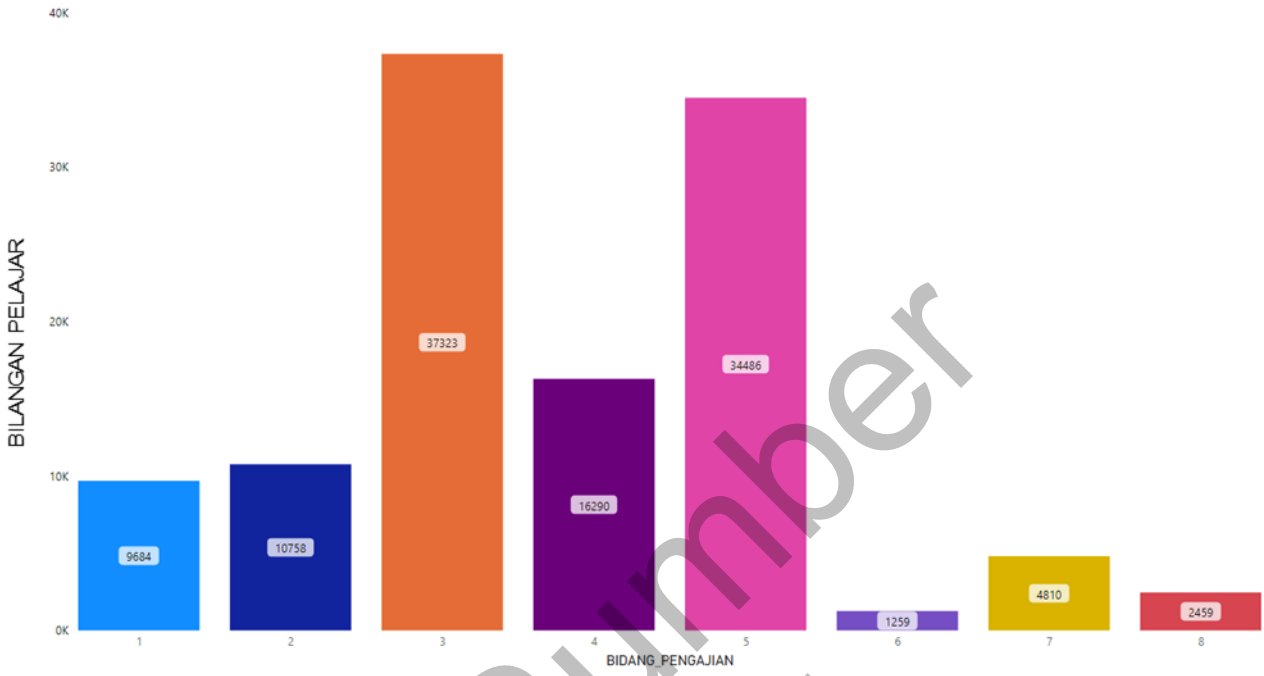
Rajah 3.9. Statistik Pelajar B40 Mengikut Kelayakan

Jadual 3.21. Jadual Statistik Pelajar B40 Mengikut Kelayakan.

Numerik	Kelayakan Masuk Universiti	Bilangan	Peratus
1	SPM, STPM dan lain-lain	44225	37.78
2	Matrikulasi dan Asasi	49380	42.18
3	Diploma	23464	20.04

Rajah 3.9 menunjukkan kelayakan masuk pelajar B40 ke universiti yang terdiri daripada tiga jenis kumpulan kelayakan. Seramai 42.18% pelajar mendaftar ke universiti menggunakan kelulusan lepasan program matrikulasi dan program asasi kelolaan IPT. Manakala seramai 37.78% pelajar menggunakan keputusan SPM, STPM dan lain-lain kelayakan untuk mendaftar ke universiti tempat pengajian mereka. Kumpulan paling kecil yang telah mendaftar masuk ke universiti pula menggunakan kelayakan diploma daripada pelbagai aliran iaitu seramai 20.04%.

g. Bidang Pengajian



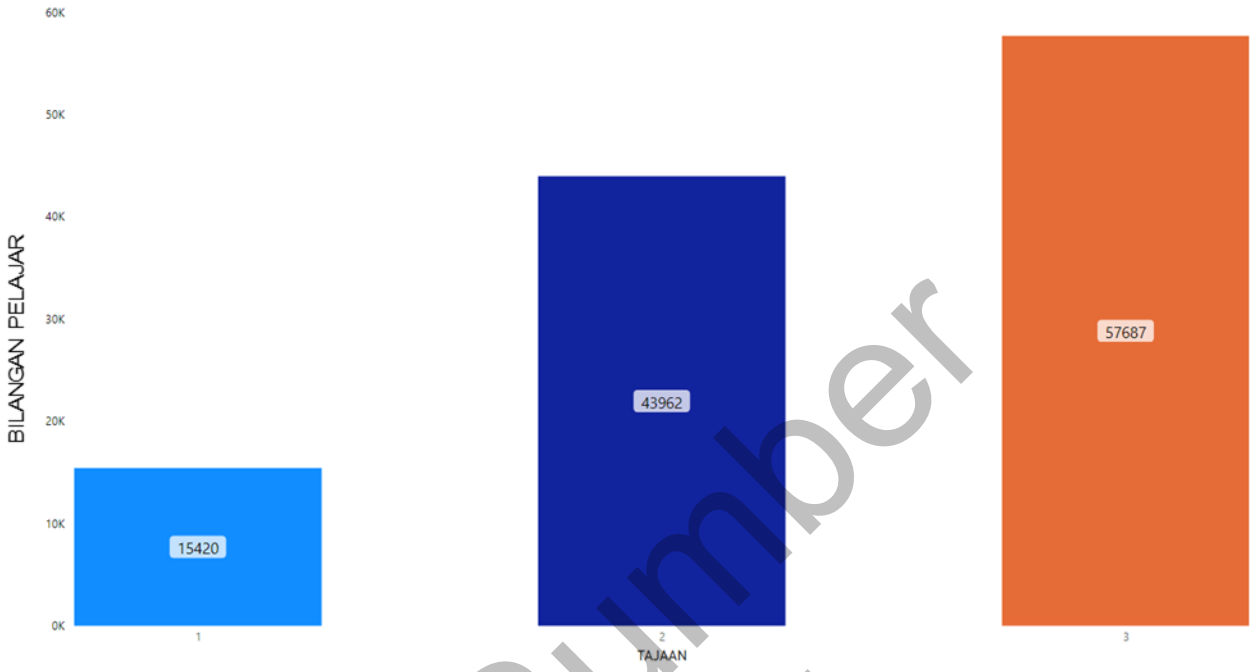
Rajah 3.10. Statistik Pelajar B40 Mengikut Bidang Pengajian.

Jadual 3.22. Jadual Statistik Pelajar B40 Mengikut Bidang Pengajian.

Numerik	Bidang Pengajian	Bilangan	Peratus
1	Pendidikan	9684	8.27
2	Sastera dan kemanusiaan	10758	9.19
3	Sains sosial, perniagaan dan perundangan	37323	31.88
4	Sains, matematik dan komputer	16290	13.91
5	Kejuruteraan, pembuatan dan pembinaan	34486	29.46
6	pertanian dan vaterinar	1259	1.08
7	Kesihatan dan kebajikan	4810	4.11
8	Perkhidmatan	2459	2.10

Mengikut statistik daripada rajah 3.10, dua kumpulan terbesar pelajar B40 telah mengikuti bidang pengajian sains sosial, perniagaan dan perundangan (31.88%) serta kejuruteraan, pembuatan dan pembinaan (29.46%). Manakala sebilangan kecil pelajar telah mengikuti bidang pengajian kesihatan dan kebajikan, perkhidmatan dan pertanian dan vaterinar.

h. Tajaan



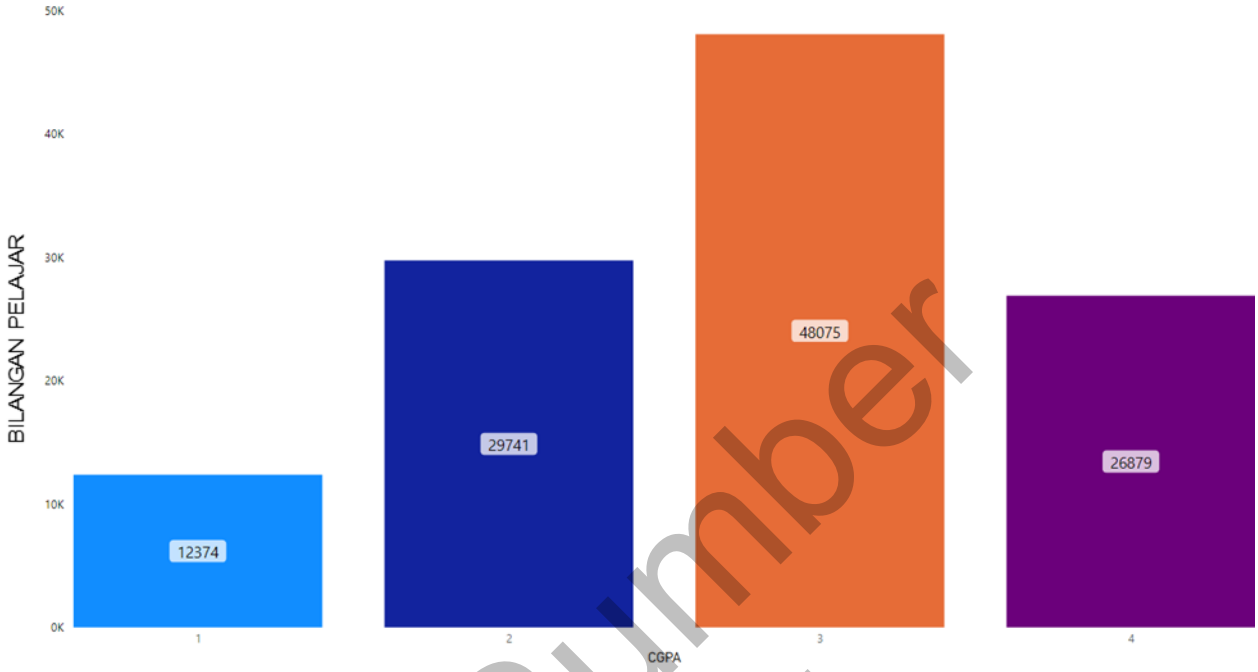
Rajah 3.11. Statistik Pelajar B40 Mengikut Tajaan.

Jadual 3.23. Jadual Statistik Pelajar B40 Mengikut Tajaan.

Numerik	Tajaan	Bilangan	Peratus
1	Biasiswa	15420	13.17
2	Persendirian	43962	37.55
3	Pinjaman	57687	49.28

Rajah 3.11 menunjukkan majoriti pelajar B40 membiayai pengajian mereka menggunakan pinjaman melalui Perbadanan Tabung Pendidikan Tinggi Negara (PTPTN) iaitu seramai 49.28%. Manakala 37.55% pelajar membiayaai sendiri pengajian mereka dan selebihnya hanya 13.17% pelajar yang berjaya mendapatkan tajaan biasiswa bagi menampung kos pengajian di IPT.

i. CGPA



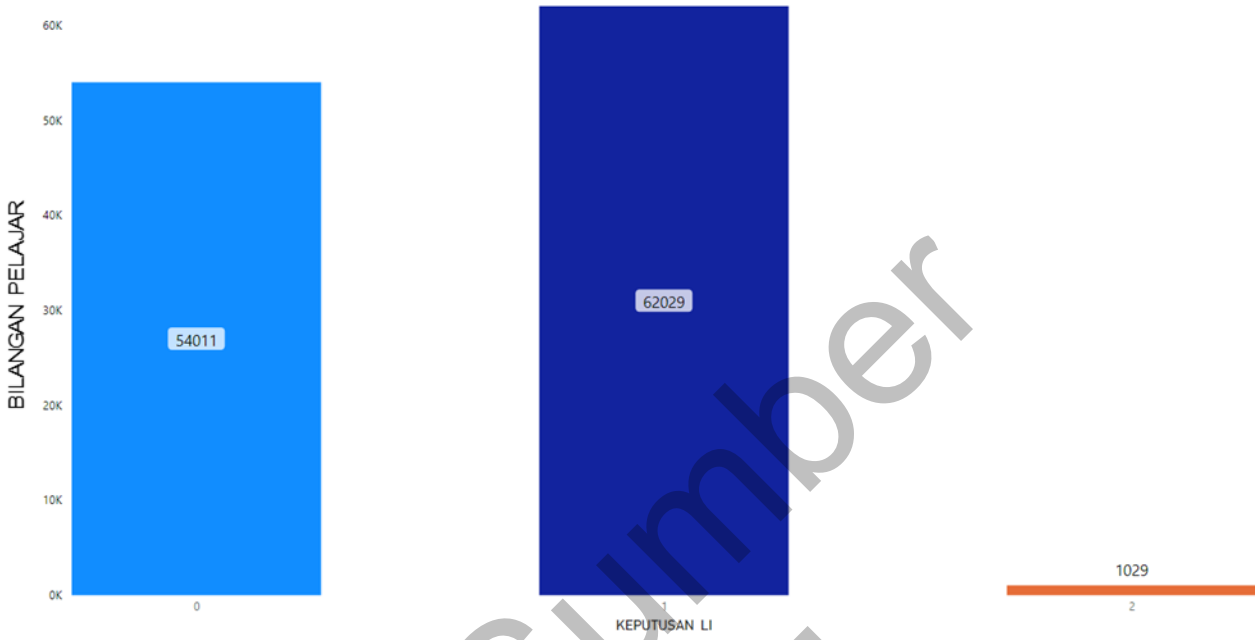
Rajah 3.12. Statistik Pelajar B40 Mengikut CGPA.

Jadual 3.24. Jadual Statistik Pelajar B40 Mengikut CGPA.

Numerik	CGPA	Bilangan	Peratus
1	Bawah 2.00	12374	10.57
2	2.01 – 2.99	29741	25.40
3	3.00 – 3.49	48075	41.07
4	3.50 – 4.00	26879	22.96

Pencapaian akademik pelajar di universiti diukur menggunakan purata nilai gred kumulatif (PNGK) atau CGPA bagi setiap semester sepanjang tempoh pengajian yang diikuti oleh pelajar. Pelajar yang mendapat CGPA kelas kedua tinggi (3.00-3.49) di akhir pengajian adalah majoriti dalam kajian ini seramai 48,075 orang pelajar bersamaan 41.07%. Kemudian diikuti dengan pelajar yang memperoleh CGPA kelas kedua dan ketiga (2.00-2.99) seramai 29,741 (25.40%) dan CGPA kelas pertama (3.50-4.00) seramai 26,879 (22.96%). Peratusan terkecil pelajar di dalam kajian ini telah mendapat PNGK gagal (bawah 1.99) di akhir pengajian mereka iaitu seramai 12,374 orang bersamaan 10.57%.

j. Keputusan Latihan Industri



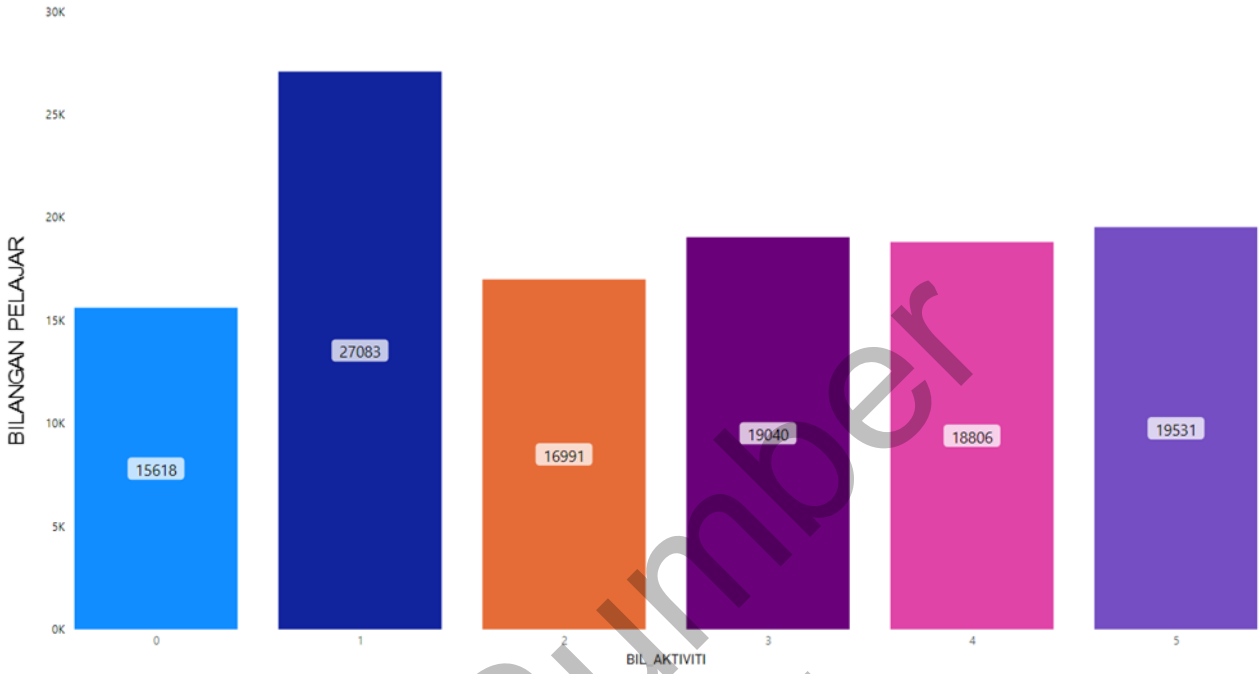
Rajah 3.13. Statistik Pelajar B40 Mengikut Keputusan Latihan Industri.

Jadual 3.25. Jadual Statistik Pelajar B40 Mengikut Keputusan Latihan Industri.

Numerik	Keputusan LI	Bilangan	Peratus
0	Tiada	54011	46.14
1	Lulus	62029	52.98
2	Gagal	1029	0.88

Rajah 3.13 menunjukkan statistik keputusan latihan industri dalam kalangan pelajar B40 di IPT Awam. Majoriti pelajar B40 telah mendapat keputusan lulus iaitu seramai 52.98% berbanding hanya 0.88% yang mencatatkan keputusan gagal. Pelajar yang tiada keputusan adalah sangat tinggi iaitu seramai 54,011 bersamaan 46.14% kerana tidak semua bidang pengajian di IPT menyenaraikan latihan industri sebagai keperluan wajib untuk bergraduat sebagai contoh bidang pendidikan, perubatan dan perundangan yang mempunyai latihannya yang tersendiri.

k. Bilangan Aktiviti



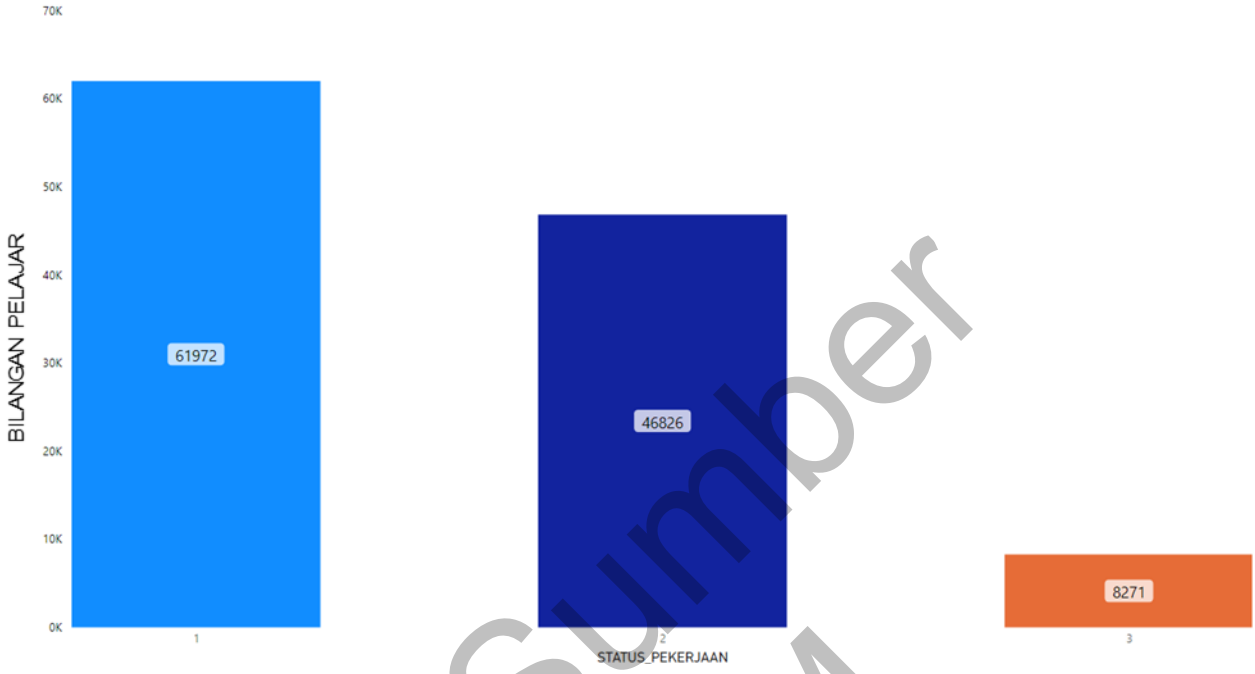
Rajah 3.14. Statistik Pelajar B40 Mengikut Bilangan Aktiviti.

Jadual 3.26. Jadual Statistik Pelajar B40 Mengikut Bilangan Aktiviti.

Numerik	Bilangan Aktiviti	Bilangan	Peratus
0	Tiada	15618	13.34
1	1	27083	23.13
2	2	16991	14.51
3	3 dan 4	19040	16.26
4	5 hingga 9	18806	16.06
5	10 ke atas	19531	16.68

Rajah 3.14 menunjukkan taburan bilangan aktiviti pelajar B40 dalam aktiviti ko-kurikulum dan majoriti pelajar terlibat dengan sekurang-kurangnya satu aktiviti iaitu seramai 27,083 orang (23.13%). Manakala frekuensi pelajar yang terlibat dengan 2, 3 dan 4, 5 hingga 9 dan 10 ke atas aktiviti adalah hampir sama. Pelajar yang tidak tidak terlibat dengan sebarang aktiviti pula mencatatkan bilangan paling rendah iaitu seramai 15,618 (13.34%).

l. Status Pekerjaan



Rajah 3.15. Statistik Pelajar B40 Mengikut Status Pekerjaan.

Jadual 3.27. Jadual Statistik Pelajar B40 Mengikut Status Pekerjaan.

Numerik	Status Pekerjaan	Bilangan	Peratus
1	Bekerja	61972	52.94
2	Tidak bekerja	46826	40.00
3	Melanjutkan pengajian	8271	7.07

Rajah 3.15 menunjukkan statistik pelajar B40 mengikut status pekerjaan setelah bergraduat daripada pengajian. Seramai 61,972 orang pelajar melaporkan telah mendapat pekerjaan setelah tamat pengajian bersamaan dengan 52.94%. Bilangan pelajar yang masih belum bekerja atau sedang menunggu penempatan pekerjaan pula adalah seramai 46,826 orang bersamaan 40% daripada bilangan keseluruhan pelajar. Sebilangan kecil pelajar pula memilih untuk melanjutkan pengajian dan meningkatkan kemahiran dalam bidang-bidang tertentu iaitu seramai 8,271 orang (7.07%).

m. Status Pengajian

Rajah 3.16. Statistik Pelajar B40 Mengikut Status Pengajian.

Jadual 3.28. Jadual Statistik Pelajar B40 Mengikut Status Pengajian.

Numerik	Status Pengajian	Bilangan	Peratus
0	Cicir	20534	17.54
1	Tamat Pengajian	96535	82.46

Daripada 117,069 orang pelajar B40 yang terlibat dalam kajian ini, seramai 96,535 orang atau 82.46% telah bergraduat daripada pengajian mereka. Manakala selebihnya iaitu seramai 20,534 atau 17.54% telah mengalami keciciran atas pelbagai sebab.

n. Jadual Analisis Deskriptif

Jadual 3.29 menunjukkan analisis deskriptif set data asal yang mengandungi jenis data numerikal dan kategorikal iaitu sebelum melalui proses transformasi data. Dapat dilihat kesemua atribut mengandungi nilai yang berbeza disebabkan oleh unit yang berlainan dan sesetengah atribut mempunyai sisihan piawai (*standard deviation*) yang tinggi berbanding atribut yang lain.

Jadual 3.29. Analisis Deskriptif Set Data Pelajar.

	Universiti	Umur Daftar	Kelayakan	Bidang Pengajian	Tajaan	CGPA	Keputusan LI	Bil Aktiviti	Status Pekerjaan	Status Pengajian
Bil.	117069	117069	117069	117069	117069	117069	117069	117069	117069	117069
min	2.162	20.005	1.823	3.773	2.361	2.832	0.547	5.411	1.541	0.825
std	0.815	1.046	0.739	1.556	0.703	1.064	0.515	8.069	0.624	0.380
min	1	16	1	1	1	0	0	0	1	0
25%	1	19	1	3	2	2.77	0	1	1	1
50%	2	20	2	4	2	3.16	1	2	1	1
75%	3	20	2	5	3	3.47	1	6	2	1
max	3	55	3	8	3	4	2	139	3	1

Manakala Jadual 3.30 pula menunjukkan analisis deskriptif terhadap set data yang telah ditransformasi kepada jenis numerikal. Perbezaan ketara yang dilihat pada jadual sebelum ini telah berkurangan seperti pada nilai sisihan piawai (*std*) dan min (*mean*). Atribut yang mempunyai nilai sisihan piawai tertinggi adalah bidang pengajian dan bilangan aktiviti iaitu masing-masing 1.556 dan 1.684. Perbezaan yang besar ini boleh dilihat dengan jelas pada Rajah 3.3 yang mana plot varians bagi kedua-dua atribut ini adalah yang tertinggi berbanding atribut yang lain.

Jadual 3.30. Analisis Deskriptif Set Data Pelajar Numerikal.

	Universiti	Umur Daftar	Kelayakan	Bidang Pengajian	Tajaan	CGPA	Keputusan LI	Bil Aktiviti	Status Pekerjaan	Status Pengajian
Bil.	117069	117069	117069	117069	117069	117069	117069	117069	117069	117069
mean	2.162	1.908	1.823	3.773	2.361	2.764	0.547	2.486	1.541	0.825
std	0.815	0.751	0.739	1.556	0.703	0.922	0.515	1.684	0.624	0.380
min	1	1	1	1	1	1	0	0	1	0
25%	1	1	1	3	2	2	0	1	1	1
50%	2	2	2	4	2	3	1	2	1	1
75%	3	2	2	5	3	3	1	4	2	1
max	3	3	3	8	3	4	2	5	3	1

d. Pembangunan Model Pengelompokan

Proses pembangunan model pengelompokan mengandungi tiga sub-proses iaitu pembangunan model pengelompokan berasaskan algoritma k-min, BIRCH dan DBSCAN bagi mengelas prestasi pelajar B40 di IPTA

a. Algoritma k-min

Teknik pengelompokan k-min adalah teknik utama yang digunakan bagi mengelompok set-set atribut pelajar B40 di dalam kajian ini. Penggunaannya adalah sangat meluas dan merupakan salah satu algoritma pembelajaran mesin tanpa selia yang sangat ringkas. Hasil pengelompokan akan disiasat bagi menentukan jenis prestasi pelajar yang telah dikelompokkan bersama. Jadual 3.31 menyenaraikan kod pseudo algoritma pengelompokan k-min.

Jadual 3.31. Kod Pseudo Algoritma Pengelompokan k-min.

Algoritma: Pengelompokan k-min
Input: ● k: bilangan kelompok ● D: satu dataset yang mengandungi n objek Output: Satu set k kelompok. Langkah-langkah: **1)** Menentukan bilangan objek k secara rawak daripada D sebagai titik tengah permulaan. **2)** Mengagihkan setiap objek ke kelompok di mana objek itu paling serupa, berdasarkan nilai min obagi bjek-objek di dalam kelompok. **3)** Mengemaskini min-min kelompok iaitu mengira nilai min bagi objek untuk setiap kelompok. **4)** Ulang langkah 2 dan 3 sehingga tiada perubahan.

b. Algoritma BIRCH

BIRCH merupakan satu algoritma pengelompokan secara hierarki yang berupaya mengelompok set data besar berdimensi tinggi dengan baik. Jadual 3.32 menyenaraikan kod pseudo algoritma pengelompokan BIRCH.

Jadual 3.32. Kod Pseudo Algoritma Pengelompokan BIRCH.

Algoritma: Pengelompokan BIRCH
1. Memasukkan data ke dalam memori: Memasukkan data ke dalam memori dengan membina pohon CF. Sekiranya memori telah penuh, bina semula pohon daripada nod daun.
2. Memampatkan data: Mengubah saiz data dengan membina pohon CF yang lebih kecil dan membuang *outlier*. Proses memampatkan data adalah atas pilihan.
3. Pengelompokan menyeluruh: Menggunakan algoritma pengelompokan sedia ada seperti k-min atau *hierarchical* ke atas pemasukan CF.
4. Penyempurnaan kelompok: Penyempurnaan kelompok adalah atas pilihan.

c. Algoritma DBSCAN

DBSCAN adalah satu algoritma pengelompokan yang berasaskan ketumpatan dan sensitif terhadap dua parameter iaitu *epsilon* dan *MinPts*. Kod pseudo bagi algorima DBSCAN ditunjukkan di dalam Jadual 3.33:

Jadual 3.33. Kod Pseudo Algoritma Pengelompokan DBSCAN.

Algoritma: Pengelompokan DBSCAN
1. Memilih titik awal secara rawak.
2. Mengisytihar pembolehubah *MinPts* dan *epsilon*
3. Mengira pembolehubah *epsilon* atau semua jarak ketumpatan terjangkau terhadap titik p menggunakan jarak *Euclidean.*
4. Sekiranya titik yang memenuhi pemboleh ubah *epsilon* melebihi nilai pemboleh ubah *MinPts*, maka titik p menjadi titik pusat dan membentuk kelompok.
5. Ulangi langkah 3 dan 4 sehingga semua titik diproses.
6. Sekiranya titik p adalah titik sempadan dan tidak ada titik ketumpatan terjangkau terhadap p, maka proses diteruskan ke titik yang lain.

Sumber: (Abdul Rahman et al., 2021)

e. Penilaian Model Pengelompokan

Dalam fasa penilaian kelompok, ketepatan atau kualiti hasil pengelompokan akan ditentukan dan disahkan. Penilaian ini merupakan satu pengukuran yang penting dalam menentukan algoritma mana yang mencapai prestasi terbaik dengan menggunakan data input bagi kajian ini. Penilaian pengelompokan adalah proses yang berdiri sendiri (*stand-alone*) dan tidak disertakan semasa proses pengelompokan. Proses ini selalu dijalankan selepas output akhir pengelompokan dihasilkan (Hassani dan Seidl 2017). Terdapat dua kaedah yang diamalkan dalam mengukur kualiti hasil kelompok iaitu pengesahsahihan dalaman (*internal validation*) dan pengesahsahihan luaran (*external validation*).

Pengesahsahihan dalaman adalah proses penilaian pengelompokan yang dibandingkan dengan keputusan pengelompokan itu sendiri iaitu perhubungan antara struktur kelompok-kelompok yang telah dibentuk. Hal ini adalah lebih realistik dan efisien dalam menyelesaikan permasalahan kajian yang melibatkan data pendidikan dengan saiz dan dimensi yang semakin meningkat saban hari.

Kajian ini menggunakan tiga jenis kaedah pengesahsahihan dalaman yang sering digunakan dalam kajian pengelompokan kebelakangan ini iaitu indeks *Davies-Bouldin* (DB), indeks pekali *silhouette* dan indeks *Calinski-Harabasz* (CH). Notasi penting yang akan digunakan dalam formula matematik bagi pengukuran penilaian pengelompokan adalah seperti berikut: *D* ialah set data input, n ialah bilangan titik data dalam *D*, *g* ialah titik tengah bagi keseluruhan set data *D*, *P* ialah bilangan dimensi *D*, *NC* ialah bilangan kelompok, C_i ialah kelompok ke-i, n_i ialah bilangan titik data dalam C_i, c_i ialah titik tengah bagi kelompok C_i dan *d(x,y)* ialah jarak antara titik x dan y (Hassani & Seidl, 2017).

a) Indeks *Davies-Bouldin*

DB ialah satu kaedah yang telah lama diperkenalkan tetapi masih digunakan secara meluas dalam pengukuran pengesahsahihan dalaman. DB menggunakan varians intra-kelompok dan jarak titik tengah inter-kelompok bagi menentukan pasangan kelompok terburuk. Jadi, pengurangan nilai indeks DB memberikan bilangan kelompok yang optimum. Formula matematik bagi DB adalah didefinisikan seperti berikut (3.4):

$$DB = \frac{1}{NC}\sum_i \frac{\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i)+\frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)}{d(c_i,c_j)} \quad (3.4)$$

b. Indeks Pekali *Silhouette*

Indeks pekali *silhouette* digunakan untuk menilai kualiti dan kekuatan sesebuah kelompok. Nilai pekali *silhouette* yang tinggi menunjukkan model dengan kelompok yang lebih baik dan memberikan isyarat yang mana sesuatu objek dipadankan dengan baik kepada kelompoknya sendiri dan tidak sepadan dengan kelompok yang berdekatan. Persamaan untuk mengira nilai pekali *silhouette* bagi sampel tunggal adalah seperti berikut (3.5):

$$s = \frac{1}{NC}\sum_i (\frac{1}{n_i}\sum_{x\in C_i} \frac{b(x)-a(x)}{max\,[b(x),a(x)]}) \quad (3.5)$$

dimana $a(x) = \frac{1}{n_i-1}\sum_{y\in C_i, y\neq x} d(x,y)$

dan $b(x) = min_{j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j} d(x,y)\right]$.

S tidak mengambil kira c_i atau g dalam pengiraan dan menggunakan jarak pasangan demi pasangan (*pairwise*) antara semua objek di dalam kelompok bagi mengira kepadatan ($a(x)$). Manakala $b(x)$ mengukur pemisahan iaitu jarak purata bagi objek-objek ke kelompok alternatif atau kelompok kedua terhampir. Daripada persamaan (3.5), julat nilai pekali *silhouette* boleh berada pada nilai antara -1 dan 1. Semakin besar nilai positif pekali, semakin tinggi kebarangkaliannya untuk dikelompokkan di dalam kumpulan yang betul. Manakala elemen dengan nilai pekali yang negatif lebih cenderung dikelompokkan di dalam kumpulan yang salah (Shutaywi & Kachouie, 2021).

c. Indeks *Calinski-Harabasz*

Indeks CH mengukur dua kriteria serentak dengan menggunakan purata hasil tambah kuasa dua antara kelompok dan purata hasil tambah kuasa dua dalam kelompok. Pengangka dalam formula menggambarkan darjah pemisahan iaitu sejauh mana titik tengah kelompok tersebar.

Manakala penyebut pula menggambarkan kepadatan iaitu sedekat mana objek-objek di dalam kelompok berkumpul di sekeliling titik tengah. Formula matematik bagi CH adalah didefinisikan seperti berikut (3.6):

$$CH = \frac{\sum_i d^2(c_i,g)/(NC-1)}{\sum_i \sum_{x \in C_i} d^2(x,c_i)/(n-NC)} \quad (3.6)$$

3.3.2 FASA OUTPUT

Fasa ini mengandungi dua sub-proses iaitu analisis kelompok dan pembangunan model pengelasan prestasi pelajar B40.

a. Analisis Kelompok

Di dalam kajian ini, model pengelompokan terbaik yang telah dibangunkan di fasa pemodelan akan dianalisis dan diperiksa secara mendalam. Analisis yang akan dilakukan adalah pengekstrakan ciri dan deskriptif statistik berdasarkan plot taburan data bagi setiap kelompok. Atribut penting yang mempengaruhi prestasi pelajar B40 akan dikenal pasti selepas analisis kelompok dijalankan.

b. Model Pengelasan Prestasi Pelajar B40

Kemudian di sub-proses kedua, set data Model B yang mengandungi label kelas kelompok akan digunakan bagi membangunkan model pengelasan prestasi pelajar B40. Model ini akan menggunakan algoritma hutan rawak, pohon keputusan dan ANN pada perisian Weka. Pemilihan tiga algoritma pengelasan ini adalah berdasarkan penggunaannya dalam kajian oleh (Sani et al., 2020) dan telah diaplikasi pada set data pelajar B40 bagi meramal kecicir an di UA.

a. Algoritma Pohon Keputusan

Algoritma pengelasan pertama yang akan digunakan bagi pembangunan model pengelasan prestasi pelajar B40 adalah algoritma pohon keputusan. Langkah-langkah pembentukan pohon keputusan adalah seperti berikut (Nadiyah, 2019):

i. Penilaian setiap atribut bagi tujuan pemilihan atribut adalah dengan menggunakan kaedah pengukuran *entropy gain* (atau *information gain*) dan indeks *Gini*.
ii. Nilai atribut tersebut akan disusun dan nilai yang akan tertinggi akan dipilih menjadi nod akar untuk memaksimumkan kadar ramalan.
iii. Seterusnya, ujian dijalankan ke atas nod akar tersebut. Sekiranya hasil ujian iaitu dahan (dalam bentuk jika-maka) menemui nod daun, algoritma akan dihentikan. Jika sebaliknya, ujian akan diteruskan dengan menggunakan atribut yang lain sehingga semua ujian menemui nod daun terakhir.
iv. Kaedah pemangkasan mencari sub-pohon atau dahan yang tidak membantu membuat ramalan dan menyebabkan terlampau padan lalu menggantikannya dengan nod daun.

b. Algoritma Hutan Rawak

Algoritma pengelasan kedua yang akan digunakan bagi pembangunan model pengelasan prestasi pelajar B40 adalah algoritma hutan rawak. Hutan rawak adalah satu kaedah pembelajaran bergabung bagi teknik pengelasan yang terdiri daripada satu set pohon keputusan. Langkah-langkah pembentukan hutan rawak adalah seperti berikut (Mamata Laxmi et al., 2020):

i. Sampel N dipilih secara rawak daripada set data.
ii. Daripada sampel-sampel N yang telah dipilih, satu pohon keputusan dibina bagi setiap sampel.
iii. Langkah (i) dan (ii) diulang dengan mengikut penetapan bilangan pohon di dalam algoritma sehingga pohon-pohon keputusan siap dibina.
iv. Pengundian akan dilakukan bagi setiap keputusan yang dikelaskan.
v. Keputusan yang menerima undian tertinggi dipilih sebagai keputusan akhir pengelasan.

c. Algoritma ANN

Algoritma pengelasan terakhir yang akan digunakan bagi pembangunan model pengelasan prestasi pelajar B40 adalah algoritma ANN. Langkah-langkah peramalan menggunakan algoritma ANN adalah seperti berikut (Nadiyah, 2019):

i. Nilai awal bagi setiap unit lapisan input berserta pemberat akan dihantar melalui rangkaian ke unit-unit di lapisan tersembunyi.

ii. Pengiraan jumlah nilai pemberat bagi setiap unit di lapisan tersembunyi akan dilakukan. Manakala nilai output akan didapati dengan cara mengenakan fungsi tidak linear seperti *sigmoid* ke atas jumlah nilai pemberat.

iii. Nilai output yang telah didapati akan dihantar ke lapisan tersembunyi seterusnya ataupun ke lapisan output. Pengiraan seperti di langkah (ii) akan dilaksanakan bagi mencari nilai output.

iv. Perbandingan antara nilai output bagi lapisan output dengan nilai sebenar kelas akan dilakukan bagi mendapatkan nilai ralat. Seterusnya nilai ralat tersebut akan dihantar semula melalui rangkaian dan nilai pemberat bagi setiap rangkaian akan diubah suai mengikut nilai ralat dan kadar pembelajaran yang telah ditetapkan.

Langkah (i) hingga (iv) akan diulang mengikut nilai kitaran pembelajaran yang telah ditetapkan sehingga tiada ralat atau nilai ralat kecil sahaja yang dihasilkan dan prestasi peramalan menjadi baik.

c. Penilaian Prestasi Model Pengelompokan

Prestasi model pengelasan yang dihasilkan akan diukur dengan metrik pengukuran ketepatan (*accuracy*). Jadual 3.34 menunjukkan sampel matrik kekeliruan yang akan digunakan dalam mengira metrik pengukuran penilaian prestasi model pengelompokan. Kedua-dua positif benar (PB) dan negatif benar (NB) menunjukkan hasil pengelasan yang betul. Positif palsu (PP) adalah di mana keputusan ramalan menunjukkan positif tetapi sebenarnya ia negatif. Manakala negatif palsu (NP) adalah kes di mana keputusan ramalan menunjukkan negatif tetapi ia sebenarnya positif.

Jadual 3.34. Sampel Matrik Kekeliruan.

		Kelas Ramalan	
		Positif	Negatif
Kelas Sebenar	Positif	Positif Benar (PB)	Negatif Palsu (NP)
	Negatif	Positif Palsu (PP)	Negatif Benar (NB)

Seperti yang ditunjukkan pada persamaan (3.7), ketepatan adalah peratusan bilangan rekod yang berjaya dikelaskan dengan betul (Ko & Leu, 2020; Nadiyah, 2019).

$$Ketepatan = \frac{PB+NB}{PB+NB+PP+NP} \tag{3.7}$$

Statistik kappa adalah satu metrik pengukuran dan didefinisikan seperti berikut (Thakar et al., 2017):

$$K = \frac{P(A)-P(E)}{1-P(E)} \tag{3.8}$$

di mana *P(A)* ialah peratus persetujuan dan *P(E)* ialah peluang persetujuan (*chance agreement*). Sekiranya nilai K=1, maka terdapat persetujuan yang ideal antara pengelas dan *ground truth*. Dengan kata lain, model yang dihasilkan mempunyai prestasi yang baik. Jika nilai K=0, ia menandakan prestasi pengelasan adalah lemah (Hussain et al., 2018). Statistik kappa adalah sangat berguna dalam pengelasan pelbagai kelas (*multi-class*) dan masalah kelas tidak seimbang.

3.3 PERISIAN

Untuk membangunkan model pembelajaran mesin bagi kajian ini, beberapa perisian telah dikenal pasti amat sesuai untuk digunakan. Perisian yang telah dipilih dapat membantu melancarkan proses-proses pra-pemprosesan data, pembangunan model pembelajaran mesin berasaskan algoritma pengelompokan dan pengelasan, membuat analisa statistik dan menganalisa data secara deskriptif. Berikut disenaraikan perisian-perisan yang digunakan dalam kajian ini.

a. Rapid Miner

Perisian Rapid Miner merupakan satu platfom pembelajaran mesin yang popular dan banyak digunakan oleh para pengkaji seluruh dunia. Antara kajian pembelajaran mesin dan pelombongan data dalam bidang pendidikan yang menggunakan perisian Rapid Miner ada

dilaporkan di dalam kajian kepustakaan yang telah dikaji (Al-Hagery et al 2020, Krizanic 2020, Amra 2017, Agrawal et al 2017, Tabrez-Nafis et al 2017).

Rapid Miner ialah satu perisian sumber terbuka di mana banyak operator dan algoritma telah disertakan didalamnya dan tidak memerlukan para pengguna melakukan pengaturcaraan. Ruang kerjanya mengandungi beberapa operator dan alatan yang mudah digunakan bagi membantu membina, melaksanakan, menilai dan visualisasi model yang akan dibangunkan (Al-Hagery 2020). Reka bentuk proses aliran kerja adalah secara visual dengan menggunakan teknik tarik dan lepas yang memudahkan pengguna melakukan ekperimentasi berkaitan projek sains data. Penggunaan Rapid Miner dalam kajian ini adalah pada peringkat awal pra-pemprosesan data di mana dapat mempercepat dan mempermudahkan urusan penyediaan data sebelum pembinaan model pembelajaran mesin.

b. Google Colab

Di samping itu, analisis statistik, pemilihan atribut dan pembinaan model pengelompokan dalam kajian ini akan menggunakan platform Google Colab berasaskan *python* di pelayar web. Platform ini membolehkan penyelidik menulis dan menjalankan kod pengaturcaraan terutamanya berkaitan dengan pembelajaran mesin dan analisis data. Platform ini tidak memerlukan sebarang perisian dipasang pada komputer dan memberikan akses percuma kepada sumber perkomputeran berasaskan awan.

Selain itu, pembangunan model pengelompokan bagi kajian ini yang menggunakan pengaturcaraan *python* telah memanfaatkan sepenuhnya modul scikit-learn. Scikit-learn adalah sebuah modul atau perpustakaan pembelajaran mesin yang percuma dengan menyediakan pelbagai algoritma pengelasan, regresi dan pengelompokan bagi kegunaan pengaturcaraan python. Kelebihan menggunakan modul scikit-learn adalah mudah diimplementasi, menyediakan banyak algoritma pengelompokan yang sesuai dengan kajian dan set data pelajar serta mudah dibuat perbandingan dari segi prestasi model yang dibangunkan.

c. SPSS

Perisian SPSS (Statistical Package for the Social Sciences) adalah satu platfom perisian statistik yang sangat berkuasa bagi membuat analisa statistik terhadap set data kajian.

Penggunaan perisian SPSS dalam kajian ini adalah bagi membuat ujian statistik iaitu ujian t berpasangan ke atas skor penilaian pengelompokan yang telah dihasilkan sejurus model pengelompokan dihasilkan.

d. Power BI

Penghasilan visual daripada set data dalam kajian ini terutama semasa membuat analisis deskriptif telah menggunakan sepenuhnya perisian Power BI. Perisian ini menyediakan pengguna dengan peralatan dan platfom untuk menganalisis dan menghasilkan visual dengan cepat serta mudah untuk dioperasikan. Visual yang dihasilkan adalah dalam bentuk plot dan graf di mana bertujuan memudahkan kita memahami kandungan di sebalik data kajian.

e. Weka

Weka merupakan satu perisian pembelajaran yang digunakan dan dilaporkan secara meluas di dalam kajian-kajian pengelasan prestasi pelajar IPT. Perisian ini adalah perisian sumber terbuka dan mengandungi koleksi algoritma pembelajaran mesin yang besar untuk digunakan bagi pembangunan model mengikut keperluan kajian. Kajian ini akan memanfaatkan kelebihan yang ada pada Weka bagi membangunkan model pengelasan prestasi pelajar B40 menggunakan algoritma pengelasan hutan rawak, pohon keputusan dan ran ANN. Weka menyediakan platfom yang membolehkan algoritma pembelajaran mesin diaplikasikan secara terus kepada set data dan ini memudahkan dan mempercepatkan proses pembangunan model dalam kajian ini.

3.4 KESIMPULAN

Bab metodologi kajian ini menerangkan dengan jelas dan menyeluruh langkah-langkah yang dilakukan termasuk fasa data, fasa pembangunan model, fasa penilaian dan fasa output. Proses pra-pemprosesan data telah dilaksanakan termasuklah pemilihan atribut secara berselia dan tanpa selia. Atribut-atribut akhir yang telah dihasilkan selepas proses pra-pemprosesan telah dijadikan tiga set model bagi dijadikan input kepada model pengelompokan di fasa pembangunan model. Atribut-atribut ini juga diterangkan secara deskriptif di dalam bab ini. Seterusnya tiga jenis algoritma pengelompokan yang digunakan dalam kajian ini juga telah

diterangkan iaitu algoritma K-min, BIRCH dan DBSCAN. Kemudian, tiga jenis metrik pengukuran bagi tujuan penilaian pengesahsahihan dalaman iaitu indeks DB, indeks pekali *silhouette* dan indeks CH diterangkan secara terperinci. Fasa pemodelan dan penilaian akan diterangkan di bab seterusnya.

BAB IV

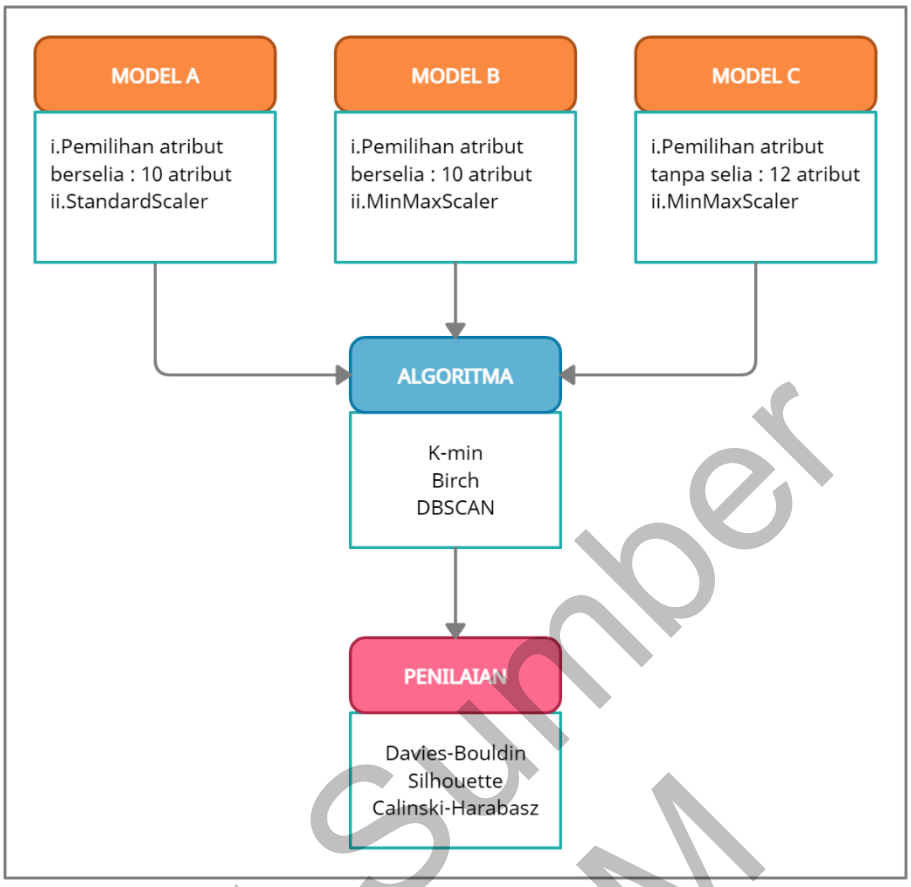
MODEL PENGELOMPOKAN PELAJAR B40

4.1 PENGENALAN

Perbincangan dalam bab ini mengkhususkan tentang keputusan yang diperolehi bagi setiap eksperimen yang dijalankan. Penetapan nilai *k* terbaik akan dilakukan pada permulaan eksperimen kerana hasil keputusannya akan digunakan di dalam eksperimen yang berikutnya Terdapat tiga model yang akan melalui eksperimen pengelompokan menggunakan tiga algoritma berbeza. Hasil pengelompokan bagi setiap model akan dinilai bagi mengenal pasti teknik yang terbaik.

4.2 PENETAPAN PARAMETER MODEL PENGELOMPOKAN

Eksperimen dijalankan ke atas 3 model di mana setiap model mempunyai tetapan parameter yang berbeza dan tetapan penormalan yang juga berbeza setelah proses pra-pemprosesan. Rajah 4.1 menunjukkan ilustrasi rekabentuk penetapan eksperimen di mana model pertama (A) merupakan set data yang mengandungi 10 atribut yang dipilih oleh pemilihan atribut berselia dan menggunakan tetapan penormalan *StandardScaler*. Model kedua (B) pula ialah set data yang mengandungi 10 atribut yang dipilih oleh pemilihan atribut berselia dan menggunakan tetapan penormalan *MinMaxScaler*. Manakala model ketiga (C) merupakan set data yang mengandungi 12 atribut yang dipilih oleh pemilihan atribut tanpa selia dan menggunakan tetapan penormalan *MinMaxScaler*.



Rajah 4.1. Rekabentuk Penetapan Eksperimen.

4.3 PEMBANGUNAN MODEL PENGELOMPOKAN

4.3.1 Teknik k-min

a. Penalaan Parameter

Algoritma k-min diuji dengan penetapan parameter seperti yang ditunjukkan di dalam Jadual 4.1 di bawah. Penalaan k bagi bilangan kelompok untuk dibentuk dan bilangan sentroid untuk dihasilkan pada nilai antara 2 hingga 10 dan akan ditetapkan setelah teknik *elbow* dan analisis *silhouette* dijalankan. Nilai *n_iter* pada asasnya menentukan berapa banyak set sentroid yang dipilih secara rawak sekiranya algoritma ini digunakan dan ditetapkan pada nilai lalai (*default value*) bersamaan 10. Manakala *max_iter* iaitu bilangan maksimum ulangan algoritma bagi sekali larian ditala pada nilai 100, 200, 300 dan 400. Pengukuran jarak *Distance metric* ditetapkan kepada *Euclidean*.

Hasil ujian penalaan parameter *max_iter* menunjukkan nilai 100, 200 dan 400 telah memberikan keputusan yang malar walaupun masa perkomputerannya berubah-ubah. Perkara ini telah menyukarkan penentuan prestasi terbaik bagi setiap nilai yang ditala. Justeru, nilai *max_iter* bersamaan 300 telah dipilih untuk ditetapkan pada algoritma k-min ketika fasa pembinaan model.

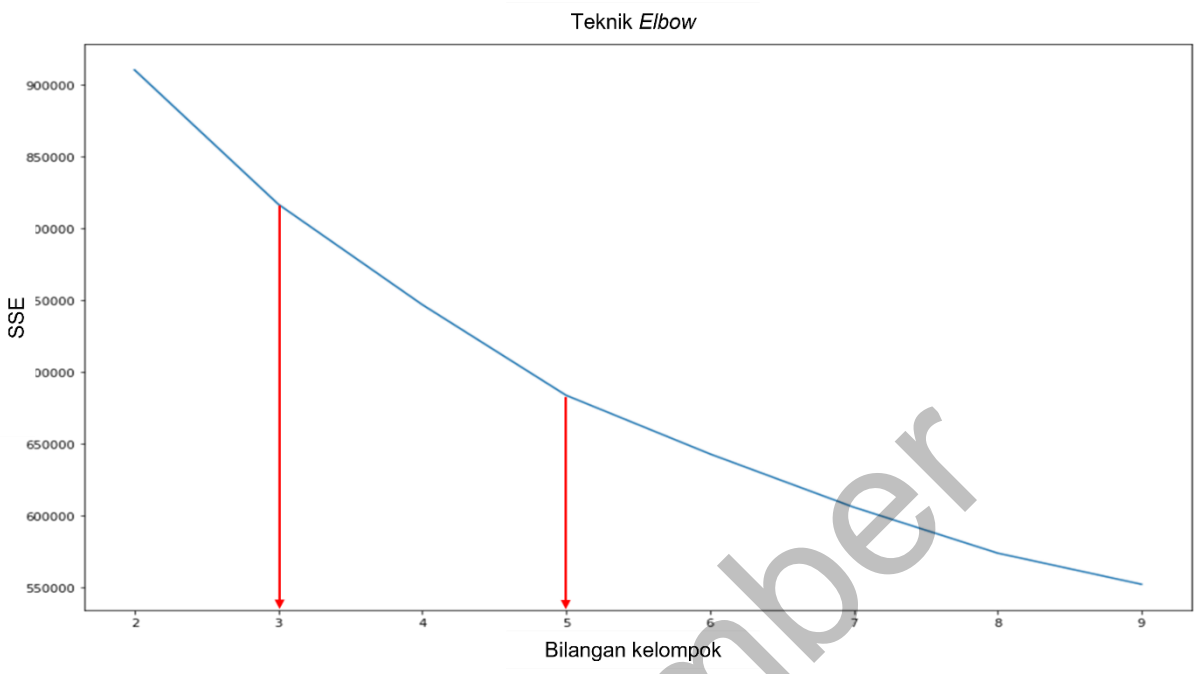
Jadual 4.1. Penalaan parameter Algoritma k-min.

Pembolehubah	**Penalaan Nilai**
k	2 hingga 10
n_iter	10
max_iter	100, 200, 300, 400
Distance metric	*euclidean*

b. Penentuan Nilai k Terbaik

Tiga teknik pengukuran digunakan bagi menghasilkan visual dan menetapkan nilai k dalam kajian ini. Teknik yang pertama ialah teknik *elbow* iaitu satu teknik secara *heuristic* dan visualisasi berdasarkan pengamatan pada graf yang diplot. Teknik ini telah digunakan dalam beberapa kajian lepas oleh (Aggarwal & Sharma, 2019; X. Li et al., 2021; Marbouti et al., 2020) dan dibuktikan berkesan dalam menyelesaikan masalah ini. Pelaksanaan teknik ini adalah dengan menguji teknik pengelompokan k-min ke atas set data pada satu julat k dan jarak purata antara setiap titik dalam kelompok dengan titik sentroid akan dikira. Dengan pertambahan bilangan kelompok (k), jarak purata akan berkurangan. Kemudian, graf purata jarak bagi setiap nilai k akan diplot bagi mencari satu lengkukan seperti siku di sepanjang garisan graf. Lengkukan tajam ataupun jatuhan purata jarak yang dalam menunjukkan nilai k yang optimum dan akan dipilih untuk menjadi bilangan kelompok dalam fasa pembinaan model.

Dalam penentuan ini, julat nilai k yang dipilih adalah antara 2 hingga 10 dan algoritma k-min telah dijalankan menggunakan setiap nilai k yang ditetapkan. Berdasarkan plot dalam rajah 4.2 di bawah, terdapat lengkukan SSE pada nilai k bersamaan k=3 dan k=5.

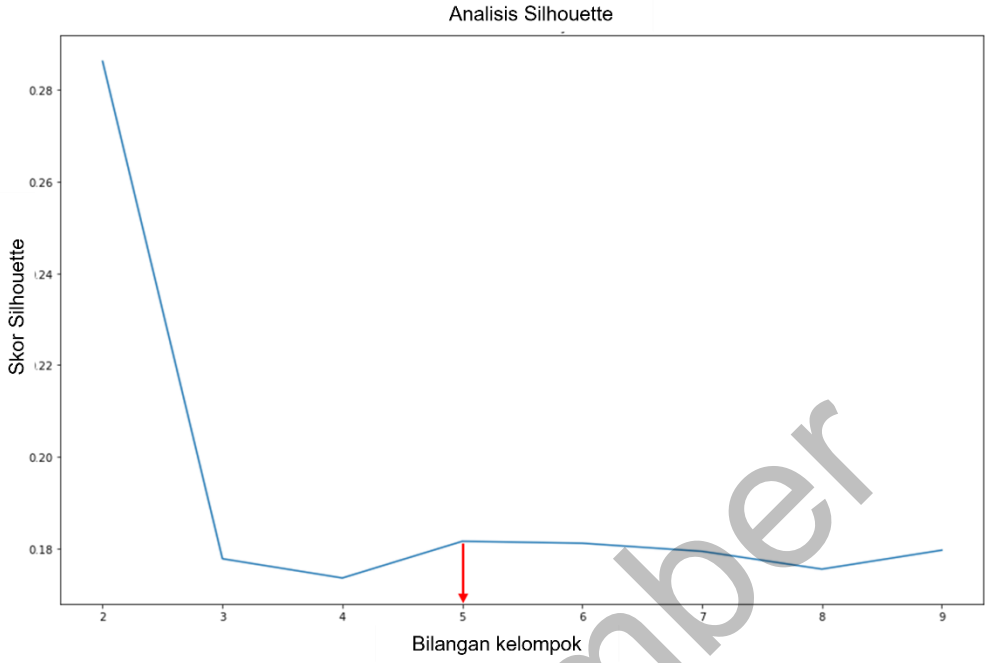


Rajah 4.2. Teknik *Elbow* Bagi Penetapan Bilangan Kelompok.

Teknik yang kedua adalah menggunakan analisis plot *silhouette* dengan mengira pekali bagi setiap titik data bagi mengukur kemiripannya dengan kelompok sendiri berbanding kelompok lain. Nilai pekali *silhouette* berapa pada julat [1, -1] di mana nilai yang tinggi menunjukkan yang objek itu dipadankan dengan baik kepada kelompoknya.

Dalam penentuan ini, julat nilai k yang dipilih adalah antara 2 hingga 10. Bagi setiap model k-min, nilai pekali *silhouette* telah diplot dan perubahan nilai bagi setiap kelompok telah diperhatikan pada rajah 4.3 di bawah.

Plot *silhouette* menunjukkan k=2 menghasilkan skor yang terbaik. Namun secara realitinya, dua kelompok bagi set data ini adalah tidak mencukupi. Jadi, nilai skor *silhouette* yang kedua tertinggi telah dipilih iaitu bersamaan k=5. Nilai ini menunjukkan persetujuan antara kedua-dua teknik bagi penentuan nilai k terbaik untuk menetapkan bilangan kelompok.



Rajah 4.3. Analisis Silhouette Bagi Penetapan Bilangan Kelompok.

Justeru, nilai k yang terbaik adalah k=5 dan akan digunakan pada ketiga-tiga algoritma pengelompokan Selain daripada itu, hasil daripada keputusan analisis di atas, penetapan nilai parameter bagi algoritma k-min ditunjukkan di Jadual 4.2.

Jadual 4.2. Penetapan Parameter Algoritma k-min.

Pembolehubah	**Penetapan Nilai**
k	2 hingga 10
n_iter	10
max_iter	300
Distance metric	euclidean

4.3.2 Teknik BIRCH

Algoritma BIRCH diuji dengan penalaan parameter seperti yang ditunjukkan di dalam Jadual 4.3 di bawah. Penetapan *n_clusters* iaitu bilangan kelompok setelah langkah pengelompokan terakhir di mana pasangan-pasangan kelompok kecil digabung menjadi kelompok besar pada nilai 5. Perubahan parameter *n_clusters* kepada integer 5 menyebabkan model pengelompokan dipadankan kepada pengelompokan *agglomerative*. Seterusnya *threshold* ditetapkan pada nilai lalai (*default value*) 0.5 dan *branching_factor* pada nilai lalai 50.

Jadual 4.3. Penetapan Parameter Algoritma BIRCH.

Pembolehubah	Penetapan Nilai
n_clusters	5
threshold	0.5
branching_factor	50

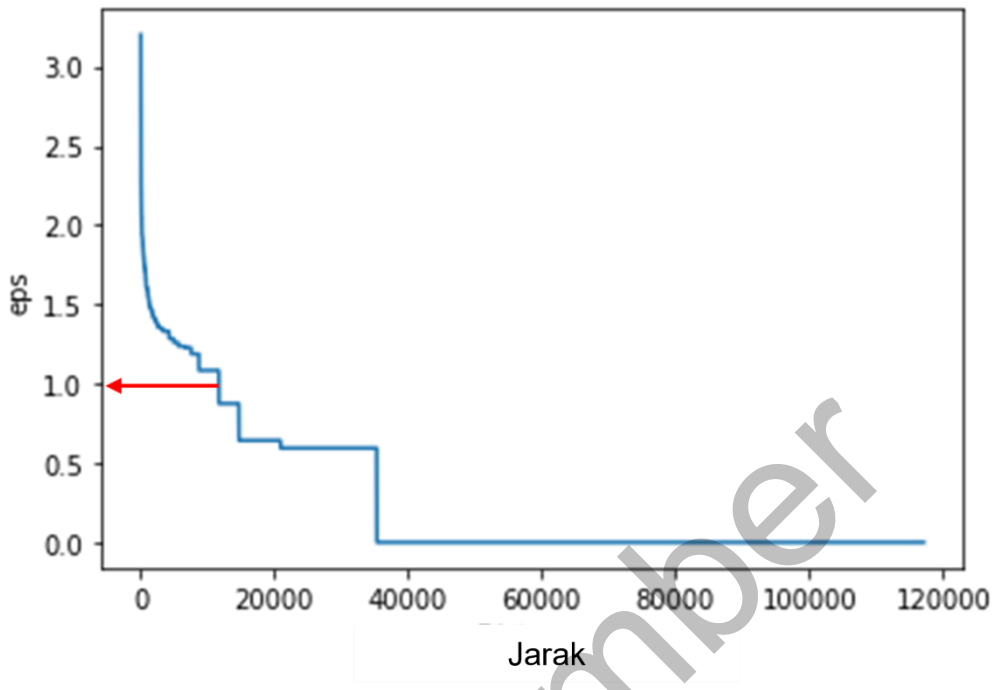
4.3.3 Teknik DBSCAN

Algoritma DBSCAN diuji dengan penetapan nilai seperti yang ditunjukkan di dalam Jadual 4.4 di bawah. Penetapan *epsilon* atau radius pada nilai lalai 0.5 bagi Model B dan C manakala 1.0 bagi Model A. Jenis pengukuran jarak iaitu *metric* ditetapkan kepada *euclidean* dan penetapan *min_samples* yang menerangkan bilangan sampel dalam suatu kawasan untuk satu titik dipertimbangkan sebagai titik utama pada nilai 1000 untuk Model B dan C manakala 1200 untuk Model A.

Jadual 4.4. Penetapan Parameter Algoritma DBSCAN.

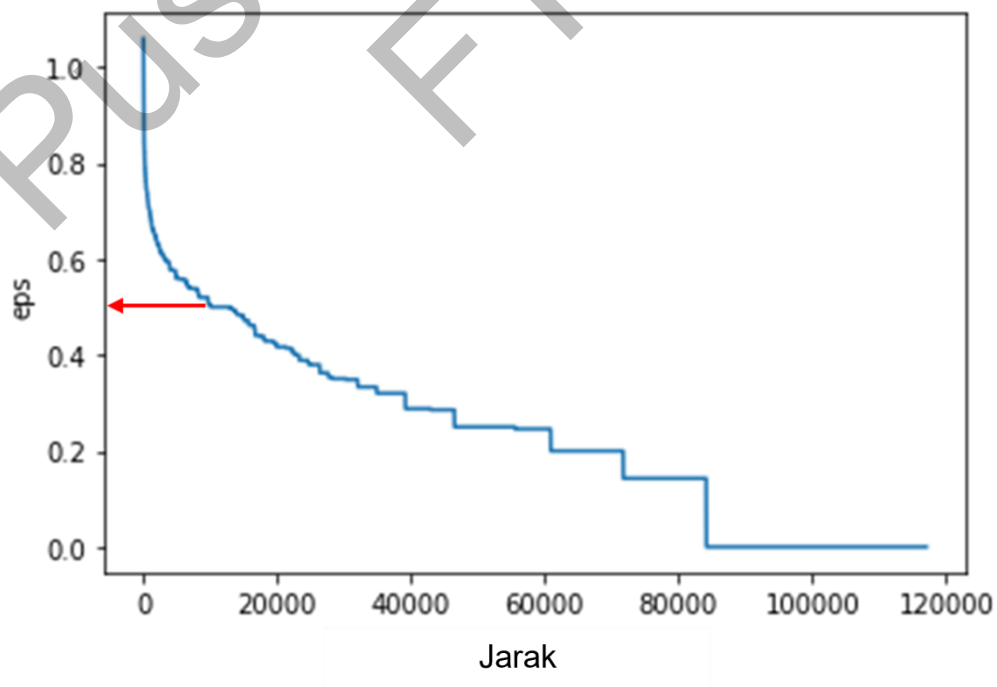
Pembolehubah	Penetapan Nilai
epsilon	0.5 (Model C) & 1.0 (Model A&B)
metric	*euclidean*
min_samples	1000 (Model B&C) & 1200 (Model A)

Seperti mana perbincangan mengenai algoritma DBSCAN di bab 2, dapat diketahui yang pembinaan model DBSCAN adalah dipengaruhi oleh penetapan nilai parameter *epsilon*. Penetapan nilai *epsilon* dalam kajian ini berdasarkan plot jarak purata setiap titik kepada k-jiran terdekatnya. Lekukan pada plot akan menunjukkan nilai parameter *epsilon* yang optimum. Lekukan adalah nilai ambang di mana berlaku perubahan besar pada kecerunan di sepanjang garisan plot. Berdasarkan Rajah 4.4, nilai epsilon bagi Model A dan B adalah pada *eps* = 1.0 setelah pemerhatian dibuat pada lekukan di sepanjang garisan plot.



Rajah 4.4. Epsilon Model A dan B.

Rajah 4.5 pula menunjukkan plot *Epsilon* bagi model C. Terdapat perbezaan plot seperti yang ditunjukkan di Rajah 4.4 dan Rajah 4.5 kerana bilangan atribut yang berlainan antara Model A, B dan C telah mempengaruhi nilai pengiraan jarak purata kepada k-jiran terdekat. Seperti dalam Rajah 4.5, nilai *epsilon* bagi Model C adalah pada *eps* = 0.5 berdasarkan garisan plot dan nilai ini adalah bersamaan dengan nilai lalai.



Rajah 4.5. Epsilon Model C.

4.3.4 Penetapan Parameter Akhir

Setelah eksperimen penalaan parameter bagi algoritma k-min, BIRCH dan DBSCAN telah selesai, nilai-nilai terbaik akan dipilih bagi tujuan tetapan parameter akhir algoritma dalam eksperimen yang seterusnya. Penetapan nilai parameter akhir adalah seperti yang ditunjukkan di dalam Jadual 4.5 di bawah.

Jadual 4.5. Penetapan Parameter Akhir

	Pembolehubah	Penetapan Nilai
k-min	*k*	5
	n_iter	10
	max_iter	300
	Distance metric	euclidean
BIRCH	*n_clusters*	5
	threshold	0.5
	branching_factor	50
DBSCAN	*epsilon*	0.5 (Model C) & 1.0 (Model A&B)
	metric	*euclidean*
	min_samples	1000 (Model B&C) & 1200 (Model A)

4.3.5 Penilaian Prestasi Model Pengelompokan

a. Keputusan Penilaian

Penunjuk yang perlu diberi perhatian semasa memeriksa skor daripada metrik penilaian pengelompokan adalah seperti berikut: nilai indeks DB yang rendah; indeks pekali silhouette dengan nilai positif yang tinggi; dan indeks CH dengan nilai yang tinggi menunjukkan prestasi pengelompokan yang baik. Jadual 4.6 menunjukkan tiga metrik penilaian pengesahsahihan dalaman yang digunakan dalam kajian ini berserta ukuran nilai indeks yang menunjukkan prestasi pengelompokan terbaik.

Jadual 4.6. Penilaian Pengesahsahihan Dalaman.

Jenis Penilaian	Nilai Indeks Terbaik
Davies-Bouldin (DB)	Nilai indeks minimum
Silhouette	Nilai indeks pekali maksimum
Calinski-Harabasz (CH)	Nilai indeks maksimum

Jadual 4.7 di bawah menyenaraikan prestasi pengelompokan berdasarkan pengesahsahihan dalaman bagi tiga algoritma pengelompokan, k-min, BIRCH dan DBSCAN yang diaplikasikan pada tiga jenis model berbeza iaitu Model A, Model B dan Model C. Metrik penilaian pengesahsahihan dalaman yang dipilih boleh menunjukkan sekiranya kelompok terpisah dengan baik dan tidak bertindih.

b. Kedudukan Model Mengikut Skor

Selepas perbandingan prestasi dilakukan, prestasi pengelompokan setiap algoritma kemudiannya disusun berdasarkan kedudukan bagi setiap model. Kedudukan akhir ditentukan berdasarkan jumlah skor kedudukan yang dikira menggunakan purata skor model bagi setiap algoritma seperti yang ditunjukkan di dalam Jadual 4.8. Model B mencatatkan jumlah skor terbaik iaitu 4, diikuti oleh Model A dengan nilai skor 6 pada tempat kedua dan Model C berada pada kedudukan terakhir dengan jumlah skor 8.

Jadual 4.7. Keputusan Penilaian Pengelompokan Antara Algoritma Yang Digunakan.

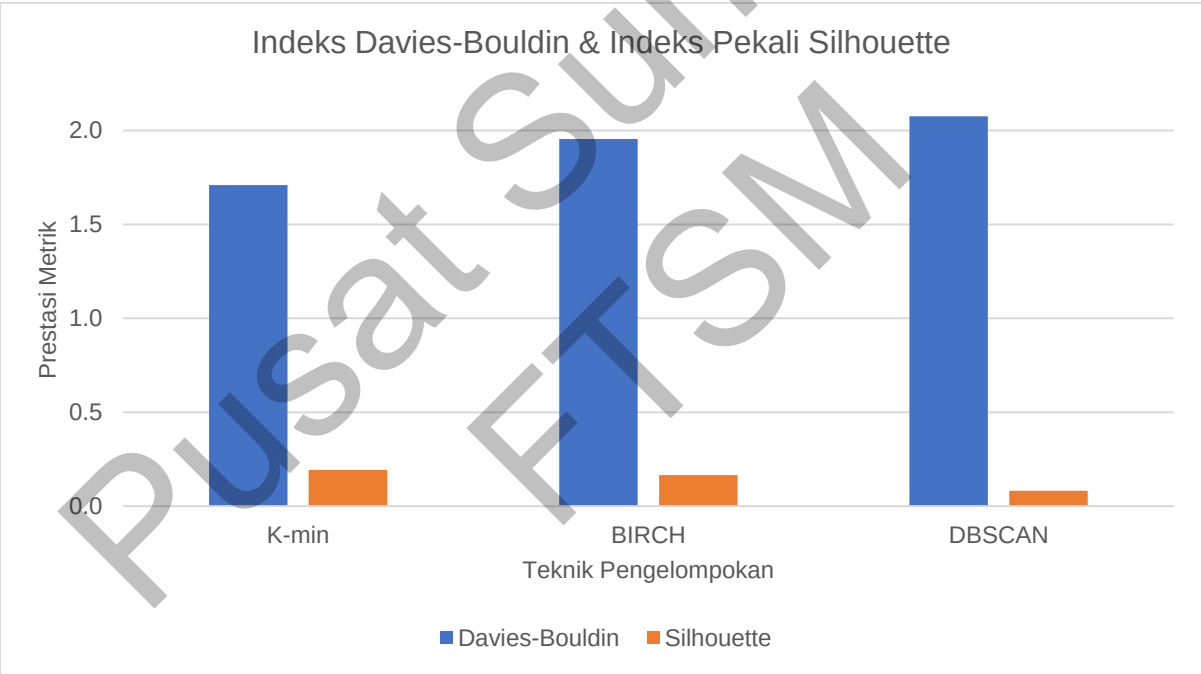
	Model A				Model B				Model C			
	DB	**Silhouette**	**CH**	**Masa (s)**	**DB**	**Silhouette**	**CH**	**Masa (s)**	**DB**	**Silhouette**	**CH**	**Masa (s)**
k-min	1.85	0.18	20846.33	238	1.71	0.192	24946.34	238	1.891	0.16	17358.46	231
BIRCH	2.08	0.14	17203.63	254	1.96	0.165	21722.28	223	2.222	0.12	14664.02	235
DBSCAN	1.43	-0.17	3034.38	269	2.08	0.082	11218.19	306	2.229	-0.02	6760.80	308

Jadual 4.8. Skor Kedudukan Setiap Model Mengikut Algoritma.

	Model A				Model B				Model C			
	DB	**Silhouette**	**CH**	**Purata Skor**	**DB**	**Silhouette**	**CH**	**Purata Skor**	**DB**	**Silhouette**	**CH**	**Purata Skor**
k-min	2	2	2	2	1	1	1	**1**	3	1	3	2
BIRCH	2	2	2	2	1	1	1	**1**	3	3	3	3
DBSCAN	1	1	3	**2**	2	2	1	**2**	3	3	2	3
Jumlah Purata Skor	6				4				8			
Kedudukan	2				**1**				3			

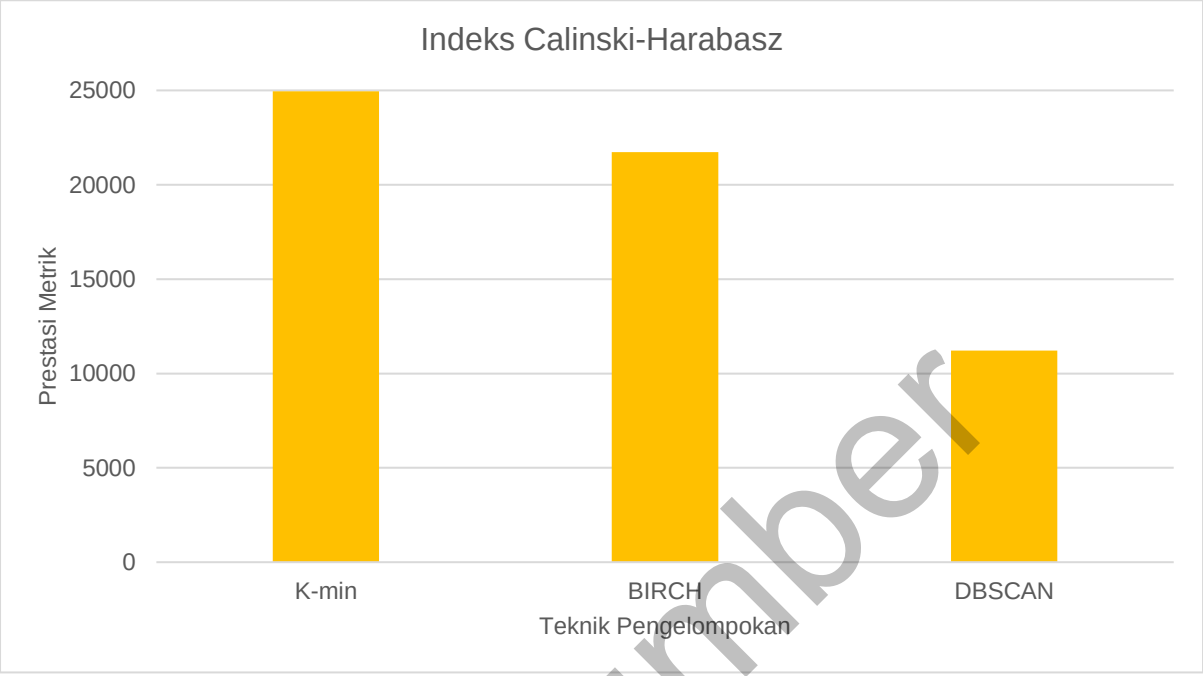
c. Perbandingan Penilaian Model B

Berdasarkan perbandingan antara model-model di atas, kita akan memfokuskan analisis penilaian terhadap tiga algoritma pengelompokan yang digunakan pada Model B kerana jelas mengatasi dua model yang lain. Pada Rajah 4.6, plot indeks DB dan indeks pekali silhouette jelas menunjukkan penguasaan algoritma k-min berbanding dua algoritma yang lain. Nilai indeks DB bagi algoritma k-min adalah 1.71 iaitu lebih rendah berbanding *BIRCH* (1.96) dan DBSCAN (2.08). Bagi indeks pekali silhouette, Rajah 4.6 menunjukkan algoritma k-min sekali lagi mencatatkan keputusan terbaik di mana nilai indeks pekali silhouette adalah yang tertinggi iaitu bersamaan 0.192 berbanding BIRCH (0.165) dan DBSCAN (0.082). Jadi, berdasarkan dua penilaian pengesahsahihan dalaman tersebut, algoritma K-min adalah yang terbaik bagi Model B.



Rajah 4.6. Perbandingan Metrik Penilaian Indeks DB Dan Indeks Pekali *Silhouette* Bagi Pengelompokan Model B.

Selain daripada itu, plot pada Rajah 4.7 di bawah memaparkan perbandingan indeks CH bagi ketiga-tiga algoritma yang dijalankan pada Model B. Sekali lagi algoritma k-min menunjukkan prestasi yang baik apabila mencatatkan nilai indeks yang tertinggi berbanding dua algoritma yang lain. Nilai indeks CH yang dicatatkan oleh ketiga-tiga algoritma ialah 24946.34 bagi k-min, 21722.28 bagi BIRCH dan 11218.19 bagi DBSCAN.



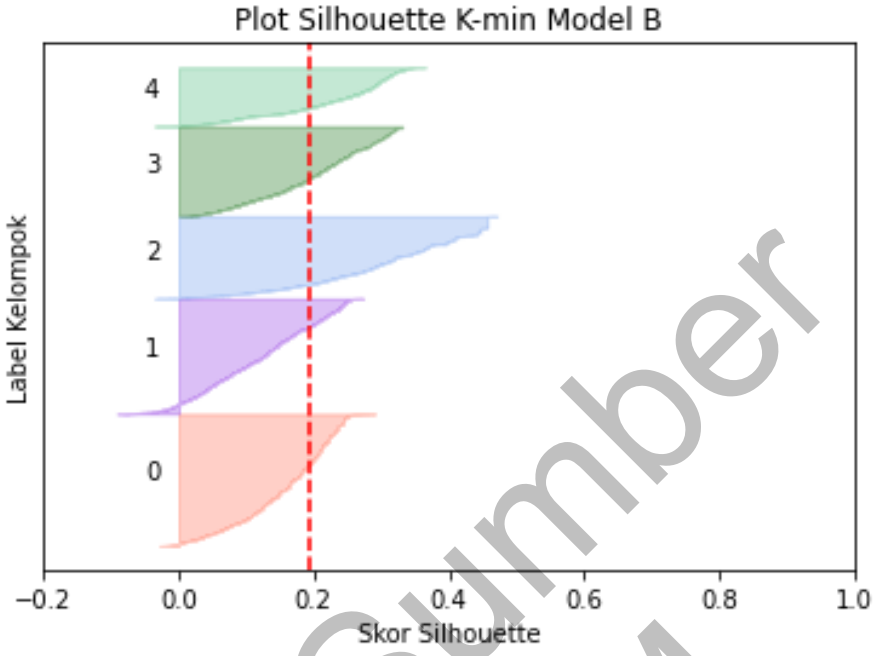
Rajah 4.7. Perbandingan Metrik Penilaian Indeks CH Bagi Pengelompokan Model B.

d. Perbandingan Penilaian Silhouette Bagi k-min Dan BIRCH

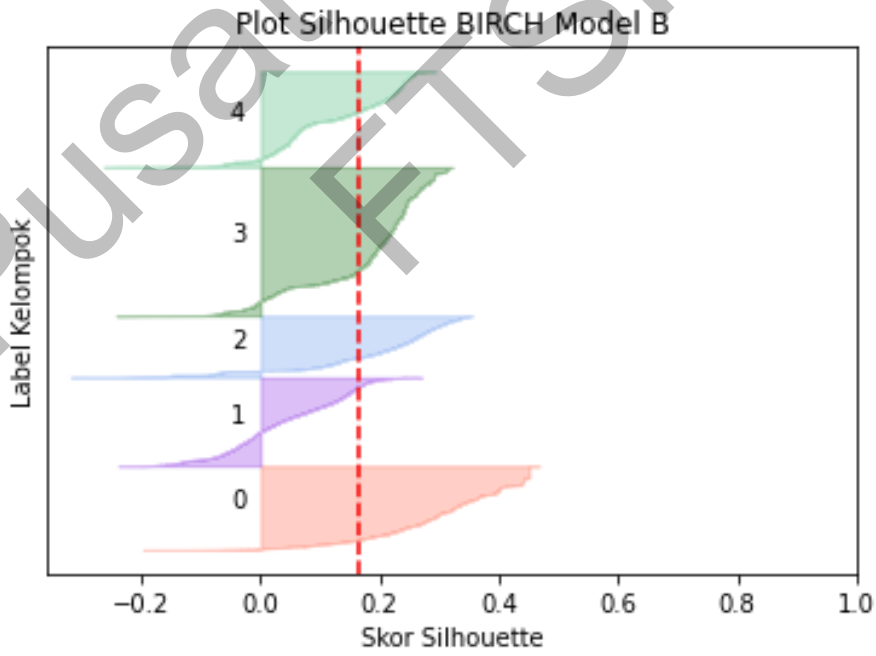
Analisis tambahan bagi menyiasat hasil pengelompokan algoritma k-min dan BIRCH adalah dengan memeriksa plot *silhouette* seperti di Rajah 4.8 dan 4.9. Plot di Rajah 4.8 menunjukkan kesemua lima kelompok yang dihasilkan oleh algoritma k-min berada di atas garisan nilai purata *silhouette* dan ini memberikan satu gambaran hasil pengelompokan yang baik. Nilai purata indeks pekali silhouette Model B bagi algoritma K-min ialah 0.192 dan ditandakan pada plot dengan garisan merah. Turun naik (*fluctuation*) saiz plot silhouette pula tidak menunjukkan perubahan yang ketara di mana kesemua kelompok mencatatkan nilai purata skor yang positif dan ini menandakan hampir kesemua nilai ditentukan ke dalam kelompok yang betul.

Plot silhouette di Rajah 4.9 menunjukkan kesemua lima kelompok yang dihasilkan oleh algoritma BIRCH berada di atas garisan nilai purata *silhouette* dan menunjukkan hasil pengelompokan yang baik. Nilai purata indeks pekali *silhouette* Model B bagi algoritma BIRCH ialah 0.165 dan ditunjukkan pada plot dengan garisan merah. Berbanding algoritma k-min, nilai pekali *silhouette* algoritma BIRCH mencatatkan sedikit penurunan dengan perbezaan sebanyak 0.027. Turun naik saiz plot pula menunjukkan perubahan besar dan ketara di mana

saiz yang kecil ditunjukkan oleh kelompok 1, 2 dan 4 manakala saiz yang besar pula ditunjukkan oleh kelompok 0 dan 3.



Rajah 4.8 . Plot *Silhouette* Bagi Pengelompokan k-min Untuk Model B.



Rajah 4.9. Plot *Silhouette* Bagi Pengelompokan BIRCH Untuk Model B.

e. Ujian Statistik

Di akhir penilaian prestasi pengelompokan, ujian statistik dilakukan terhadap kualiti prestasi yang telah dihasilkan oleh kedua-dua algoritma. Ujian-t berpasangan telah dijalankan ke atas skor indeks *silhouette* bagi kedua-dua teknik k-min dan BIRCH seperti yang ditunjukkan di Rajah 4.5 dan 4.6 menggunakan perisian SPSS. Hipotesis bagi ujian-t berpasangan di dalam kajian ini adalah seperti berikut:

H_0: Min indeks pekali *silhouette* BIRCH lebih besar daripada min indeks pekali *silhouette* k-min. Ini bermaksud nilai purata pekali indeks *silhouette* BIRCH > nilai purata pekali indeks *silhouette* k-min

H_1: Min indeks pekali silhouette BIRCH lebih kecil daripada min indeks pekali silhouette k-min. Ini bermaksud nilai purata indeks pekali *silhouette* BIRCH < nilai purata indeks pekali *silhouette* k-min

Penentuan penerimaan hipotesis nol (H_0) adalah berdasarkan aras keyakinan 95% dan $\alpha = 0.05$. Sekiranya nilai $p < \alpha$, maka hipotesis nol akan ditolak. Berdasarkan Jadual 4.9 di bawah, keputusan ujian menunjukkan nilai $p = 0$ dan ini bermakna hipotesis nol ditolak dan hipotesis alternatif (H_1) diterima. Ini menunjukkan algoritma k-min menghasilkan prestasi yang lebih baik berbanding BIRCH kerana nilai purata indeks pekali *silhouette* k-min adalah lebih besar dan keputusan ujian adalah signifikan secara statistik pada aras 0.05.

Jadual 4.9. Keputusan Ujian-T Berpasangan Bagi Min Skor Silhouette Untuk Algoritma k-min Dan BIRCH.

Min	Sisihan Piawai	Min Ralat Piawai	Selang Keyakinan 95% Perbezaan		t	df	Sig. (*2-tailed*)
			Bawah	Atas			
-0.027	0.117	0.0003	-0.028	-0.026	-79.114	117068	0.000

f. Perbincangan Keputusan Penilaian

Algoritma k-min menghasilkan pengelompokan terbaik berbanding algoritma BIRCH dan DBSCAN. Ini kerana sifat algorima k-min yang mudah mengelompok data numerikal yang berdimensi tinggi. Implementasi k-min juga membantu memberikan label pada kelompok