MODEL RAMALAN KEBOLEHPASARAN PELAJAR TAJAAN KERAJAAN MENGGUNAKAN KAEDAH KLASIFIKASI

MOHD RIDHWAN BIN ABDULLAH SANI

UNIVERSITI KEBANGSAAN MALAYSIA

MODEL RAMALAN KEBOLEHPASARAN PELAJAR TAJAAN KERAJAAN MENGGUNAKAN KAEDAH KLASIFIKASI

MOHD RIDHWAN BIN ABDULLAH SANI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH IJAZAH
SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

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

24 Jun 2024

MOHD RIDHWAN BIN ABDULLAH SANI
P127015

PENGHARGAAN

Segala puji bagi Allah SWT atas limpah kurnia-Nya yang telah mengurniakan kekuatan, ketabahan dan ilham kepada saya untuk menyiapkan disertasi bertajuk "Model Ramalan Kebolehpasaran Pelajar Tajaan Kerajaan Menggunakan Kaedah Klasifikasi" ini. Semoga segala usaha dan penat lelah ini mendapat keberkatan dan keredhaan-Nya.

Dengan penuh rasa kesyukuran, saya merakamkan setinggi-tinggi penghargaan dan terima kasih kepada semua pihak di atas sokongan, bimbingan dan dorongan yang diberikan dalam perjalanan menyiapkan disertasi ini. Terutamanya kepada Pensyarah Penyelia, Prof. Madya Dr. Mohd Ridzwan bin Yaakub yang telah memberikan bimbingan, nasihat dan sokongan yang tidak berbelah bahagi sepanjang projek ini dilaksanakan. Tanpa bimbingan dan panduan beliau, projek ini tidak akan dapat disiapkan dengan jayanya.

Saya juga mengucapkan ribuan terima kasih kepada Dekan dan warga Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia (FTSM, UKM) khususnya pensyarah-pensyarah Sarjana Sains Data yang telah memberikan kerjasama, ilmu dan panduan berharga sepanjang pengajian saya. Tidak lupa juga kepada pihak UKM Shape yang telah menawarkan program ini, terima kasih atas bantuan dan sokongan anda.

Penghargaan juga ditujukan kepada Jabatan Perkhidmatan Awam (JPA) khususnya Bahagian Pembangunan Modal Insan (BMI) yang telah memberi peluang kepada saya untuk melanjutkan pengajian di peringkat sarjana melalui Hadiah Latihan Persekutuan (HLP) dan turut membenarkan data penajaan digunakan untuk projek ini. Sokongan dan tajaan ini amat bermakna dan sangat saya hargai.

Tidak dilupakan, saya merakamkan penghargaan yang mendalam kepada isteri saya, Farahazfa binti Mohd Sapari, yang sentiasa memberikan sokongan moral dan dorongan yang tidak putus-putus. Terima kasih juga kepada ibu saya, Hjh. Shariffah binti Endut, anak-anak dan semua ahli keluarga yang sentiasa memberikan galakan dan doa.

Akhir sekali, terima kasih kepada rakan-rakan seperjuangan yang telah bersama-sama melalui pelbagai cabaran dan suka duka sepanjang tempoh pengajian ini. Sokongan anda semua amat saya hargai.

ABSTRAK

Kajian ini bertujuan untuk menangani isu kebolehpasaran graduan tajaan kerajaan dengan membangunkan model ramalan menggunakan kaedah klasifikasi untuk meningkatkan pulangan pelaburan kerajaan dalam penajaan pendidikan. Model ramalan dibangunkan menggunakan data sejarah rekod penajaan JPA dari tahun 2016 sehingga 2022. Sebanyak 48,952 data digunakan yang merangkumi pelbagai universiti dan bidang pengajian merentasi pelbagai negara. Model CRISP-DM digunakan untuk pembangunan model merangkumi pengumpulan data, pra-pemprosesan, pemodelan, penilaian dan penerapan. Sebanyak 15 ciri dikenalpasti untuk pembangunan model iaitu Peringkat Pengajian, Julat Umur, Bidang Perincian NEC, Program Penajaan, Nama Institusi, Bidang NEC, Negara Pengajian, Jenis Institusi, Status Bumiputra, Pekerjaan Bapa, Negeri Tetap, Jantina, Pekerjaan Ibu dan Julat Pendapatan. Lima model klasifikasi dibangunkan iaitu Regresi Logistik, Naïve Bayes, Perhutanan Rawak, Pohon Keputusan dan Mesin Galakan Kecerunan. Hasil kajian menunjukkan bahawa model Perhutanan Rawak mencapai prestasi terbaik dengan Ketepatan 73.65%, Kejituan 72.93%, Sensitiviti 73.66%, Skor F1 sebanyak 70.96% dan skor AUC-ROC sebanyak 76.16%. Analisis mendapati bahawa peringkat pengajian adalah faktor paling berpengaruh, diikuti oleh julat umur, bidang pengajian, program penajaan dan reputasi institusi pendidikan. Faktor pendidikan dan akademik memainkan peranan utama dalam kebolehpasaran graduan, mengatasi faktor demografi dan sosioekonomi. Kajian ini menyumbang kepada pemahaman yang lebih mendalam mengenai faktor-faktor kebolehpasaran pelajar tajaan JPA dan menyediakan panduan praktikal untuk menilai dan meningkatkan keberkesanan program penajaan, menyelaraskan penajaan dengan keperluan industri, serta memaksimumkan peluang pekerjaan graduan. Implikasi kepada dasar termasuk penggunaan hasil kajian untuk merancang intervensi bersasar dan mengoptimakan pelaburan kerajaan dalam pembangunan modal insan negara.

PREDICTION MODEL OF GOVERNMENT-SPONSORED STUDENTS' EMPLOYABILITY USING CLASSIFICATION METHODS

ABSTRACT

This study aims to address the issue of employability among government-sponsored graduates by developing a predictive model using classification methods to enhance the return on investment in government-funded education. The predictive model is developed using historical data from the JPA's sponsorship records from 2016 to 2022. A total of 48,952 records encompassing various universities and fields of study across different countries were used. The CRISP-DM model was employed for model development, which includes data collection, preprocessing, modeling, evaluation, and deployment. Fifteen features were identified for model development, including Level of Education, Age Range, NEC Field Details, Sponsorship Program, Institution Name, NEC Field, Country of Study, Type of Institution , Bumiputra Status, Father's Occupation, Permanent State, Gender, Mother's Occupation and Income Range. Five classification models were developed: Logistic Regression, Naïve Bayes, Random Forest, Decision Tree, and Gradient Boosting Machine. The results showed that the Random Forest model achieved the best performance with an accuracy of 73.65%, precision of 72.93%, recall of 73.66%, F1 score of 70.96% and an AUC-ROC score of 76.16%. The analysis found that the level of education is the most influential factor, followed by age range, field of study, type of sponsorship program, and reputation of the educational institution. Educational and academic factors play a primary role in graduate employability, surpassing demographic and socioeconomic factors. This study contributes to a deeper understanding of the factors affecting the employability of JPA-sponsored students and provides practical guidance for assessing and improving the effectiveness of sponsorship programs, aligning sponsorship with industry needs, and maximizing graduate employment opportunities. Policy implications include using study findings to plan targeted interventions and optimize government investment in human capital development.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

AI	Kecerdasan Buatan
ANN	Rangkaian Neural Buatan
AUC	Luas di Bawah Keluk (*Area Under the Curve*)
BMI	Bahagian Pembangunan Modal Insan
CGPA	Purata Nilai Gred Kumulatif
DT	Pohon Keputusan (*Decision Tree*)
GPA	Purata Nilai Gred
GOT	Bergraduat Menepati Masa
HKA	Had Kecemerlangan Akademik
ICT	Teknologi Maklumat dan Komunikasi (*Information and Communication Technology*)
ILMIA	Institut Maklumat dan Analisis Pasaran Buruh
IPT	Institut Pengajian Tinggi
JPA	Jabatan Perkhidmatan Awam
KNN	K-Nearest Neighbors
KPM	Kementerian Pendidikan Malaysia
KPT	Kementerian Pengajian Tinggi
LR	Regresi Logistik (*Regresi Logistik*)
MARA	Majlis Amanah Rakyat
MMU	Universiti Multimedia Malaysia
NB	Naive Bayes
NEC	Kod Pendidikan Nasional
PTPTN	Perbadanan Tabung Pendidikan Tinggi Nasional
RF	Perhutanan Rawak (*Perhutanan Rawak*)
ROC	Ciri Operasi Penerima (*Receiver Operating Charateristic*)

SPM	Sijil Pelajaran Malaysia
SKPG	Sistem Kajian Pengesanan Graduan
SVM	Mesin Sokongan Vektor (*Support Vector Machine*)
UKM	Universiti Kebangsaan Malaysia
UM	Universiti Malaya
UPM	Universiti Putra Malaysia
USM	Universiti Sains Malaysia
UTM	Universiti Teknologi Malaysia
WEKA	*Waikato Environment for Knowledge Analysis*
XGB	Penggalak Kecerunan Ekstreme / *Extreme Gradient Boosting*

BAB I

PENGENALAN

1.1 PENDAHULUAN

Model ramalan adalah suatu teknik yang digunakan untuk membuat jangkaan atau unjuran mengenai data masa hadapan berdasarkan analisis data sejarah. Teknik ini menerapkan kaedah statistik dan pembelajaran mesin untuk mencari pola dan hubungan dalam data yang akan digunakan untuk meramalkan hasil. Ia memainkan peranan penting dalam proses membuat keputusan berpandukan data. Keputusan berpandukan data dapat meminimakan risiko dan meningkatkan produktiviti dalam sesebuah organisasi (Rustagi & Goel 2022).

Model ramalan telah digunakan secara meluas dalam pelbagai bidang seperti kesihatan, insurans, ekonomi dan pendidikan. Dalam bidang kesihatan, model ini digunakan untuk meramal potensi penyakit seperti risiko penyakit kardiovaskular atau kanser payudara. Ia membantu pengamal perubatan untuk memberikan khidmat nasihat dan rawatan kepada pesakit (Ranapurwala et al. 2019). Dalam industri insurans, model ramalan membantu dalam pengelasan risiko dengan menggunakan data sejarah tuntutan dan keadaan kesihatan semasa pelanggan untuk menetapkan kadar premium yang sesuai (Singla et al. 2020). Begitu juga dalam bidang ekonomi, model ramalan menggunakan teknik statistik yang berasaskan kecerdasan buatan (AI) untuk meramalkan Indeks Harga Pengguna (CPI) yang mencerminkan keadaan ekonomi dan kadar inflasi sesebuah negara (Oh et al. 2021). Manakala dalam bidang akademik pula, model ramalan digunakan untuk meramalkan prestasi akademik, kadar keciciran dan kebolehpasaran pelajar dengan matlamat untuk menambah baik program akademik dan membantu pelajar lemah. Pendekatan ini melibatkan proses pengumpulan dan analisis

pelbagai jenis data berkaitan pelajar seperti demografi, ciri psikologi, prestasi akademik dan latar belakang keluarga (Gafarov et al. 2023).

Contoh-contoh yang dinyatakan ini menggambarkan penggunaan yang luas dan signifikan model ramalan merentasi pelbagai bidang. Sehubungan itu, kajian ini bertujuan untuk membangunkan model ramalan bagi meramal kebolehpasaran pelajar di peringkat pengajian tinggi yang menerima penajaan kerajaan menggunakan kaedah klasifikasi dengan memanfaatkan data penajaan dan data kebolehpasaran di dalam pangkalan data Jabatan Perkhidmatan Awam (JPA). Kajian ini akan memberikan input bermakna kepada penaja untuk menguruskan sumber secara lebih berkesan dan merangka tindakan yang proaktif selaras dengan hasrat kerajaan yang menggalakkan penggunaan digital dan AI dalam urusan kerajaan.

1.2 LATAR BELAKANG

Perbelanjaan kerajaan dalam pendidikan memainkan peranan penting terhadap pembangunan ekonomi negara. Kajian mendapati bahawa perbelanjaan ini memberi kesan positif terhadap pertumbuhan ekonomi dan merupakan salah satu komponen perbelanjaan yang paling berpengaruh di Malaysia (Kamis et al. 2020). Pelaburan berterusan dalam pendidikan adalah penting untuk pembangunan ekonomi Malaysia kerana ia meningkatkan modal insan, menyokong pertumbuhan pendapatan dan menangani isu pengangguran, sekali gus menyumbang kepada kestabilan ekonomi dan pertumbuhan jangka panjang negara.

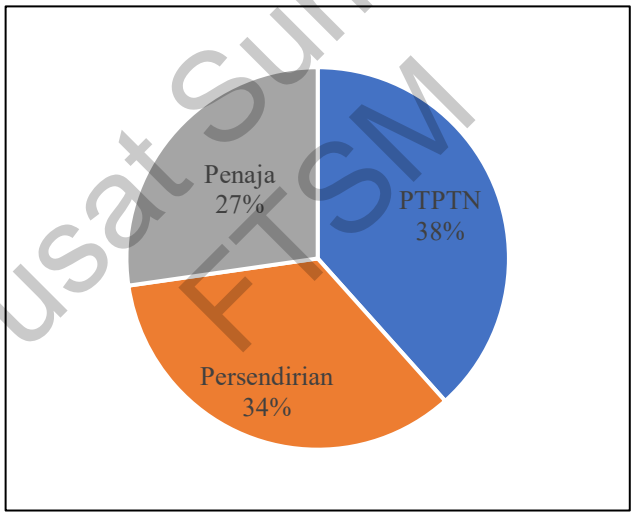
Modal insan merupakan salah satu aset ekonomi dan dikenal pasti sebagai pembolehubah paling penting untuk pertumbuhan ekonomi jangka panjang di Malaysia (Et.al 2021). Modal insan dianggap sebagai tonggak utama untuk meningkatkan taraf hidup. Sesebuah negara sukar untuk mencapai kemajuan tanpa modal insan yang unggul, berpengetahuan luas dan berkemahiran tinggi. Modal insan yang berpendidikan tinggi juga diakui mempunyai hubungan yang positif dengan kadar pertumbuhan ekonomi (Abdullah & Abd Majid 2022).

Pelaburan ke atas modal insan sangat penting kerana pengetahuan dan kemahiran seseorang tidak boleh dipindahkan daripada seorang individu kepada

individu yang lain tetapi memerlukan usaha untuk memperolehnya. Pelaburan modal insan meliputi penajaan pendidikan, latihan pra-pekerjaan, latihan pembangunan profesionalisme dan mobilasi bagi meningkatkan taraf hidup seseorang individu. (Abdullah & Abd Majid 2022).

1.2.1 Penajaan Pendidikan

Penajaan pendidikan merupakan pelaburan strategik kerajaan dalam pembangunan modal insan negara. Matlamat penajaan adalah untuk memberi peluang pendidikan yang sama rata disamping meningkatkan bilangan pakar dan profesional di dalam negara. Penajaan ini mempunyai fokus dan tujuan tersendiri sama ada sebagai obligasi sosial untuk mendukung agenda pembangunan negara ataupun berfokus untuk perancangan sumber manusia akan datang (Ab Hamid 2021).



Rajah 1.1 Pembiayaan Pendidikan di Malaysia

Sumber: Laporan Kajian Pengesanan Graduan 2018 - 2022

Berdasarkan Rajah 1.1, 38% graduan di Malaysia mendapat pembiayaan pendidikan melalui Perbadanan Tabung Pendidikan Tinggi Nasional (PTPTN), 34% secara persendirian dan 27% melalui penajaan pendidikan. Terdapat banyak penaja di Malaysia yang menawarkan tajaan kepada pelajar di peringkat pengajian tinggi seperti Jabatan Perkhidmatan Awam (JPA), Kementerian Pengajian Tinggi (KPT), Majlis Amanah Rakyat (MARA), Kementerian Pelajaran Malaysia (KPM) dan Kerajaan Negeri melalui yayasan pendidikan yang ditubuhkan.

Berdasarkan Buku Arahan Pentadbiran Bersepadu Bahagian Pembangunan Modal Insan 2022 yang dikeluarkan oleh JPA, penajaan ertinya pembiayaan bagi pengajian sama ada dilaksanakan menggunakan kaedah pemberian biasiswa, pinjaman, dermasiswa ataupun tajaan mengikuti sesuatu kursus dengan kadar elaun kelayakan tertentu dan terikat dengan syarat-syarat yang ditetapkan. Ia merupakan bantuan dalam bentuk kewangan seperti bayaran yuran pengajian, elaun sara hidup bulanan, elaun pakaian panas, elaun penempatan, elaun buku dan alat perkakasan. Mengambil kira kos penajaan yang tinggi, adalah menjadi keutamaan kepada pihak penaja untuk memastikan pulangan pelaburan adalah selaras dengan objektif penajaan bagi memastikan kepentingan semua pihak dipenuhi serta meneruskan kesinambungan penajaan di masa akan datang.

1.2.2 Bahagian Pembangunan Modal Insan

JPA menerusi Bahagian Pembangunan Modal Insan (BMI) telah diberi mandat untuk melaksanakan fungsi penajaan pendidikan kepada pelajar-pelajar cemerlang lepasan Sijil Pelajaran Malaysia (SPM) dan juga pegawai awam untuk mengikuti pengajian dalam pelbagai peringkat di institusi pengajian tinggi (IPT) dalam dan luar negara. Peranan ini adalah selaras dengan peruntukan Fasal 153 Perlembagaan Persekutuan dan Perintah Menteri-Menteri Kerajaan Persekutuan yang meletakkan perkara-perkara berkaitan penajaan (biasiswa persekutuan dan latihan) adalah di bawah bidang tugas JPA dan tanggungjawab Ketua Pengarah Perkhidmatan Awam untuk melaksanakannya.

Saban tahun BMI menaja hampir 10,000 orang pelajar untuk melanjutkan pengajian di peringkat pengajian tinggi. Objektif penajaan ini adalah untuk memastikan keperluan strategik sektor awam dan negara dapat dipenuhi melalui latihan dan program penajaan yang dinamik. Secara kasarnya, jumlah tajaan kerajaan bagi seorang pelajar di dalam negara adalah sekitar RM22,000.00 – RM40,000.00 manakala tajaan di luar negara adalah sekitar RM300,000.00 – RM500,000.00 tertakluk kepada bidang dan institusi pengajian. Berdasarkan rekod penajaan JPA, purata perbelanjaan yang dikeluarkan kerajaan kepada pelajar tajaan yang bergraduat pada tahun 2019 sehingga tahun 2022 adalah sekitar RM835.1 juta setahun.

Mengambil kira implikasi kewangan yang besar dalam usaha membangunkan modal insan, JPA telah menetapkan beberapa perkara sebagai ukuran dan penanda aras terhadap keberkesanan program penajaan yang dilaksanakan. Perkara-perkara yang dinilai adalah kecemerlangan akademik, peratusan pelajar tamat pengajian tanpa pelanjutan dan kebolehpasaran graduan yang ditaja. Perkara yang dinilai ini adalah selaras dengan objektif penajaan dan juga untuk memastikan pulangan pelaburan kerajaan dapat dimaksimumkan.

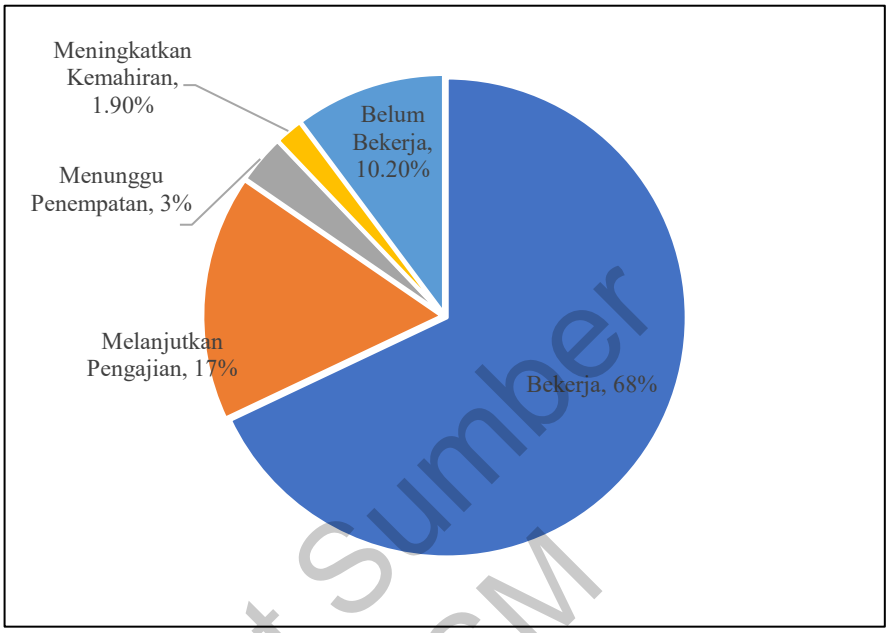
1.2.3 Kebolehpasaran Graduan

Pemahaman terhadap kebolehpasaran graduan adalah aspek kritikal dalam memahami dinamika pasaran pekerjaan di Malaysia yang sentiasa berubah. Menurut Institut Maklumat dan Analisis Pasaran Buruh (ILMIA), dinamika ini mencerminkan kepelbagaian cabaran dan peluang dalam konteks pekerjaan di negara ini. Berdasarkan kepada *The National Employability Blueprint 2012 – 2017*, kebolehpasaran graduan bermaksud keupayaan untuk dipasarkan di dalam industri atau mempunyai kecekapan bagi mendapat pekerjaan dan mengekalkan pekerjaan yang bersesuaian. Ianya dipengaruhi oleh banyak faktor melangkaui kelayakan akademik seperti kemahiran, pengetahuan, sifat peribadi dan kemampuan untuk meneroka pasaran pekerjaan secara efektif (Maaliw et al. 2022).

Kadar kebolehpasaran graduan berupaya memberikan penunjuk tentang keupayaan ekosistem pendidikan tinggi negara dalam mengeluarkan graduan yang berkualiti serta memenuhi permintaan tenaga kerja dari pihak industri. Ia menjadi faktor signifikan dalam menaikkan reputasi sesebuah institusi (Haque et al. 2024). Semakin tinggi kadar kebolehpasaran graduan, semakin tinggi kualiti graduan yang dihasilkan oleh IPT yang berupaya memenuhi permintaan semasa industri. Selain itu juga, ia boleh menjadi ukuran terhadap keberkesanan program penajaan pendidikan.

Bagi mengukur kadar kebolehpasaran graduan di Malaysia, KPT telah melaksanakan kajian pengesanan graduan sejak tahun 2006 bagi mengesan status kebolehpasaran graduan selepas mereka menamatkan pengajian melalui Sistem Kajian Pengesanan Graduan (SKPG). Rajah 1.2 menunjukkan ukuran kebolehpasaran graduan di Malaysia berdasarkan Laporan Kajian Pengesanan Graduan 2022. Kadar

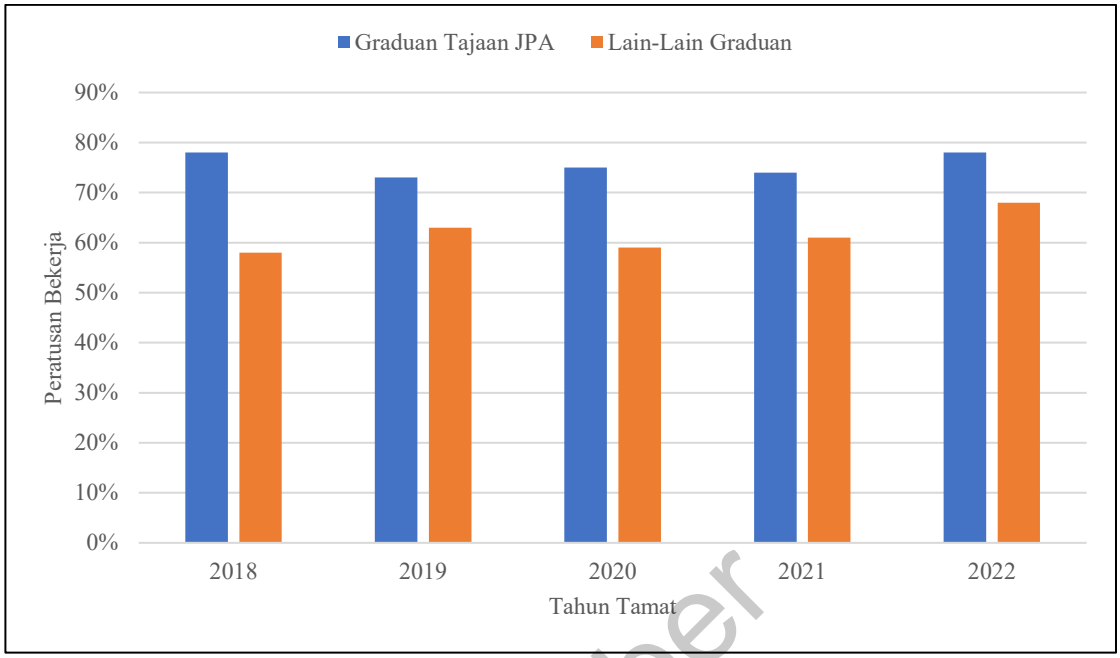
kebolehpasaran graduan adalah 89.8% merangkumi graduan telah bekerja (67.9%), graduan melanjutkan pengajian (16.6%), graduan menunggu penempatan pekerjaan (3.3%) dan graduan yang sedang mengikuti pelbagai program meningkatkan kemahiran (1.9%). Selebihnya iaitu 10.2% graduan masih belum bekerja.



Rajah 1.2 Kebolehpasaran Graduan di Malaysia

Sumber: Laporan Kajian Pengesanan Graduan 2022

Kajian ini menjadi rujukan dan penanda aras oleh kebanyakan agensi dan IPT. Walau bagaimanapun kajian ini adalah terhad kepada graduan yang menamatkan pengajian di dalam negara sahaja, tidak termasuk graduan-graduan Malaysia yang menamatkan pengajian di luar negara. Sehubungan itu, JPA selaku penaja turut melaksanakan kajian pengesanan graduan untuk mendapatkan gambaran keseluruhan kebolehpasaran graduan tajaan JPA. Berbanding kajian yang dibuat oleh KPT, kebolehpasaran graduan tajaan JPA hanya merujuk kepada graduan yang telah bekerja sahaja supaya selaras dengan objektif penajaan seperti yang telah dinyatakan dalam Bab 1.2.2. Perbandingan peratusan bekerja bagi graduan tajaan JPA dan lain-lain graduan di Malaysia adalah seperti Rajah 1.3. Berdasarkan rajah ini didapati bahawa graduan tajaan JPA mempunyai peratusan bekerja yang lebih baik berbanding lain-lain graduan yang tidak mendapat tajaan JPA bagi tempoh 5 tahun kebelakangan ini.



Rajah 1.3 Perbandingan Peratusan Bekerja Graduan Tajaan JPA dan Lain-Lain Graduan

Sumber : Kajian Pengesanan Graduan JPA dan Laporan SKPG

1.3 PENYATAAN MASALAH

Kerajaan melalui JPA khususnya telah mengeluarkan perbelanjaan yang besar untuk menaja pelajar-pelajar cemerlang dengan harapan supaya bakat terbaik ini akan menyumbang bakti kepada negara sejurus menamatkan pengajian. Data terkini menunjukkan bahawa kadar kebolehpasaran graduan tajaan JPA secara purata adalah sekitar 75% dalam tempoh lima tahun terakhir. Walaupun peratusan ini lebih baik berbanding lain-lain graduan secara keseluruhannya, tumpuan perlu diberikan kepada baki 25% graduan yang tidak mendapat pekerjaan ini kerana mereka merupakan kelompok pelajar cemerlang dari segi akademik, sahsiah dan mendapat bantuan kewangan daripada kerajaan. Kegagalan mereka untuk memperolehi kerja selepas menamatkan pengajian merupakan kerugian kepada pelaburan kerajaan dan pembaziran bakat (Haque et al. 2024).

Sehubungan itu, untuk meningkatkan kebolehpasaran pelajar tajaan, tindakan intervensi awal perlu dilakukan oleh penaja untuk mengenalpasti punca kegagalan pelajar-pelajar ini mendapatkan pekerjaan walaupun terdiri daripada kelompok pelajar cemerlang. Kegagalan ini boleh berpunca daripada kesilapan program penajaan,

kesilapan pemilihan pelajar atau faktor-faktor lain daripada luar. Tindakan intervensi ini perlulah dilaksanakan lebih awal iaitu semasa pemilihan pelajar untuk ditaja dan sebelum pelajar menamatkan pengajian.

Pada masa ini, kebolehpasaran graduan hanya diukur selepas pelajar menamatkan pengajian iaitu melalui analisis deskriptif berbanding ramalan. Maklumat ini adalah sangat berharga untuk membangunkan model ramalan bagi mengenalpasti kelompok pelajar yang menghadapi kesukaran untuk mendapatkan pekerjaan. Maklumat ini boleh membantu penaja untuk merangka strategi dan intervensi yang perlu untuk memperbaiki prospek kerjaya pelajar tajaan (Haque et al. 2024).

Selain itu, pembangunan model ramalan boleh membantu penaja untuk melaksanakan program intervensi bersasar untuk meningkatkan kebolehpasaran pelajar. Pada masa ini, program intervensi pelajar seperti program pemadanan pekerjaan, karnival kerjaya, program sangkutan industri dan jaringan industri dibuat secara umum dan bergantung kepada permintaan industri. Ini menyebabkan hanya kelompok pelajar tertentu yang menerima manfaat dan mudah menembusi pasaran pekerjaan. Melalui program intervensi bersasar, kelompok pelajar yang dikenalpasti sukar mendapatkan pekerjaan berdasarkan model ramalan dapat dikenalpasti dan dibantu lebih awal (K & H K 2020). Tindakan proaktif ini mampu membantu meningkatkan kebolehpasaran pelajar dan juga memastikan objektif penajaan dapat dicapai.

1.4 JURANG PENYELIDIKAN

Kebanyakan model ramalan yang dibangunkan sebelum ini khususnya di Malaysia menggunakan data yang diambil daripada pangkalan data SKPG semata-mata. Kebergantungan kepada data SKPG ini menyebabkan model ramalan sebelum ini tidak dapat digeneralisasikan memandangkan terdapat atribut-atribut yang digunakan hanya diperolehi selepas pelajar menamatkan pengajian dan cenderung kepada institusi atau bidang tertentu sahaja. Sebagai contohnya, Rahman et al. (2017) menggunakan data SKPG graduan UTM, UKM, USM, UM dan UPM sahaja, Othman et al. (2018) pula membangunkan model ramalan menggunakan data SKPG graduan UKM sahaja dan Haque et al. (2024) pula membangunkan model dari data SKPG graduan MMU sahaja.

Manakala Maaliw et al. (2022) pula memfokuskan kepada graduan bidang kejuruteraan elektronik sahaja, Raman & Pramod (2022) kepada graduan Sarjana Pentadbiran Perniagaan dan Baffa et al. (2023) kepada graduan Sains Komputer sahaja.

Tiada lagi model ramalan yang dibangunkan menggunakan pangkalan data penaja untuk meramal kebolehpasaran. Pangkalan data penaja menyimpan banyak data merangkumi data pelajar, data akademik, data demografi, data penajaan dan data sosioekonomi keluarga. Data ini telah dikumpulkan semasa pelajar memohon untuk mendapatkan penajaan dan menandatangani perjanjian. Memanfaatkan data penajaan ini boleh membantu penaja meramalkan kebolehpasaran pelajar lebih awal sebelum pelajar menamatkan pengajian. Ia boleh membantu penaja merangka strategi lebih efektif dalam membantu pelajar sedia ada dan pemilihan pelajar baru di masa akan datang.

Kajian ini juga melibatkan atribut-atribut yang lebih meluas dan merangkumi pelbagai bidang pengajian, universiti, peringkat, dan negara. Ini berbeza daripada kajian-kajian sebelumnya yang cenderung menumpukan kepada satu bidang pengajian atau universiti tertentu. Penerokaan kepada kepada atribut lebih pelbagai dan lebih generik akan memberikan pemahaman yang lebih holistik tentang faktor-faktor yang mempengaruhi kebolehpasaran graduan, memberikan gambaran yang lebih jelas dan komprehensif kepada penaja dalam membuat keputusan penajaan.

1.5 PERSOALAN KAJIAN

Persoalan yang perlu dijawab dalam kajian ini untuk membangunkan model ramalan kebolehpasan pelajar tajaan kerajaan adalah seperti berikut:

1. Apakah data yang perlu dikumpulkan dan dilombong untuk membangunkan model ramalan kebolehpasaran pelajar tajaan kerajaan?
2. Apakah teknik atau model pengelasan berasaskan pembelajaran mesin yang terbaik untuk meramal kebolehpasaran pelajar tajaan kerajaan?
3. Apakah faktor penentu yang mempengaruhi kebolehpasaran pelajar tajaan kerajaan?

1.6 OBJEKTIF KAJIAN

Tujuan kajian ini adalah untuk meramalkan kebolehpasaran pelajar tajaan kerajaan dengan memanfaatkan data penajaan JPA menggunakan model ramalan klasifikasi. Bagi mencapai tujuan ini, objektif kajian ditetapkan seperti berikut:

1. Mengenalpasti dan memilih atribut penentu daripada data penajaan untuk membangunkan model ramalan.
2. Membangunkan model ramalan kebolehpasaran pelajar tajaan kerajaan menggunakan kaedah klasifikasi.
3. Mengenalpasti faktor penentu kebolehpasaran pelajar tajaan JPA berdasarkan ciri yang digunakan dalam model ramalan.

1.7 SKOP KAJIAN

Skop kajian ditetapkan untuk memberi fokus pada kajian supaya tidak tersasar dari tujuan kajian dan mencapai objektifnya. Beberapa batasan kajian telah dikenalpasti iaitu:

1. Kajian ini menggunakan data pengesanan graduan tajaan JPA dari tahun 2016 sehingga 2022 yang dilabelkan sebagai bekerja dan tidak bekerja sahaja.
2. Maklumat graduan diperolehi daripada pangkalan data JPA melalui Sistem eSILA sahaja merangkumi maklumat peribadi, maklumat penajaan, maklumat akademik, maklumat perhubungan dan maklumat waris.
3. Kajian ini menggunakan garis panduan dalaman JPA iaitu Arahan Pentadbiran Bersepadu Bahagian Pembangunan Modal Insan 2022 sebagai rujukan polisi berkaitan penaja.
4. Model yang dibangunkan adalah terhad untuk meramal kebolehpasaran pelajar tajaan JPA sahaja.

1.8 KEPENTINGAN KAJIAN

Kajian ini penting untuk memberikan panduan yang lebih tepat dan relevan kepada penaja untuk membuat keputusan penajaan yang strategik dan berkesan. Melalui model ramalan yang tepat, penaja dapat mengoptimumkan pelaburan mereka dan meningkatkan kebolehpasaran pelajar. Model ramalan ini juga membantu mengenal pasti pelajar yang berisiko tinggi mengalami kesukaran dalam mendapatkan pekerjaan, membolehkan intervensi yang lebih tepat dan berkesan.

Ramalan kebolehpasaran pelajar untuk tahun-tahun akan datang menyediakan data penting bagi perancangan jangka panjang, termasuk perancangan sumber manusia, pembangunan ekonomi, dan strategi industri. Penggunaan teknik perlombongan data dan model klasifikasi dalam kajian ini turut menyumbang kepada pengetahuan dalam bidang ini dan mendorong penggunaan sains data dalam konteks pendidikan.

Dengan pemahaman yang lebih mendalam mengenai kebolehpasaran pelajar, penaja dapat memaksimumkan pulangan dan menyokong pembangunan modal insan negara. Pembangunan model ramalan yang tepat memastikan kebolehpasaran graduan melepasi purata semasa, meningkatkan keberkesanan program tajaan pendidikan oleh kerajaan. Kajian ini juga menyumbang kepada pembangunan ekonomi dan sosial negara melalui perancangan yang lebih baik dan penggunaan data yang lebih efektif.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Bab ini memerihalkan kajian-kajian lepas yang telah dibuat berkaitan dengan ramalan kebolehpasaran graduan untuk dijadikan sebagai panduan dalam kajian ini.

2.2 FAKTOR MEMPENGARUHI KEBOLEHPASARAN GRADUAN

Kebolehpasaran graduan di Malaysia dipengaruhi oleh pelbagai faktor yang kompleks seperti kemahiran insaniah, prestasi akademik, penguasaan bahasa dan keadaan ekonomi sekitaran. Kemahiran insaniah seperti komunikasi, kemahiran menganalisa, kerja berpasukan dan nilai positif adalah sangat penting dan mempengaruhi kebolehpasaran graduan sehingga 85.5%. Hasil ini diperolehi melalui suatu kajian yang memetakan kemahiran insaniah kepada status pekerjaan di kalangan graduan (Basir et al. 2022). Malah terdapat kajian lain yang menyatakan bahawa sifat peribadi dan kemahiran khusus seperti kemahiran menyelesaian masalah dan kepimpinan juga mempengaruhi kebolehpasaran graduan secara langsung (Rogis et al. 2022).

Prestasi akademik mempunyai impak yang signifikan terhadap kebolehpasaran graduan di Malaysia, tetapi ia bukanlah penentu tunggal. Kajian mendapati, walaupun memperoleh purata nilai gred kumulatif (CGPA) yang tinggi adalah faktor positif, ia mesti dilengkapi dengan kemahiran dan kompetensi lain untuk bersaing dalam pasaran kerja. Ketidaksepadanan antara kemahiran yang diajar di universiti dan kemahiran yang diperlukan oleh majikan mencadangkan bahawa prestasi akademik sahaja tidak mencukupi untuk mendapatkan pekerjaan. Ini disokong oleh dapatan yang menyatakan bahawa pelajar dan pensyarah sering terlebih nilai tentang kesediaan pelajar untuk

memasuki dunia pekerjaan dengan lebih menumpukan pada pencapaian akademik berbanding kemahiran praktikal (Ong et al. 2022).

Penguasaan Bahasa Inggeris memainkan peranan penting dalam kebolehpasaran graduan terutamanya dalam konteks Asia Tenggara. Kajian mendapati kelemahan berkomunikasi dalam Bahasa Inggeris merupakan faktor utama pengangguran dikalangan graduan di mana sektor swasta menekankan penguasaan Bahasa Inggeris untuk meningkatkan kecekapan dan produktiviti (Jawing & Kamlun 2022). Kajian menunjukkan terdapat korelasi yang signifikan di antara kemahiran Bahasa Inggeris dengan kebolehpasaran di mana graduan yang mahir berkomunikasi dalam Bahasa Inggeris lebih mudah mendapatkan pekerjaan (Hiew et al. 2021). Ini juga menjadi pendorong kepada syarikat swasta untuk memberi keutamaan kepada calon yang berpendidikan luar negara kerana kemahiran berbahasa dan perspektif global yang dibawa mereka.

Keadaan ekonomi turut memberi impak yang signifikan terhadap kebolehpasaran graduan di Malaysia terutamanya melalui interaksi antara kemelesetan ekonomi, permintaan industri dan ketidaksepadanan antara kemahiran graduan dengan keperluan pasaran. Ketidakstabilan ekonomi juga menekan graduan untuk mempersiapkan diri dengan kemahiran lain seperti ketahanan, pengurusan masa dan pengalaman penyelidikan untuk lebih kompetetif dalam pasaran pekerjaan (Abd Majid et al. 2020). Selain itu, faktor demografi seperti lokasi, jantina dan pendapatan keluarga juga mempengaruhi kebolehpasaran graduan (Tengku Mohamed & Lee 2021).

Peluang pekerjaan mengikut kawasan adalah berbeza-beza dengan negeri-negeri seperti Kedah, Perak, dan Selangor menawarkan lebih banyak prospek pekerjaan yang boleh mempengaruhi strategi pencarian pekerjaan graduan dari pelbagai jurusan (Sani & Jamil 2019).

Satu kajian mendapati bahawa graduan wanita menghadapi kadar pengangguran yang lebih tinggi berbanding lelaki, merangkumi 54.2% daripada jumlah graduan yang menganggur pada tahun 2022 (Sarhan & Ab. Aziz 2023). Diskriminasi dalam pasaran

pekerjaan jelas kelihatan, di mana wanita perlu mengemukakan lebih banyak permohonan untuk menerima jumlah panggilan temu duga yang sama seperti lelaki. Graduan wanita juga cenderung menerima gaji yang lebih rendah berbanding lelaki, terutamanya pada peringkat yang lebih tinggi walaupun mempunyai kelayakan yang sama. Dalam bidang vokasional pula, graduan lelaki dianggap mempunyai kecekapan yang lebih tinggi dalam pekerjaan berbanding graduan wanita sekaligus mempengaruhi prospek pekerjaan mereka (Poon & Leeves 2022).

Pendapatan keluarga turut menjadi faktor kepada kebolehpasaran graduan di Malaysia sama ada secara langsung atau tidak langsung. Pendapatan keluarga yang tinggi memberi akses yang lebih baik kepada sumber pengetahuan yang boleh meningkatkan prestasi akademik dan penguasaan kemahiran sekaligus meningkatkan prospek kebolehpasaran. Satu kajian telah mengenal pasti pendapatan keluarga sebagai faktor penting yang mempengaruhi kebolehpasaran graduan, bersama-sama dengan pembolehubah lain seperti jantina, CGPA dan kemahiran komunikasi (Basir et al. 2022). Selain itu, pendapatan keluarga juga mempengaruhi kebarangkalian graduan terlibat dalam aktiviti keusahawanan kerana kestabilan kewangan boleh memberikan modal dan toleransi risiko yang diperlukan untuk memulakan perniagaan. Ini disokong oleh dapatan bahawa keluarga memainkan peranan yang lebih penting daripada universiti dalam membentuk persepsi keusahawanan di kalangan pelajar (Shahadan et al. 2019).

Jenis universiti dan bidang pengajian juga mempengaruhi faktor kebolehpasaran graduan dengan bidang dan institusi tertentu memberikan prospek pekerjaan yang lebih baik (Tengku Mohamed & Lee 2021). Pandemik COVID-19 telah meningkatkan cabaran dalam mencari pekerjaan kerana wujudnya persaingan antara graduan dan pekerja yang diberhentikan. Graduan memerlukan inisiatif tambahan seperti menghadiri program peningkatan kemahiran untuk meningkatkan prospek pekerjaan (Rahman et al. 2022). Selain itu, trend yang membimbangkan di mana pelajar lepasan sekolah memilih untuk tidak melanjutkan pendidikan tinggi dan keperluan institusi pendidikan tinggi untuk lebih selari dengan keperluan industri semakin merumitkan landskap kebolehpasaran (Baharin & Abdullah 2022).

Jadual 2.1 Faktor Mempengaruhi Kebolehpasaran Graduan di Malaysia

Bil.	Faktor	**Perincian**
1.	Faktor Peribadi	Kemahiran insaniah, kemahiran bahasa, pengalaman dan ciri kepimpinan.
2.	Faktor Demografi	Umur, jantina, lokasi tempat tinggal dan bidang pengajian
3.	Faktor Persekitaran	Keadaan ekonomi, sosioekonomi dan polisi kerajaan

Jadual 2.1 menunjukkan ringkasan faktor yang mempengaruhi kebolehpasaran graduan di Malaysia. Berdasarkan kajian terhadap faktor kebolehpasaran graduan, dapat diringkaskan bahawa kebolehpasaran graduan dipengaruhi oleh faktor peribadi, faktor demografi dan faktor luaran. Faktor peribadi adalah seperti prestasi kemahiran insaniah, Faktor demografi pula murujuk kepada umur, jantina, lokasi tempat tinggal dan bidang pengajian. Manakala faktor persekitaran yang mempengaruhi kebolehpasaran ialah keadaan ekonomi, sosioekonomi dan polisi kerajaan. Memahami faktor-faktor ini adalah penting semasa memilih atribut untuk pembangunan model ramalan.

2.3 MODEL KLASIFIKASI

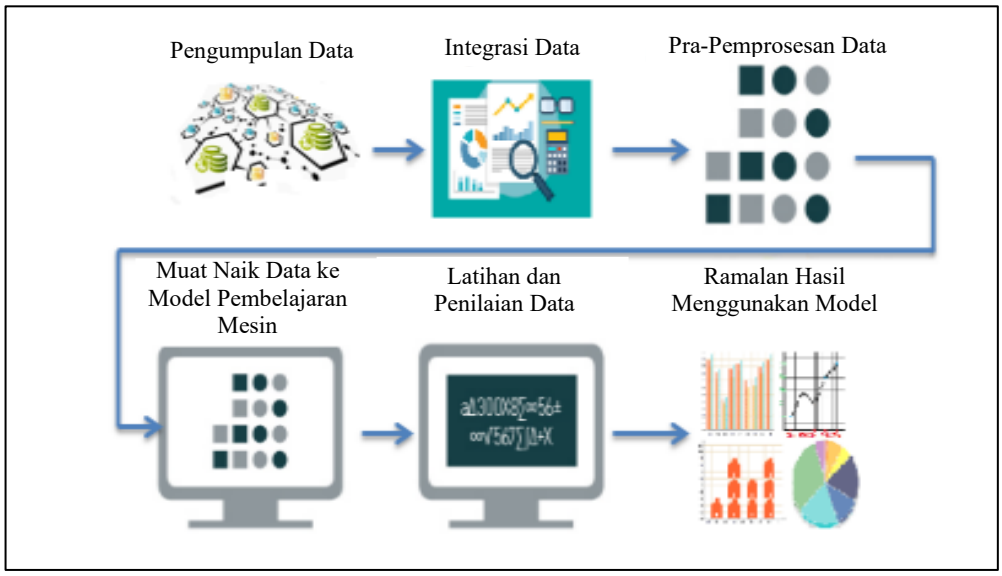
Klasifikasi adalah proses untuk mengkategorikan data ke dalam kumpulan yang ditetapkan berdasarkan ciri-ciri mereka. Kumpulan ini sering dirujuk sebagai sasaran, label, kelas atau kategori. Matlamat utama klasifikasi adalah untuk mencari kumpulan atau sasaran yang sesuai dengan data baru (K & H K 2020). Dalam konteks kajian ini, model klasifikasi digunakan untuk mengkelaskan kebolehpasaran bakal graduan sama ada akan bekerja atau tidak bekerja. Proses pengelasan biasanya melibatkan beberapa langkah iaitu pengambilan data, perlombongan data, analisis ciri, pemilihan algoritma pembelajaran mesin, latihan model, pengesahan dan ramalan.

Perlombongan data merupakan proses penting dalam mencari dan mengenalpasti corak dan pengetahuan dari data yang besar. Perlombongaan data menganalisis data-data yang lalu bagi tujuan meramal masa depan. Perlombongan data melibatkan pelbagai bidang iaitu teknologi pangkalan data, statistik, visualisasi maklumat dan pembelajaran mesin dan kecerdasan buatan. Ia melibatkan banyak

tugasan seperti konsep penerangan, pengelasan dan ramalan, analisis statistik, analisis data terasing, analisis tren, regresi dan sebagainya (Mumtaz et al. 2023).

Pembelajaran mesin adalah bidang saintifik yang menggunakan model statistik dan algoritma untuk melaksanakan tugas oleh sistem tanpa arahan eksplisit, sebaliknya bergantung kepada inferens dan corak. Oleh itu, pembelajaran mesin beroperasi dalam sistem yang mampu mengenal pasti dan memahami corak data dan menggunakannya untuk membuat keputusan secara autonomi (Haque et al. 2024). Dengan mempelajari data yang telah diberikan kepada sesebuah program, pembelajaran mereka dari semasa ke semasa akan ditingkatkan (Black et al. 2023).

Pembelajaran mesin memberikan kemampuan baru secara menyeluruh kepada sistem komputer. Rajah 2.1 menunjukkan gambaran keseluruhan pembelajaran mesin dalam model ramalan. Data mentah dikumpulkan dan digabungkan dari pelbagai sumber. Data yang diperolehi ini kemudiaanya melalui pra-pemprosesan data untuk pembersihan, penyeragaman dan visualisasi yang membantu dalam memahami hubungan antara parameter-parameter tersebut. Data yang telah diproses ini dimuat naik ke dalam model pembelajaran mesin bagi tujuan latihan dan mengembangkan model bagi tujuan ramalan dan penilaian. Akhirnya model yang telah dilatih ini digunakan untuk meramal data baru (K & H K 2020).



Rajah 2.1 Gambaran Pembelajaran Mesin

Sumber: K & H K 2020

Pemilihan algoritma sering bergantung pada jenis data sama ada data nominal, ordinal, nisbah atau interval yang akan digunakan. Pelbagai algoritma boleh digunakan untuk tugas klasifikasi seperti LR yang sesuai untuk masalah klasifikasi binari atau model yang lebih kompleks seperti DT, RF, ANN, SVM dan NB (Rahman et al. 2017).

2.3.1 Regresi Logistik

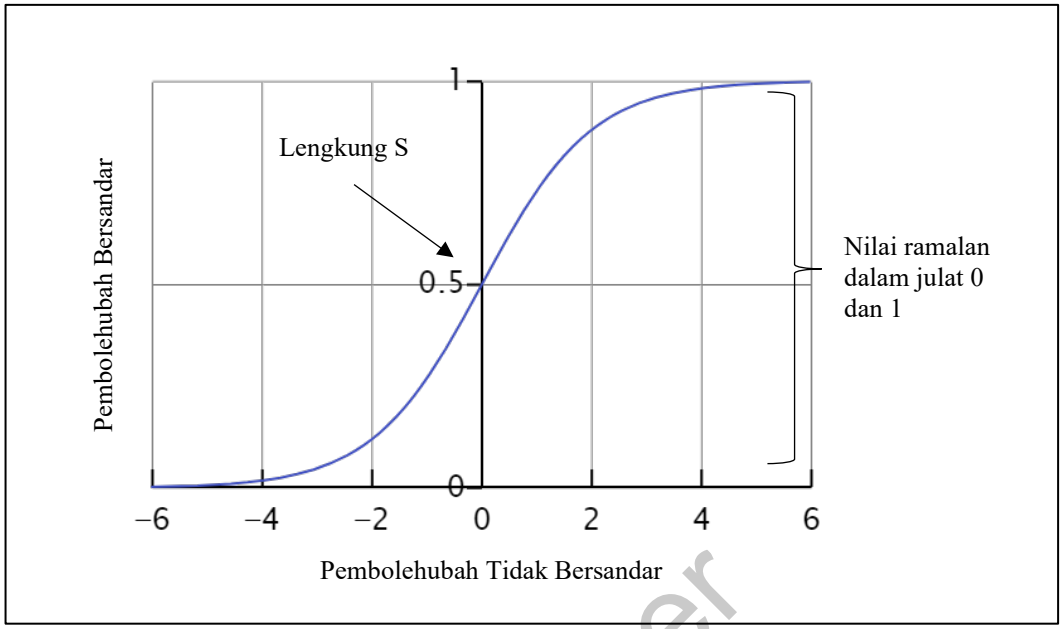
Regresi Logistik (LR) merupakan algoritma pembelajaran mesin yang banyak digunakan terutamanya untuk tugas pengelasan binari. Objektif LR adalah untuk mencari hubungan yang terbaik antara pembolehubah bersandar dan set pembolehubah tidak bersandar. Ia memodelkan kebarangkalian hasil binari berdasarkan satu atau lebih pembolehubah peramal dengan memadankan fungsi logistik kepada data, yang menukar nombor nilai sebenar menjadi nilai antara 0 dan 1. Fungsi logistik ditransformasi melalui fungsi sigmoid untuk memberikan kebarangkalian hasil.

$$P\,(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k)}} \qquad \ldots(2.1)$$

Di mana:

- P adalah kebarangkalian hasil.
- β_0 adalah pemalar (pintasan)
- $\beta_1, \beta_1, \ldots, \beta_k$ adalah koefisien regresi
- *X1, X2,..., Xk* adalah pembolehubah tidak bersandar (ciri)

Persamaan 2.1 menunjukkan fungsi sigmoid yang digunakan di dalam LR iaitu nilai input disatukan secara linear menggunakan nilai pemberat atau koefisien untuk meramalkan hasil. Fungsi sigmoid ini memastikan bahawa kebarangkalian yang diramalkan berada dalam julat 0 hingga 1, dengan bentuk yang ditunjukkan seperti Rajah 2.2.



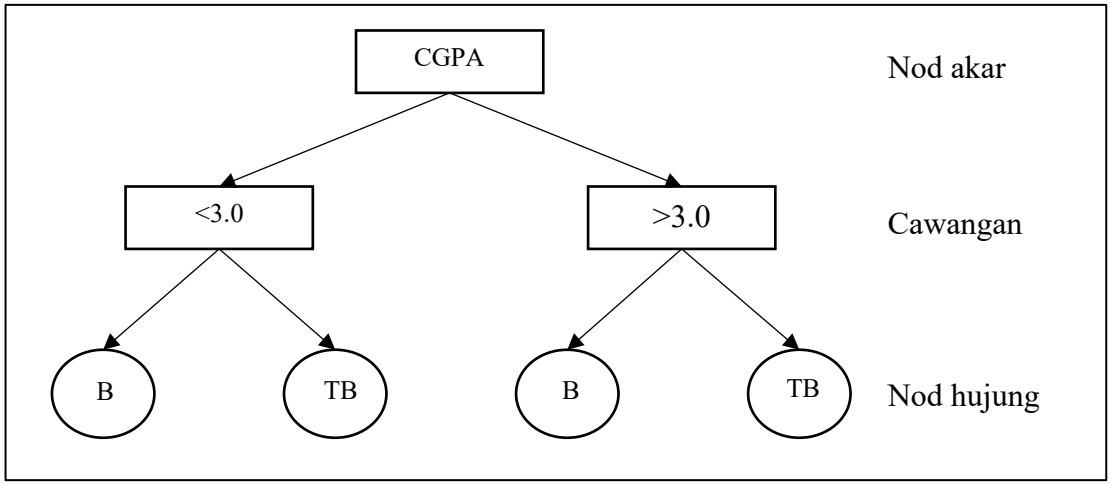
Rajah 2.2 Fungsi Sigmoid

Sumber: Zaidi & Al Luhayb 2023

Algoritma ini sering digunakan oleh penyelidik kerana mudah ditafsirkan, ringkas dan mudah diproses (Liu 2023). Ia juga fleksibel dan boleh digunakan untuk pelbagai jenis data menjadikannya sesuai untuk anggaran kebarangkalian (Zaidi & Al Luhayb 2023). Kelemahan algoritma ialah sensitif terhadap nilai terpencil, hanya sesuai untuk hasil binari dan kurang berkesan untuk data besar (K & H K 2020).

2.3.2 Pohon Keputusan

Pohon Keputusan (DT) adalah model pembelajaran mesin yang banyak digunakan untuk menunjukkan kemungkinan bagi setiap keputusan dalam struktur seperti pokok. Ia menggunakan peraturan *IF-THEN* untuk membina struktur cawangan dan nod berdasarkan nilai ciri input (Salame 2023). DT beroperasi dengan membahagikan set data secara berulang kepada subset dengan tujuan mewujudkan kumpulan homogen terhadap pembolehubah sasaran (Slavutskaya 2023). Ia mengandungi nod akar, cawangan dan nod hujung. Setiap nod akar mewakili ujian atas ciri, setiap cawangan mewakili hasil ujian dan setiap nod hujung menyimpan label kelas. Rajah 2.3 merupakan konsep DT dalam konteks kebolehpasaran graduan sama ada bekerja atau tidak bekerja berdasarkan pencapaian CGPA (Madhavi Girase 2018).



Rajah 2.3 Konsep Pohon Keputusan

Sumber: Madhavi Girase 2018

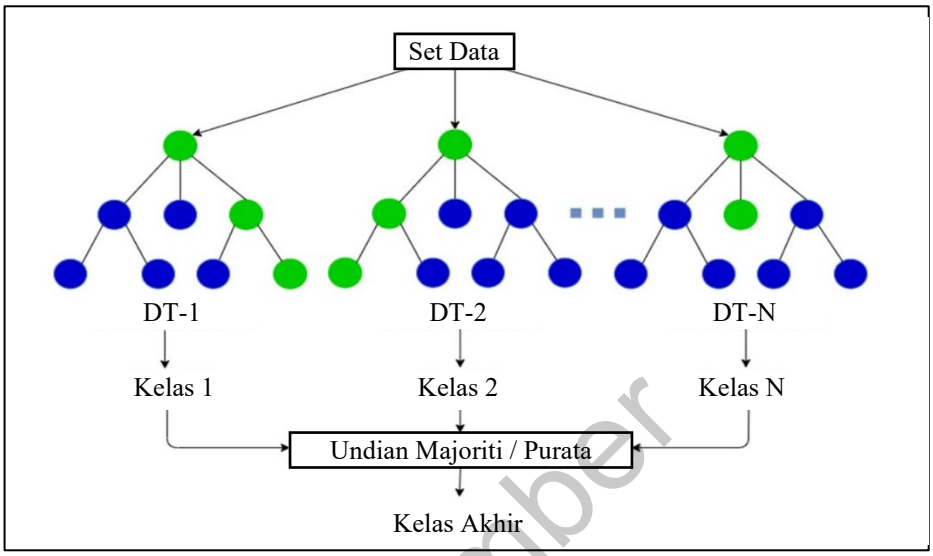
DT boleh digunakan untuk memahami ketaklinearan dan menyediakan algoritma yang meramalkan pilihan terbaik secara statistik. Kaedah ini amat berkesan dalam menangani data berbilang dimensi menjadikannya sangat mudah difahami dan sesuai untuk tugasan klasifikasi. Walau bagaimanapun, algoritma DT cenderung untuk terlebih penyesuaian (*overfitting*) terutamanya jika pohon terlalu kompleks. Ia juga tidak stabil di mana sedikit perubahan kepada data boleh menyebabkan perubahan besar pada struktur pohon dan tidak efisyen untuk data yang bersaiz besar.

2.3.3 Perhutanan Rawak

Perhutanan Rawak (RF) memperluaskan konsep DT dengan membina beberapa DT dan menggabungkan keputusan mereka untuk menghasilkan ramalan yang lebih baik dan stabil. RF menggabungkan dua konsep utama iaitu pembelajaran gabungan (*ensemble learning*) dan pengambilan sampel berulang (*bootstrapping*).

Pembelajaran gabungan adalah teknik yang menggabungkan keputusan beberapa model untuk meningkatkan prestasi keseluruhan. Dalam RF, model yang digabungkan ialah DT. Setiap pohon dalam RF dibina menggunakan subset rawak dari data latihan dengan penggantian, yang bermaksud bahawa beberapa sampel boleh muncul lebih daripada sekali dalam satu subset, sementara yang lain mungkin tidak muncul sama sekali. Di setiap nod dalam setiap pohon keputusan, subset kecil ciri (atribut) dipilih secara rawak untuk menentukan pemisahan terbaik, mengurangkan

korelasi antara pohon dan meningkatkan kepelbagaian mereka (Raposo et al. 2020). Model RF ditunjukkan dalam Rajah 2.4.



Rajah 2.4 Model Perhutanan Rawak

Dua formula penting yang digunakan dalam proses ini adalah Maklumat Keuntungan (*Information Gain*) dan Indeks Gini. Kedua-dua formula ini digunakan untuk menentukan cara terbaik untuk membahagikan data di setiap nod pohon keputusan. Maklumat Keuntungan digunakan untuk memilih atribut yang memberikan pemisahan terbaik pada setiap nod dalam pohon keputusan. Ia mengukur pengurangan dalam ketidaktentuan (entropi) setelah membahagikan data berdasarkan atribut tertentu seperti yang ditunjukkan dalam Persamaan 2.2.

$$Maklumat\ Keuntungan\ (D, A) = H(D) - \sum_{v \in NilaiA} \frac{|D_v|}{|D|} H(D_v) \quad \ldots(2.2)$$

Di mana:

- D adalah set data semasa
- A adalah atribut yang sedang dipertimbangkan untuk pemisahan
- H(D) adalah entropi set data D
- D_v adalah subset data D di mana atribut A mempunyai nilai *v*

- Nilai (A) adalah set semua nilai yang mungkin diambil oleh atribut A
- $\frac{|D_v|}{|D|}$ adalah sebahagian subset D_v terhadap set data D

Indeks Gini digunakan untuk mengukur ketidaksamaan atau ketidaktentuan dalam pembahagian data pada setiap nod. Ia mengukur kebarangkalian bahawa elemen yang dipilih secara rawak akan salah klasifikasi. Indeks Gini yang rendah menandakan pembahagian data adalah lebih baik. Persamaan Indeks Gini adalah seperti Persamaan 2.3.

$$Gini\ (D) = 1 - \sum_{i=1}^{n} p_i^2 \quad \ldots(2.3)$$

Di mana:

- D adalah set data semasa
- n adalah jumlah kelas dalam set data D
- p_i adalah kebarangkalian elemen daripada kelas i dalam set data D

Dengan menggabungkan keputusan dari pelbagai pohon, RF dapat mengatasi masalah terlebih penyesuian berbanding DT. RF juga dapat mampu memberikan ketepatan ramalan yang baik dalam keadaan set data yang tidak lengkap ataupun data hilang. Algoritma ini efisien dan dapat mengendalikan dengan set data yang besar serta mempunyai banyak ciri. Selain itu, RF juga menghasilkan model yang lebih pelbagai dan tahan terhadap variasi data kerana menggunakan subset ciri secara rawak (Little et al. 2022). RF adalah algoritma yang fleksibel dan boleh digunakan dengan pelbagai jenis data termasuk data berstruktur, data kategori, data numerik, data berkala, data teks dan data gambar.

Walau bagaimanapun, membina dan menggabungkan banyak DT memerlukan pengiraan yang lebih intensif berbanding model yang lebih sederhana. Berbanding dengan pohon DT, hasil dari perhutanan rawak lebih sukar untuk ditafsirkan kerana ia

adalah gabungan dari banyak pohon. Penyimpanan semua pohon keputusan dalam perhutanan rawak juga boleh memerlukan jumlah memori yang besar.

2.3.4 Naive Bayes

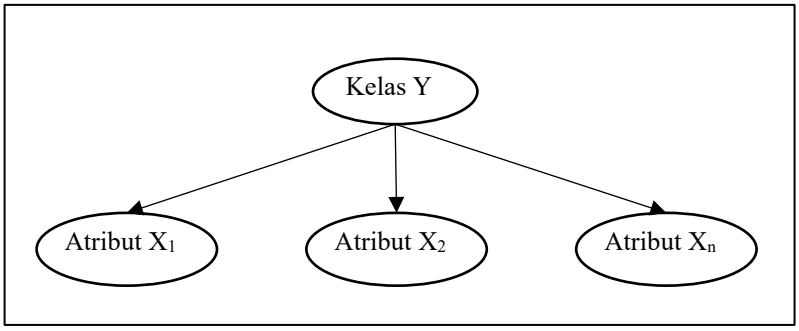
Algoritma Naive Bayes (NB) adalah kaedah klasifikasi kebarangkalian yang berasaskan teorem Bayes. Teorem Bayes mencari kebarangkalian satu peristiwa berlaku dengan mengambil kira kebarangkalian peristiwa lain yang telah berlaku. Kebarangkalian ini ditunjukkan dalam Persamaan 2.4.

$$P(A|B) = \frac{P(A|B).P(A)}{P(B)} \quad \ldots(2.4)$$

Di mana:

- P(A|B) adalah kebarangkalian kejadian A diberikan bahawa B benar
- P(B|A) adalah kebarangkalian kejadian B diberikan bahawa A benar
- P(A) adalah kebarangkalian awal kejadian A
- P(B) adalah kebarangkalian awal kejadian B

NB mengandaikan kewujudan setiap ciri atribut adalah tidak bersandar antara satu sama lain dan setiap atribut menyumbang secara bebas kepada kebarangkalian kelas. Pengiraan kebarangkalian untuk setiap hipotesis adalah berasingan, menjadikan algoritma ini cekap dari segi pengiraan dan sesuai untuk masalah berdimensi tinggi (K & H K 2020). Rajah 2.5 menunjukkan Model NB menggunakan Teorem Bayes di mana Atribut X menyumbang secara bebas dan berasingan kepada Kelas Y.



Rajah 2.5 Model Naive Bayes menggunakan Teorem Bayes

Sumber: Mivule 2014

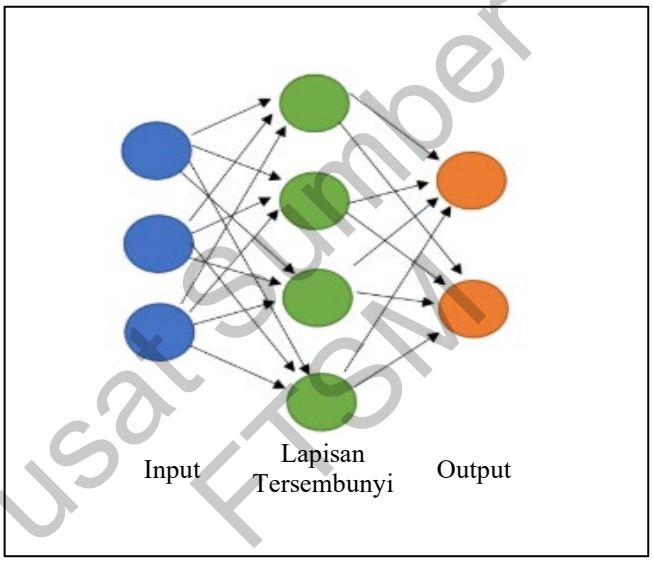
Kelebihan utama NB ialah efisien dan pantas membuat ramalan, menjadikannya salah satu algoritma klasifikasi yang paling cepat dan sangat sesuai untuk dataset yang besar. NB mudah difahami dan digunakan, yang menjadikannya pilihan yang baik dalam kebanyakan kajian awal pembelajaran mesin. Selain itu, algoritma ini tidak memerlukan banyak penalaan parameter. NB adalah algoritma yang fleksibel dan boleh digunakan dengan pelbagai jenis data, tetapi ia lebih sesuai dengan jenis data kategori. NB juga berfungsi dengan baik dengan data yang tidak seimbang, kerana ia mengira kebarangkalian untuk setiap kelas secara terpisah.

Kelemahan model ini ialah anggapan bahawa semua ciri adalah bebas antara satu sama lain adalah secara teori sahaja dan jarang berlaku dalam dunia nyata, hal ini boleh mempengaruhi ketepatan model dalam sesetengah kes. Selain itu, NB tidak mempertimbangkan interaksi atau hubungan antara ciri, yang boleh menyebabkan kehilangan maklumat penting. Walaupun NB boleh digunakan dengan jenis data berterusan, prestasinya mungkin tidak sebaik algoritma lain yang lebih khusus untuk jenis data berterusan. Satu lagi kelemahan adalah masalah dengan frekuensi sifar, di mana kategori dalam ciri yang tidak muncul dalam data latihan akan memberikan kebarangkalian sifar, yang boleh menyebabkan masalah dalam membuat ramalan.

2.3.5 Artificial Neural Network

Rangkaian Neural Buatan (ANN) ialah model simulasi komputer yang terinspirasi oleh cara sistem saraf otak manusia memproses maklumat. ANN merupakan pemodelan matematik dengan matlamat mencipta sistem yang boleh menghasilkan dan menemukan maklumat baru secara automatik tanpa pengaturcaraan berasingan. Ia

terdiri daripada sejumlah besar elemen pemprosesan yang berhubungan antara satu sama lain yang dipanggil neuron, yang bekerja bersama-sama untuk menyelesaikan masalah tertentu seperti pengenalan corak, pengelasan data dan ramalan (Thorat et al. 2022). ANN meniru keupayaan otak untuk belajar dari contoh melalui proses yang melibatkan penyesuaian hubungan sinaptik di antara neuron sebagaimana sistem biologi. ANN berkesan dalam mengendalikan data yang tidak linear dan kompleks. Struktur ANN biasanya mengandung input, pemberat, fungsi penambahan, fungsi pengaktifan dan output, dengan pembelajaran dicapai melalui penyesuaian berulang parameter rangkaian berdasarkan kesilapan antara output ramalan dan output sebenar.



Rajah 2.6 Model Ringkas ANN

Sumber: Thorat et al. 2022

Rajah 2.6 menunjukan model ringkas ANN. Di dalam model ini, lapisan input menerima data mentah yang akan diproses oleh rangkaian neural. Setiap unit dalam lapisan input mewakili satu ciri atau atribut dalam set data. Lapisan tersembunyi terdiri daripada beberapa neuron yang mengubah dan memproses input melalui bobot dan bias yang dikaitkan dengan setiap sambungan neuron. Lapisan ini boleh terdiri daripada satu atau lebih lapisan tersembunyi. Lapisan output menghasilkan ramalan akhir model. Setiap neuron dalam lapisan output mewakili kelas atau nilai yang mungkin.

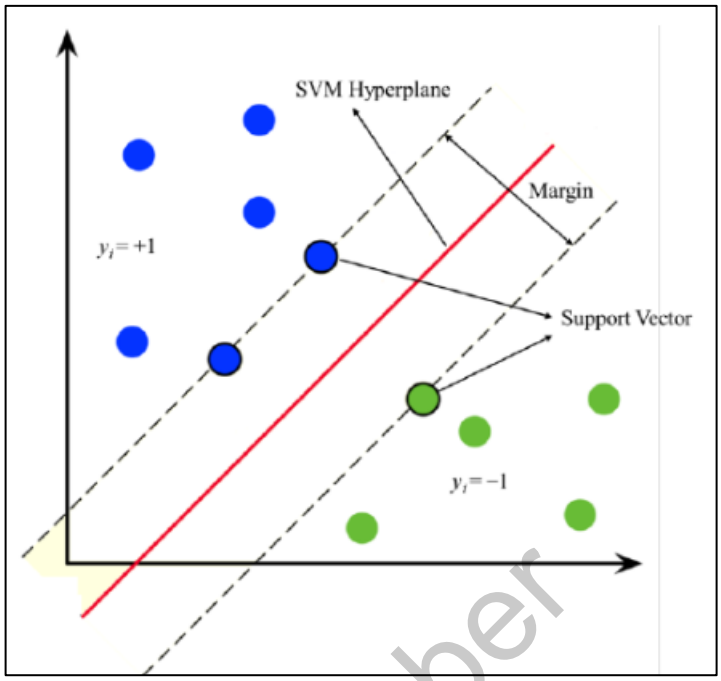
Kelebihan utama ANN termasuk kemampuannya untuk belajar dan mengenali corak yang kompleks, menjadikannya sesuai untuk pelbagai tugas pembelajaran mesin

seperti klasifikasi, regresi dan pemprosesan gambar. ANN juga fleksibel dan boleh digunakan untuk pelbagai jenis data dan aplikasi, termasuk pengenalan suara, pengenalan wajah, dan pemprosesan bahasa semula jadi (NLP). ANN mampu terus belajar dan memperbaiki prestasinya seiring dengan penambahan data baharu, yang membolehkan pembelajaran berterusan. Namun, ANN juga mempunyai beberapa kelemahan kerana ia memerlukan data yang besar untuk latihan yang berkesan, dan proses latihannya memerlukan sumber daya pengkomputeran yang tinggi. Selain itu, model ANN sering dianggap sebagai "kotak hitam" kerana kesukaran dalam mentafsirkan bagaimana mereka membuat keputusan, menjadikannya kurang sesuai dalam situasi di mana interpretasi model adalah kritikal.

2.3.6 Mesin Sokongan Vektor

Mesin Sokongan Vektor (SVM) ialah algoritma pembelajaran mesin yang diselia, digunakan untuk model klasifikasi dan regresi. SVM berkembang daripada teori pembelajaran statistik dengan tujuan untuk menyelesaikan masalah tanpa menyebabkan masalah yang lebih besar sebagai langkah pertengahan (Othman et al. 2018).

SVM boleh mengendalikan data berdimensi tinggi dan sangat berkesan dalam keadaan di mana bilangan dimensi melebihi bilangan sampel. Prestasi algoritma ini sangat bergantung pada pemilihan parameter, terutamanya fungsi kernel yang mengubah data input ke ruang berdimensi lebih tinggi untuk menjadikannya boleh dipisahkan secara linear. Keupayaan SVM untuk memaksimumkan margin antara kelas dan kesesuaian dengan pelbagai jenis data menjadikannya algoritma yang berkuasa dalam pembelajaran mesin (Yang & Prayogo 2022).



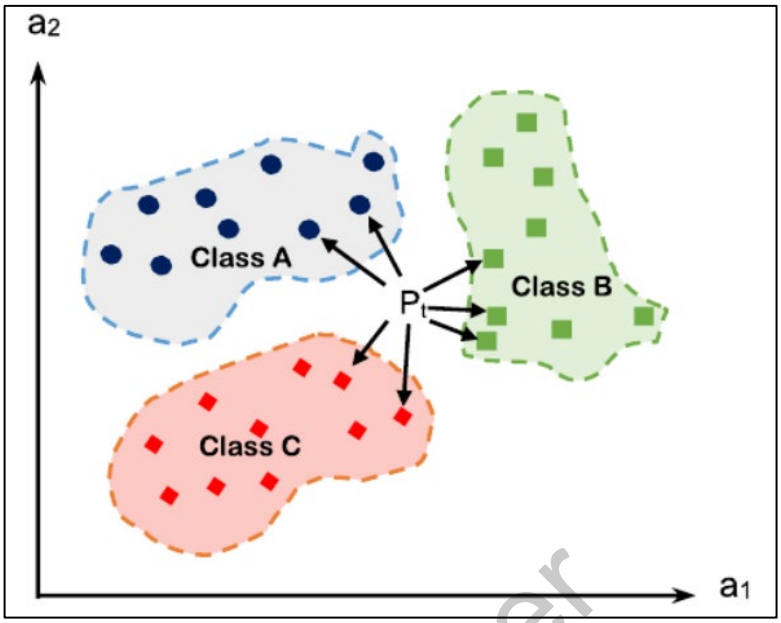
Rajah 2.7 Model SVM

Sumber: Yang & Prayogo 2022

Rajah 2.7 merupakan Model SVM di mana SVM beroperasi dengan mencari *hyperplane* optimal yang memisahkan titik data daripada kelas berbeza dengan margin maksimum. *Hyperplane* ini adalah subruang berdimensi (N-1) dalam ruang ciri berdimensi N dan titik data yang paling hampir dengan *hyperplane* ini dipanggil vektor sokongan (Feizi & Nazemi 2022).

2.3.7 KNN

K-Nearest Neighbors (KNN) adalah teknik statistik yang terkenal untuk pengenalan corak dan ramalan. KNN menyimpan semua data input dan kemudian mengelaskan serta mencipta titik data baru. Titik data baru ini dicipta berdasarkan kesamaan yang wujud dalam input dan hubungannya dengan output. Algoritma ini menentukan kelas bagi satu titik data tertentu dengan mengenal pasti 'K' jiran terdekat dan menggunakan undian majoriti atau merata-ratakan nilai mereka (Raman & Pramod 2022). Operasi Model KNN ditunjukkan pada Rajah 2.8.



Rajah 2.8 Model KNN

Sumber: Atallah et al. 2019

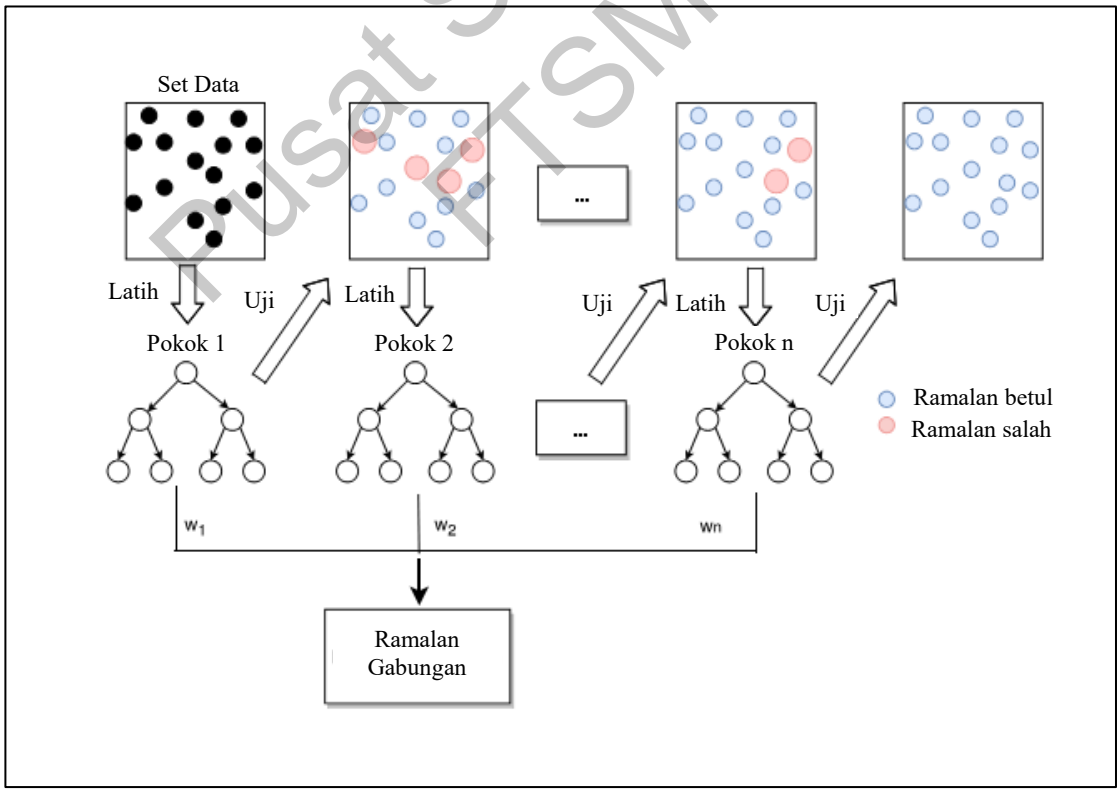
Salah satu kelebihan utama KNN adalah kesederhanaannya, kerana ia tidak memerlukan fasa latihan yang kompleks. Semua pengiraan dilakukan semasa fasa ujian dengan mencari 'K' jiran terdekat bagi sampel baru. KNN juga fleksibel dan boleh digunakan untuk tugas klasifikasi dan regresi, serta mampu menangani jenis data kategori dan nombor.

Kelemahan utama KNN ialah keperluan pengiraan yang tinggi terutamanya untuk set data yang besar, kerana ia perlu mengira jarak antara sampel ujian dan semua sampel latihan. Selain itu, KNN memerlukan jumlah memori yang signifikan kerana semua data latihan perlu disimpan. Algoritma ini juga sensitif kepada ciri-ciri yang tidak relevan dan skala ciri, jadi proses transformasi ciri adalah penting sebelum menggunakan KNN. Prestasi KNN sangat bergantung kepada pilihan parameter K, parameter K yang tidak sesuai boleh menyebabkan model mengalami terlebih atau terkurang penyesuaian.

2.3.8 Mesin Galakan Kecerunan

Mesin Galakan Kecerunan (GBM) membina model ramalan dengan menggabungkan kekuatan beberapa model yang lebih lemah, seperti DT untuk membentuk model yang lebih kuat. Pendekatan ini dikenali sebagai penggalakan (*boosting*) dan ia berfungsi dengan menambah model-model yang baru secara berturut-turut untuk memperbaiki kesilapan yang dibuat oleh model-model sebelumnya.

Penggalakan adalah teknik gabungan yang menggabungkan beberapa model lemah untuk membentuk model yang lebih kuat. Dalam konteks GBM, setiap model baru ditambahkan untuk memperbaiki kesalahan yang dibuat oleh model-model sebelumnya. GBM menggunakan teknik penurunan kecerunan (*gradient descent*) untuk mengoptimumkan fungsi kerugian. Model baru ditambahkan sedemikian rupa sehingga mengurangkan kerugian secara berturut-turut. Rajah 2.9 menunjukkan model pembelajaran GBM.



Rajah 2.9 Model Pembelajaran GBM

Sumber: Zhang et al. 2021

Kelebihan utama GBM ialah berprestasi tinggi dalam meramalkan data kompleks, keupayaannya untuk menangani data tidak seimbang dan fleksibilitinya untuk digunakan dengan pelbagai jenis fungsi kerugian. Dengan menambah model lemah secara berturut-turut dan memperbaiki kesilapan model sebelumnya menggunakan teknik gradient descent, GBM dapat menghasilkan model yang sangat tepat. Selain itu, GBM juga menawarkan mekanisme regularisasi seperti kadar pembelajaran dan pengebangan pohon yang tidak terlalu dalam, yang membantu dalam mengurangkan risiko terlebih penyesuaian.

GBM juga mempunyai beberapa kelemahan. Salah satu kelemahannya adalah keperluan pengiraan yang tinggi, yang boleh menyebabkan masa latihan yang panjang dan penggunaan sumber pengkomputeran yang besar, terutama untuk set data yang besar. Selain itu, model yang dihasilkan oleh GBM lebih sukar untuk ditafsirkan berbanding model linear atau DT tunggal, menjadikannya kurang sesuai untuk aplikasi di mana interpretasi model adalah kritikal. GBM juga sensitif terhadap penalaan parameter seperti jumlah pohon, kedalaman pohon, dan kadar pembelajaran, yang memerlukan penalaan yang teliti untuk mencapai prestasi terbaik. GBM sesuai untuk meramalkan pelbagai jenis data, termasuk data berstruktur dan tidak berstruktur.

2.3.9 Perbandingan Model Klasifikasi

Jadual 2.2 menunjukkan perbandingan model klasifikasi yang dibangunkan menggunakan algoritma Regresi Logistik (LR), Pohon Keputusan (DT), Perhutanan Rawak (RF), Naive Bayes (NB), Rangkaian Neural Tiruan (ANN), *K-Nearest Neighbors* (KNN), Mesin Sokongan Vektor (SVM) dan *Gradient Boosting Machine* (GBM):

Jadual 2.2 Perbandingan Model Klasifikasi

Model	Kelebihan	Kelemahan	Jenis Data
Regresi Logistik (LR)	Mudah ditafsirkan, ringkas dan mudah diproses. Berfungsi baik dengan data linear.	Sensitif terhadap nilai terpencil. Tidak sesuai untuk hubungan tidak linear. Kurang berkesan dengan data yang banyak ciri atau kompleks.	Data kategori, numerik, berstruktur.

bersambung...

…sambungan

Pohon Keputusan (DT)	Mudah ditafsirkan, memahami data tidak linear, menangani data berbilang dimensi.	Cenderung untuk terlebih penyesuaian. Tidak efisyen untuk data bersaiz besar.	Data kategori, numerik, berstruktur.
Perhutanan Rawak (RF)	Menangani masalah terlebih penyesuaian. Berkesan dengan data besar dan mempunyai banyak ciri	Keperluan pengiraan yang tinggi. Memerlukan memori yang besar. Sukar untuk ditafsirkan.	Data kategori, numerik, berstruktur.
Naive Bayes (NB)	Cepat, efisien dan mudah difahami. Berfungsi baik dengan data tidak seimbang.	Tiada interaksi dan hubungan antara ciri.	Data kategori, teks, binari.
Rangkaian Neural Tiruan (ANN)	Mampu menangani hubungan non-linear dan corak yang kompleks. Fleksibel untuk pelbagai aplikasi.	Memerlukan data dan pengiraan yang banyak. Sukar ditafsirkan.	Data berstruktur, tidak berstruktur, gambar, teks.
K-Nearest Neighbors (KNN)	Tidak memerlukan fasa latihan.	Keperluan memori dan pengiraan yang tinggi. Kurang berkesan dengan data berdimensi tinggi.	Data kategori, numerik, berstruktur.
Mesin Sokongan Vektor (SVM)	Berkesan untuk data berdimensi tinggi. Prestasi tinggi dengan margin pemisahan yang baik.	Keperluan pengiraan yang tinggi. Sukar ditafsirkan.	Data kategori, numerik, berstruktur.
Gradient Boosting Machine (GBM)	Prestasi tinggi, menangani data tidak seimbang dengan baik. Fleksibel dengan pelbagai fungsi kerugian.	Keperluan pengiraan yang tinggi. Sensitif terhadap penalaan hiperparameter.	Data berstruktur, kategori, masa.

2.4 PENILAIAN MODEL

Penilaian model adalah proses penting dalam pembangunan model ramalan. Ia bertujuan untuk menilai prestasi model dalam menghasilkan ramalan yang tepat dan berguna (Naidu et al. 2023). Selain itu, penilaian model dilakukan untuk membandingkan prestasi antara model dengan algoritma yang berbeza. Perkara ini penting untuk mengenal pasti model terbaik dan sesuai dengan objektif kajian. Beberapa komponen utama dalam penilaian model ialah pembahagian data, matriks kekeliruan dan matriks penilaian.

2.4.1 Pembahagian Data

Kaedah ini membahagikan set data kepada dua bahagian iaitu data latihan dan data ujian. Tujuan pembahagian data adalah untuk menilai prestasi dan kemampuan generalisasi model ramalan. Dengan membahagikan set data ini, model dilatih dan diuji

pada subset data yang berasingan bagi memastikan matriks penilaian benar-benar menilai kemampuan model untuk mengendalikan data yang tidak pernah dilihat. Kaedah ini akan membantu mengatasi masalah terlebih penyesuaian terhadap model (Jain et al. 2022). Teknik yang biasa digunakan dalam pembahagian data ialah Pemisahan Data dan Pengesahan Bersilang

Pemisahan data ialah memisahkan data sebelum model dibangunkan. Nisbah pemisahan yang sering digunakan ialah 80:20 iaitu 80% data latihan dan 20% data ujian. Lain-lain nisbah yang biasa digunakan ialah 70:30, 60:40 dan 50:50 (Joseph 2022). Kelebihan teknik ini ialah prestasi model dinilai pada data yang tidak pernah dilihat semasa latihan.

Pengesahan Bersilang-K (*K-Fold Cross Validation*) adalah teknik yang lebih canggih di mana set data dibahagikan secara rawak kepada *k* subset yang mempunyai saiz yang sama dan ujian akan diulang sebanyak *k* ulangan. Di setiap ulangan, satu subset data akan menjadi data ujian manakala selebihnya akan menjadi data latihan. Proses ini akan diulang sehingga setiap subset menjadi data ujian (Widodo et al. 2022). Teknik ini memberikan anggaran prestasi model yang lebih stabil dan tidak berat sebelah. Ia juga memastikan setiap data digunakan sebagai data latihan dan ujian pada satu masa, memaksimumkan penggunaan set data.

2.4.2 Matriks Kekeliruan

Matriks Kekeliruan merupakan salah satu kaedah yang boleh digunakan untuk mengukur prestasi model ramalan. Matriks Kekeliruan mempunyai maklumat yang membandingkan ramalan klasifikasi yang dilakukan oleh model (ramalan) dengan hasil klasifikasi yang sepatutnya (sebenar). Terdapat empat istilah yang digunakan untuk menunjukkan hasil klasifikasi iaitu Positif Sebenar (TP), Negatif Sebenar (TN), Positif Palsu (FP) dan Negatif Palsu (FN) (Putrisanni et al. 2019).

Sebenar	Ramalan	
	Ramalan Positif	**Ramalan Negatif**
Positif Sebenar	TP	FN
Negatif Sebenar	FP	TN

Rajah 2.10 Matriks Kekeliruan

Sumber: He et al. 2022

Rajah 2.10 menunjukkan matriks kekeliruan untuk pengelasan binari. Positif Sebenar (TP) ialah bilangan kes di mana model meramalkan positif dan hasil sebenar juga positif. Positif Palsu (FP) ialah bilangan kes di mana model meramalkan positif tetapi hasil sebenar adalah negatif. Manakala Negatif Palsu (FN) adalah bilangan kes di mana model meramalkan negatif tetapi hasil sebenar adalah positif dan Negatif Sebenar (TN) adalah bilangan kes di mana model meramalkan negatif dan hasil sebenar juga negatif. Daripada matriks kekeliruan ini, beberapa matriks penilaian lain boleh dikira iaitu ketepatan, kejituan, sensitiviti dan skor F1 (He et al. 2022).

2.4.3 Matriks Penilaian

Matriks penilaian adalah sekumpulan metrik yang digunakan untuk menilai prestasi model pembelajaran mesin. Metrik-metrik ini memberikan pandangan yang lebih komprehensif tentang bagaimana model berfungsi, termasuk sejauh mana model membuat ramalan yang betul, jenis kesilapan yang dibuat, dan sejauh mana model dapat generalisasi kepada data baru. Berikut adalah penjelasan beberapa metrik penilaian utama:

1. Ketepatan ialah nisbah ramalan betul kepada jumlah keseluruhan ramalan. Ketepatan mudah difahami dan digunakan untuk set data yang seimbang. Jika nilai ketepatan adalah ialah 0.8, bermakna 80% ramalan model adalah tepat. Formula ketepatan adalah seperti Persamaan 2.5 (ElSharkawy et al. 2022).

$$Ketepatan = \frac{TP + TN}{TP + FP + TN + FN} \quad \ldots(2.5)$$

2. Kejituan adalah nisbah ramalan betul kategori positif kepada jumlah ramalan kategori positif. Nilai kejituan mengukur kebolehpercayaan ramalan positif model iaitu menunjukkan prestasi model meramalkan sesuatu data itu positif.

Matriks ini tidak mengambil kira kesalahan negatif (FN). Jika nilai kejituan adalah ialah 0.8, bermakna 80% daripada ramalan positif adalah betul. Formula kejituan adalah seperti Persamaan 2.6 (ElSharkawy et al. 2022).

$$Kejituan = \frac{TP}{TP + FP} \quad \ldots(2.6)$$

3. Sensitiviti (*Recall*) ialah nisbah ramalan betul kategori positif kepada jumlah sebenar kategori positif. Sensitiviti mengukur keupayaan model untuk mengenal pasti semua kes positif yang sebenar. Matriks ini tidak mengambil kira kesalahan positif (FP). Jika nilai sensitiviti adalah 0.8, bermakna 80% daripada kes positif sebenar dikenalpasti dengan betul oleh model. Formula sensitiviti adalah seperti Persamaan 2.7 (ElSharkawy et al. 2022).

$$Sensitiviti = \frac{TP}{TP + FN} \quad \ldots(2.7)$$

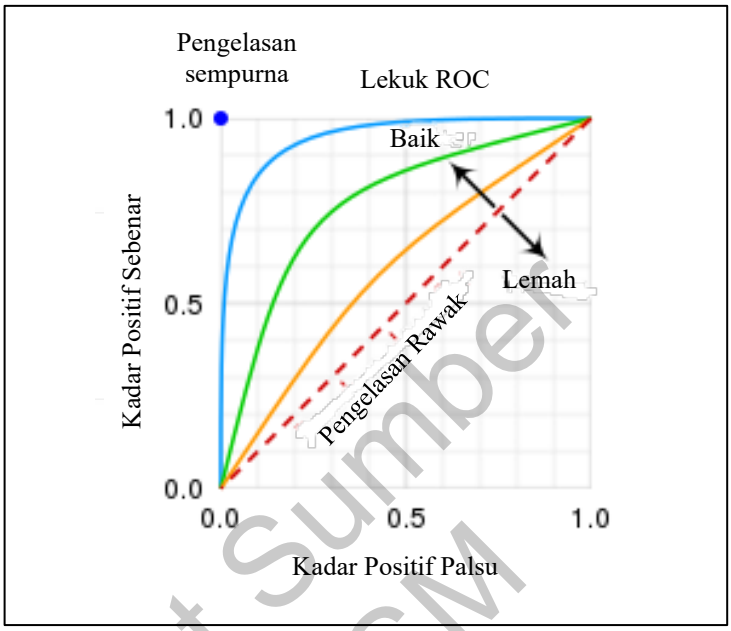
4. Skor F1 adalah purata harmonik bagi kejituan dan sensitiviti. Skor F1 memberikan keseimbangan antara kejituan dan sensitiviti. Digunakan untuk set data yang tidak seimbang. Jika nilai Skor F1 adalah 0.8, bermakna keseimbangan antara kebolehpercayaan ramalan positif (kejituan) dan keupayaan mengenal pasti kes positif (sensitiviti) ialah 80%. Formula skor F1 adalah seperti Persamaan 2.8 (ElSharkawy et al. 2022).

$$Skor\ F1 = \frac{2 * Kejituan * Sensitiviti}{Kejituan + Sensitiviti} \quad \ldots(2.8)$$

5. Kawasan di Bawah Keluk Penerima Operasi (AUC-ROC) adalah ukuran keupayaan model untuk membezakan antara kelas positif dan negatif. ROC adalah graf yang menunjukkan perbandingan antara Kadar Positif Sebenar (TPR) dan Kadar Positif Palsu (FPR) pada pelbagai ambang keputusan seperti yang ditunjukkan pada Rajah 2.11. Semakin tinggi nilai AUC, semakin baik model dalam membezakan antara kelas positif dan negatif. Matriks ini sesuai untuk set data yang tidak seimbang. Formula TPR dan FPR adalah seperti Persamaan 2.9 dan 2.10 (Othman et al. 2018).

$$TPR = \frac{TP}{TP + FN} \quad \ldots(2.9)$$

$$FPR = \frac{FP}{FP + TN} \quad \ldots(2.10)$$



Rajah 2.11 Keluk ROC

2.5 KAJIAN LEPAS BERKAITAN MODEL RAMALAN KEBOLEHPASARAN

Model ramalan kebolehpasaran pelajar telah menjadi fokus utama dalam penyelidikan akademik dan industri, bertujuan untuk menyediakan alat yang berkesan dalam meramal peluang pekerjaan pelajar. Pelbagai model statistik dan pembelajaran mesin telah dibangunkan untuk meramalkan kebolehpasaran, dengan mempertimbangkan pelbagai faktor demografi, peribadi dan luaran. Kajian literatur ini akan meneliti model-model ramalan yang telah dibangunkan dalam kajian-kajian terdahulu, menilai pendekatan dan metodologi yang digunakan, serta mengkaji keberkesanan dan ketepatan model-model tersebut.

2.5.1 Model Ramalan Kebolehpasaran di Malaysia

Kebanyakan kajian yang dijalankan di Malaysia menggunakan data SKPG sebagai sumber data sejarah untuk pembangunan model ramalan kebolehpasaran. Sapaat et al.

(2011) menggunakan data pengesanan graduan 2009 dari SKPG untuk membuat perbandingan di antara model ramalan dari kelompok Algoritma Bayes seperti Naive Bayes (NB), *Simple* NB dan *Bayes Net* serta kelompok Algoritma Pohon seperti Pohon Keputusan (DT) dan Perhutanan Rawak (RF). Hasil perbandingan ini mendapati varian Algoritma Pohon mendapat ketepatan paling tinggi iaitu 92.3% berbanding purata Algoritma Bayes 91.3%. Kajian ini juga menggunakan teknik pemilihan atribut Maklumat Keuntungan (*Information Gain*) untuk mengenal pasti dan memilih ciri utama untuk membangunkan model ramalan. Kajian ini mendapati sektor pekerjaan, status pekerjaan dan bidang merupakan faktor utama yang mempengaruhi kebolehpasaran graduan (Sapaat et al. 2011).

Rahman et al. (2017) mengkaji model klasifikasi untuk meramal dan menilai atribut pelajar dalam bidang akademik yang memenuhi keperluan industri. Kajian tersebut menggunakan data sejarah kebolehpasaran graduan Universiti Teknologi Malaysia (UTM), Universiti Sains Malaysia (USM), Universiti Kebangsaan Malaysia (UKM), Universiti Putra Malaysia (UPM) dan Universiti Malaya (UM) tahun 2014 – 2016 yang diperolehi daripada SKPG. 41 daripada 68 atribut asal dikenal pasti untuk membangunkan model klasifikasi seperti program pengajian, jantina dan CGPA. Model ini meramalkan lima kelas iaitu bekerja, tidak bekerja, melanjutkan pengajian, meningkatkan kemahiran dan menunggu penempatan. Enam model klasifikasi dibandingkan dalam kajian ini iaitu *K-Nearest Neighbors* (KNN), Regresi Logistik (LR), Mesin Sokongan Vektor (SVM), Rangkaian Neural Buatan (ANN), NB dan DT. Hasil kajian mendapati bahawa model klasifikasi KNN mencapai skor ketepatan tertinggi iaitu 97.78% berbanding model klasifikasi lain. Kajian ini mencadangkan supaya model ini digunakan untuk membantu pengurusan universiti merangka pelan jangka panjang bagi menghasilkan pelajar yang mempunyai kemahiran dan pengetahuan yang diperlukan oleh industri. Kajian juga mencadangkan supaya, atribut tambahan seperti gred bagi subjek umum yang diambil semasa tempoh pengajian dimasukkan sebagai atribut penilaian untuk menentukan kesannya terhadap kebolehpasaran pelajar (Rahman et al. 2017).

Othman et al. (2018) turut membangunkan model ramalan kebolehpasaran pelajar menggunakan data SKPG tetapi menumpukan kepada graduan UKM sahaja dari

tahun 2011 – 2017 berjumlah 43 868. Fokus kajian ini adalah untuk mengenalpasti faktor yang mempengaruhi kebolehpasaran pelajar dan membandingkan 3 model klasifikasi iaitu DT, ANN dan SVM. Daripada 357 atribut yang diperolehi daripada data SKPG, hanya 9 atribut yang digunakan untuk pembangunan model ramalan selepas melalui proses pemilihan ciri iaitu umur, fakulti, bidang pengajian, pendapatan keluarga, ko-kurikulum, status perkahwinan, latihan industri, kemahiran Bahasa Inggeris dan status kebolehpasaran sebagai kelas sasaran iaitu bekerja atau tidak bekerja. Mekanisme pengujian yang digunakan ialah pengesahan silang 10 lipatan. Hasil kajian ini mendapati model DT menggunakan algoritma J48 memberikan hasil ramalan yang lebih baik berbanding ANN dan SVM dengan ketepatan ramalan sebanyak 66%. Kajian ini juga mendapati atribut umur, latihan industri dan fakulti merupakan faktor penentu kebolehpasaran graduan UKM (Othman et al. 2018).

Seterusnya ialah kajian oleh Pauzi et al. (2021) yang menggunakan data kebolehpasaran SKPG tahun 2015 – 2018 sebanyak 375,507. Disebabkan jumlah data yang besar dan pelbagai, teknik perlombongan data terselia dan tidak terselia telah digunakan untuk mengenal pasti pembolehubah dan maklumat tersembunyi bagi memastikan ramalan lebih tepat. Kajian ini bertujuan untuk meramalkan status pekerjaan pelajar berdasarkan faktor kebolehpasaran seperti umur, kursus, CGPA, tahun konvokesyen, kecekapan bahasa Inggeris, kecekapan bahasa ketiga, pengalaman dengan kaunseling, fasilitator, kemudahan, pembiayaan, jantina, kemahiran ICT, akses perpustakaan, status OKU, bangsa, mod pengajian, sistem dan pengalaman kerja. Ia juga memprofilkan kepuasan pelajar terhadap aktiviti kokurikulum dan kemahiran ICT. Model DT, LR dan ANN digunakan untuk pembangunan model ramalan. Manakala teknik pengelompokan *K-Means* digunakan untuk mendedahkan corak tersembunyi dan mengkategorikan graduan kepada tujuh kluster berdasarkan tahap kepuasan mereka terhadap aktiviti kokurikulum dan kemahiran ICT. Model LR dengan pemilihan pemboleh ubah menggunakan LR *stepwise* dikenal pasti sebagai model terbaik untuk meramalkan status pekerjaan pelajar. LR *stepwise* mencapai ketepatan 61.02% dan faktor penentu kebolehpasaran yang dikenal pasti ialahh kursus, umur, pengalaman kerja, bangsa dan tahun tamat pengajian (Pauzi et al. 2021).

Kajian terkini oleh Haque et al. (2024) telah memberi tumpuan untuk mengenal pasti set atribut yang mempengaruhi ketepatan model ramalan. Kajian ini menggunakan gabungan data graduan Universiti Multimedia (MMU) dan data kajian pengesanan SKPG tahun 2021 mengandungi maklumat demografi, maklumat akademik, maklumat SKPG, maklumat GPA dan maklumat gred subjek. Maklumat-maklumat ini kemudiannya dibentuk menjadi 8 set data yang mengandungi gabungan maklumat yang berbeza. Model ramalan dibangunkan daripada 8 set data ini menggunakan model Penggalak Kecerunan Ekstrem (XBG), LR, NB, RF, SVM dan ANN. Setiap model dilatih sebanyak 8 kali menggunakan set data eksperimen sebagai perbandingan terhadap atribut yang digunakan. Hasil kajian mendapati dataset D6 yang merangkumi maklumat demografi, maklumat akademik, GPA dan maklumat dari SKPG mengandungi 52 atribut menunjukkan prestasi terbaik dengan model klasifikasi ANN memberikan ketepatan ramalan tertinggi iaitu 80% diikuti oleh SVM, RF dan XGB. Kajian ini turut mendedahkan bahawa kepuasan pelajar terhadap kemudahan universiti seperti perpustakaan dan kuliah adalah penting untuk kebolehpasaran mereka pada masa depan. (Haque et al. 2024)

2.5.2 Model Ramalan Kebolehpasaran di Luar Negara

Selain daripada itu, terdapat beberapa kajian terkini oleh penyelidik-penyelidik di luar negara berkaitan ramalan kebolehpasaran graduan. Casuat et al. (2020) menggunakan pendekatan pembelajaran mesin bagi meramalkan kebolehpasaran pelajar di Filipina. Set data kajian ini dikumpul daripada agensi berbeza yang mengandungi penilaian temuduga latihan iaitu keterampilan, cara bercakap, keadaan fizikal, kesedaran mental, keyakinan, kemampuan menyampaikan pendapat, kemahiran komunikasi, tahap prestasi, GPA, nama, program dan nombor pelajar. Kajian tersebut menggunakan enam model pembelajaran mesin iaitu RF, DT, LR, NB, SVM, ANN dan XGB untuk mendapatkan maklumat tentang kebolehpasaran pelajar. Dalam kajian tersebut, SVM memberikan ketepatan sebanyak 91% lebih tinggi berbanding model lain dengan DT sebanyak 85% dan RF sebanyak 84% (Casuat et al. 2020).

Seterusnya ialah kajian terhadap graduan Kejuruteraan Sains Maklumat dan Kejuruteraan Sains Komputer di India untuk meramalkan kebolehpasaran graduan menggunakan algoritma pembelajaran mesin bertujuan untuk merapatkan jurang antara

keputusan akademik dengan kebolehpasaran. Kajian ini membantu dalam intervensi khusus kepada pelajar yang memperolehi keputusan akademik yang rendah dan sukar untuk mendapatkan pekerjaan. Pengkaji menggunakan pelbagai model klasifikasi pembelajaran mesin untuk membangunkan model ramalan kebolehpasaran seperti ANN, LR, DT, KNN, SVM dan NB. Kajian ini mendapati ANN mencatatkan ketepatan yang paling tinggi iaitu sebanyak 87.42%.(K & H K 2020)

Kajian yang dijalankan oleh Bai dan Hira (2021) pula memperkenalkan model hibrid menggunakan ANN dan regresi *Softmax* untuk meramalkan kebolehpasaran pelajar. Untuk meningkatkan ketepatan model ramalan, teknik pemilihan ciri berasaskan algoritma carian burung gagak digunakan. Model pemilihan ciri ini membantu mengenal pasti subset ciri yang optimum daripada set asal, yang secara signifikan menyumbang kepada ramalan kebolehpasaran pelajar. Ciri-ciri optimum yang dipilih digunakan sebagai input untuk model ANN, yang membolehkan pembelajaran ciri intrinsik untuk menangkap representasi tahap tinggi. Seterusnya, regresi *Softmax* digunakan untuk meramalkan sama ada pelajar akan mendapat pekerjaan atau tidak. Model hibrid yang dicadangkan ini memberikan ketepatan ramalan sebanyak 98%. Analisis simulasi statistik untuk model hibrid ini dijalankan di MATLAB menggunakan dataset yang dikumpul melalui soal selidik, yang merangkumi atribut akademik asas CGPA dan atribut intelektual pelajar seperti kemahiran pengaturcaraan (Bai & Hira 2021).

Kajian oleh He et al. (2022) mencadangkan kaedah baru untuk meramal kebolehpasaran graduan menggunakan teknik *LightGBM,* satu rangka kerja Peningkatan Kecerunan Berpusat pada algoritma DT untuk mengukur faktor-faktor kebolehpasaran. Kaedah yang dicadangkan melibatkan analisis maklumat ciri klasifikasi yang pelbagai dalam data peribadi pelajar, mengikat ciri yang saling mengecualikan untuk membolehkan maklumat interaksi modal dan membina sistem ramalan kebolehpasaran berdasarkan *LightGBM*. Bagi tujuan pembangunan model, sebanyak 8 atribut telah dipilih dari 92 atribut asal berdasarkan kepentingan ciri yang dinilai menggunakan teknik RF. Atribut yang terlibat ialah jantina, major pengajian, akademik, pandangan politik, pendidikan, kesukaran pekerjaan, pelajar bandar atau luar bandar dan bangsa. Hasil kajian ini mendapati penggunaan teknik LightGBM

memberikan ketepatan ramalan yang baik sebanyak 83% berbanding model RF dan SVM yang memberikan ketepatan ramalan sekitar 60%. Kajian ini juga mendapati bahawa atribut jantina, major atau kepakaran bidang dan kolej memberikan pengaruh besar terhadap kebolehpasaran graduan (He et al. 2022).

Selain itu, terdapat juga kajian yang membangunkan model ramalan kebolehpasaran menggunakan data graduan dalam bidang statistik dan berkaitan dengannya dari universiti-universiti di Thailand. Kajian ini menggunakan 3,696 sampel dan 6 atribut dipilih untuk pembangunan model. Untuk meningkatkan prestasi model klasifikasi dan mengelakkan masalah penyesuaian berlebihan, pendekatan pemisahan data digunakan untuk memperbaiki model. Data latihan dan data ujian dipisahkan dalam nisbah 70:30. Model ramalan dibangunkan menggunakan 4 algoritma iaitu LR, DT, RF dan KNN. Kajian ini menyimpulkan bahawa model yang dibangunkan menggunakan kaedah DT dan KNN boleh digunakan untuk meramalkan kebolehpasaran graduan dengan ketepatan dan kejituan sehingga 70%. Kajian ini juga mendapati faktor penentu kebolehpasaran graduan dipengaruhi oleh kedudukan universiti, peringkat pengajian, major pengajian, jenis institusi, kategori institusi dan wilayah tempat tinggal. (Panityakul et al. 2022)

Dalam kajian yang dijalankan kepada graduan Kejuruteraan Elektronik di Filipina, pengkaji telah meneroka corak yang tidak diketahui untuk meramalkan kebolehpasaran graduan daripada data kognitif dan bukan kognitif. Berbeza dengan penyelidikan terdahulu, kajian ini menggunakan pendekatan model klasifikasi gabungan untuk meningkatkan prestasi keseluruhan model. Tujuh model klasifikasi digunakan dalam kajian ini iaitu *Multilayer Perceptron* (MLP), RF, SVM, DT, KNN, NB dan LR. Model RF memberikan ketepatan paling tinggi sebanyak 91.36% kerana keupayaannya untuk menangani masalah berbilang kelas diikuti dengan SVM (90.83%) dan NB (89.37%). Ketiga-tiga model ini digabungkan untuk membentuk model gabungan dan memberikan ketepatan semak silang sehingga 93.33%. Kajian ini mengenal pasti kemahiran teknikal, sijil profesional dan ketahanan diri sebagai peramal penting bagi kebolehpasaran graduan. Kajian ini memberikan perspektif unik mengenai kepentingan membangunkan ketahanan diri bersama kemahiran kognitif dalam tempoh

pengajian untuk mempersiapkan mereka menghadapi kerjaya profesional dalam pasaran kerja yang kompetitif (Maaliw et al. 2022).

ElSharkawy et al. (2022) telah membangunkan model ramalan kebolehpasaran graduan Teknologi Maklumat di Mesir untuk merapatkan jurang antara institusi pendidikan dan industri. Set data yang digunakan dalam kajian ini dikumpulkan melalui tinjauan yang diedarkan kepada graduan dan majikan merangkumi maklumat pelbagai kategori seperti latihan, kemahiran insaniah, kemahiran teknikal dan kemahiran bekeperluan tinggi. Masing-masing dengan nilai tertentu yang menunjukkan tahap latihan yang diterima oleh graduan. Lima model klasifikasi pembelajaran mesin digunakan iaitu DT, Gaussian NB, LR, RF dan SVM. Model DT mencapai ketepatan tertinggi dalam meramalkan kebolehpasaran diikuti oleh LR dan SVM, menunjukkan keberkesanan model ramalan dalam menyelaraskan graduan dengan permintaan pasaran kerja. Walaubagaimanapun, bilangan dataset yang kecil merupakan limitasi terhadap kajian ini (ElSharkawy et al. 2022).

Kajian oleh Usita (2022) menganalisis kebolehpasaran graduan melalui model klasifikasi untuk mengukur kejayaan program Pendidikan Tinggi. Dengan menggunakan teknik perlombongan data iaitu *Bayes Net* dan DT, kajian ini membina model untuk analisis pekerjaan graduan, membandingkan pelbagai algoritma klasifikasi menggunakan *Waikato Environment for Knowledge Analysis* (WEKA) pada set data yang terdiri daripada 1,489 graduan selama tiga tahun. Selain itu, penerokaan data menggunakan teknik Peraturan Sekutuan dengan *Apriori* memberikan maklumat lanjut mengenai kebolehpasaran graduan, menawarkan maklumat berharga mengenai ramalan, visualisasi dan algoritma pengelasan untuk menganalisis status pekerjaan graduan merentasi pelbagai sektor. Penyelidikan ini menekankan kepentingan menggunakan WEKA untuk klasifikasi dan visualisasi set data bagi mengekstrak pengetahuan untuk meningkatkan prestasi pelajar dan mencadangkan mengaitkan kebolehpasaran graduan dengan penilaian kurikulum dan penilaian prestasi untuk memperbaiki dasar pendidikan. Kajian ini mendapati DT memberikan ketepatan ramalan sebanyak 73.8% berbanding Bayes Net iaitu 73.6% (Usita 2022).

Raman dan Pramod (2022) menggunakan model ramalan untuk mengenalpasti faktor utama yang mempengaruhi peluang pekerjaan kepada graduan Sarjana Pentadbiran Perniagaan di India. 13 model ramalan klasifikasi dibangunkan menggunakan 7 atribut binari. Kajian ini mendapati model klasifikasi gabungan seperti *Bagging, Boosting* dan *ExtraTree* memberikan ketepatan ramalan yang lebih baik (90.30%) dan setara berbanding model klasikasi tunggal seperti NB dan LR (88%). Analisa juga menunjukkan kemahiran insaniah dan penglibatan kokurikulum merupakan faktor major kepada kebolehpasaran. Kajian ini juga membuktikan bahawa kelayakan akademik sahaja tidak memadai untuk meningkatkan peluang pekerjaan dan memperolehi gaji yang lebih baik (Raman & Pramod 2022).

Oumaima et al. (2022) memperkenalkan kaedah baharu untuk meramal kebolehpasaran pelajar dengan mempertimbangkan kedua-dua faktor berkaitan pelajar dan latihan industri, menjadikan proses ramalan lebih tepat dan selari dengan konteks kajian. Kajian ini menggunakan model ramalan yang lebih kompleks iaitu Model Penggalakan Kecerunan seperti *XGBoost*, *CatBoost*, dan *LightGBM* untuk meningkatkan ketepatan ramalan kebolehpasaran. Data yang digunakan ialah data tinjauan pelajar dari Universiti Princess Nourah bint Abdulrahman, Arab Saudi tahun 2019 – 2021. Tinjauan ini merangkumi 3 bahagian iaitu maklumat pelajar, maklumat latihan industri dan maklumat pekerjaan mengandungi 18 atribut keseluruhannya. Hasil kajian mendapati Model *LightGBM* menunjukkan prestasi terbaik dalam meramalkan kebolehpasaran graduan, terutamanya apabila menggunakan ciri-ciri berkaitan latihan industri, dengan ketepatan yang lebih tinggi (77.53%) berbanding model lain seperti *XGBoost* dan *CatBoost*. Atribut berkaitan latihan industri seperti gred latihan industri, tempoh dan tawaran pekerjaan dari tempat latihan industri adalah yang paling penting dalam meramalkan kebolehpasaran, lebih penting daripada ciri-ciri berkaitan pelajar. Dengan memahami atribut latihan industri yang paling signifikan mempengaruhi kebolehpasaran, kajian ini membantu universiti memperbaiki program latihan industri mereka, menjadikannya lebih berkesan dalam mempersiapkan pelajar untuk pasaran pekerjaan (Oumaima et al. 2022).

Kajian oleh Baffa et al. (2023) menyatakan bahawa IPT di Nigeria semakin memberi tumpuan kepada kebolehpasaran pelajar disebabkan kepentingan

menyediakan tenaga kerja terlatih untuk pembangunan ekonomi. Sehubungan itu satu model ramalan dibangunkan menggunakan atribut akademik dan aktiviti kokurikulum seperti CGPA, keputusan latihan industri, tempat latihan industri, jantina, kumpulan kesatuan dan tahun tamat pengajian untuk meramalkan kebolehpasaran pelajar. Analisa kepentingan ciri menggunakan algoritma RF mendapati keputusan latihan industri, CGPA dan tempat latihan industri sebagai atribut paling signifikan dalam meramalkan kebolehpasaran pelajar. Algoritma RF memberikan ketepatan ramalan sebanyak 98% berbanding LR dan DT. Sumbangan kajian ini membantu dalam meramalkan kebolepasaran pelajar sebelum tamat pengajian dan menekankan kepentingan prestasi akademik serta latihan industri sebagai faktor utama dalam meningkatkan peluang pekerjaan pelajar (Baffa et al. 2023).

Jadual 2.3 menunjukkan ringkasan kajian literatur berkaitan ramalan kebolehpasaran pelajar seperti yang dikupas lanjut dalam Bab 2.5.1 dan 2.5.2 ini.

Jadual 2.3 Ringkasan Kajian Literatur Ramalan Kebolehpasaran Graduan

Rujukan/ Tahun	Objektif Kajian	Algoritma Kajian	Sumber Data	Atribut Pilihan	Algoritma Terbaik	Ketepatan
(Sapaat et al. 2011)	Perbandingan ketepatan antara kelompok Algoritma Bayes dan Algoritma Pohon	Bayes (WAODE, AODE, NB, *Simple* NB, *Bayes Net*, AODEsr, HNB) Pohon (J48, RF, LAD *Tree*, REP*Tree*, *Random Tree*)	Data SKPG 2009	Umur, Jantina, Universiti, Peringkat, Bidang, CGPA, Status Pekerjaan, Kemahiran IT, BM, BI, Pengetahuan AM, Sektor, Negeri, Lain-lain Bahasa, Komunikasi, Analitikal, Penyelesaian Masalah, Pemikiran Kritis, Nilai Positif	DT (J48)	92.3%
(Rahman et al. 2017)	Model klasifikasi untuk meramal dan menilai atribut graduan dalam bidang akademik yang memenuhi keperluan industri	KNN NB DT ANN LR SVM	Data SKPG UTM, UKM, USM, UM dan UPM 2014 – 2015	41 atribut dari SKPG	KNN	97.78%
(Othman et al. 2018)	Mengenalpasti faktor yang mempengaruhi kebolehpasaran dan perbandingan model klasifikasi	DT (J48) ANN (MLP) SVM (SMO)	Data SKPG UKM 2011 – 2017	Umur, fakulti, bidang pengajian, kokurikulum, status perkahwinan, latihan industri dan kemahiran bahasa Inggeris.	DT (J48)	66%
(Pauzi et al. 2021)	Meramalkan status pekerjaan berdasarkan faktor kebolehpasaran serta memprofilkan kepuasan pelajar terhadap aktiviti kokurikulum dan kemahiran ICT	LR DT ANN	Data SKPG 2015 – 2018	Umur, kursus, CGPA, tahun konvokesyen, bahasa Inggeris, bahasa ketiga, pengalaman dengan kaunseling, fasilitator, kemudahan, pembiayaan, jantina, kemahiran ICT, akses perpustakaan, status OKU, bangsa, mod pengajian, sistem, dan pengalaman kerja.	LR	61.02%

bersambung…

...sambungan

(Haque et al. 2024)	Menyelidik ciri-ciri yang signifikan yang mempengaruhi keupayaan pelajar mendapat pekerjaan.	LR RF NB SVM XGB ANN	Data SKPG MMU 2021 Data Graduan MMU 2021	Maklumat demografi, maklumat akademik, GPA dan maklumat SKPG	ANN	80%
(Casuat et al. 2020)	Meramal kebolehpasaran melalui pendekatan pembelajaran	LR RF DT NB SVM XGB ANN	Data agensi latihan di Filipina	Keterampilan, cara bercakap, keadaan fizikal, kesedaran mental, keyakinan, kemampuan menyampaikan pendapat, kemahiran komunikasi, tahap prestasi, GPA, nama, program dan nombor pelajar.	SVM	91.22%
(K & H K 2020)	Membangunkan model ramalan kebolehpasaran berasaskan pencapaian akademik dan kemahiran mendapatkan kerja.	ANN LR DT KNN SVM NB	Data graduan Kejuruteraan Sains Maklumat dan Kejuruteraan Sains Komputer, India	CGPA, kursus atas talian dilengkapkan, latihan industri, projek teknikal (tesis).	ANN	87.42%
(He et al. 2022)	Meramal status pekerjaan dan menentukan faktor mempengaruhi kebolehpasaran pelajar.	Light GBM RF SVM	Data graduan *Graduate School of Hunan Normal University*	Jantina, major, akademik, pandangan politik, pendidikan, kesukaran pekerjaan, pelajar bandar/ luar bandar, bangsa	Light GBM	83%

bersambung…

...sambungan

(Panityakul et al. 2022)	Meramal kebolehpasaran graduan menggunakan statistik dan model pembelajaran mesin	LR DT RF KNN	Data graduan bidang statistik dan berkaitan di universiti-universiti Thailand tahun 2013-2017	Kedudukan institusi, tahap pendidikan, major, jenis institusi, kategori institusi dan wilayah	RF	70.55%
(Maaliw et al. 2022)	Meramal kebolehpasaran graduan daripada data kognitif dan bukan kognitif menggunakan model gabungan	RF SVM MLP DT KNN NB LR	Data graduan Kejuruteraan Elektronik di Filiphina 2014 – 2019	Grit, sijil profesional, kemahiran teknikal, kemahiran komunikasi, tabiat dan sikap kerja, kepimpinan.	RF RF+SVM+NB	91.36% 93%
(ElSharkawy et al. 2022)	Membangun model ramalan kebolehpasaran untuk merapatkan jurang antara IPT dan majikan.	DT Gaussioan NB LR RF SVM	Data graduan IT dan majikan di Mesir	Latihan, kemahiran insaniah, kemahiran teknikal, kemahiran spesifik	DT	100%
(Usita 2022)	Analisa kebolehpasaran graduan sebagai penilaian program Pendidikan Tinggi	*Bayes Net* DT	Data graduan College of Arts, Science, and Technology of Occidental Mindoro State College 2016 – 2018		DT	75.76%

bersambung…

...sambungan

(Raman & Pramod 2022)	Mengenalpasti faktor utama yang mempengaruhi peluang pekerjaan graduan menggunakan model ramalan	Bagging, Gradient Boosting, ExtraTrees, Linear SVC, AdaBoost, Ridge, Bernoulli DT KNN RF SVC LR Gaussion NB	Data graduan Sarjana Pentadbiran Perniagaan di India	Bidang pengajian, skor kelas 10, skor kelas 12, skor graduasi, kemahiran insaniah, penglibatan ko-kurikulum, penglibatan sosial	Bagging	95.84%
(Oumaima et al. 2022)	Meramal kebolehpasaran berdasarkan hubungan latihan industri dan profil akademik.	LightGBM CatBoost XGBoost AdaBoost DT LR KNN SVM NB	Data tinjauan graduan Universiti Princess Nourah bint Abdulrahman 2019 - 2021	GPA, LinkedIn, Sijil Profesional, bilangan aktiviti kokurikulum, gred, bidang, tempoh, jenis, metodologi, organisasi, sektor latihan industri, tawaran pekerjaan dan kepuasan menjalani latihan industri	LightGBM	77.53%
(Baffa et al. 2023)	Meramalan kebolehpasaran pelajar menggunakan atribut akademik dan aktiviti kokurikulum.	LR DT RF	Graduan Fakulti Komputer, Universiti Dutse 2016 – 2021	CGPA, penempatan latihan industri, keputusan latihan industri, jantina, penglibatan dalam kesatuan	RF	98%

2.6 RUMUSAN

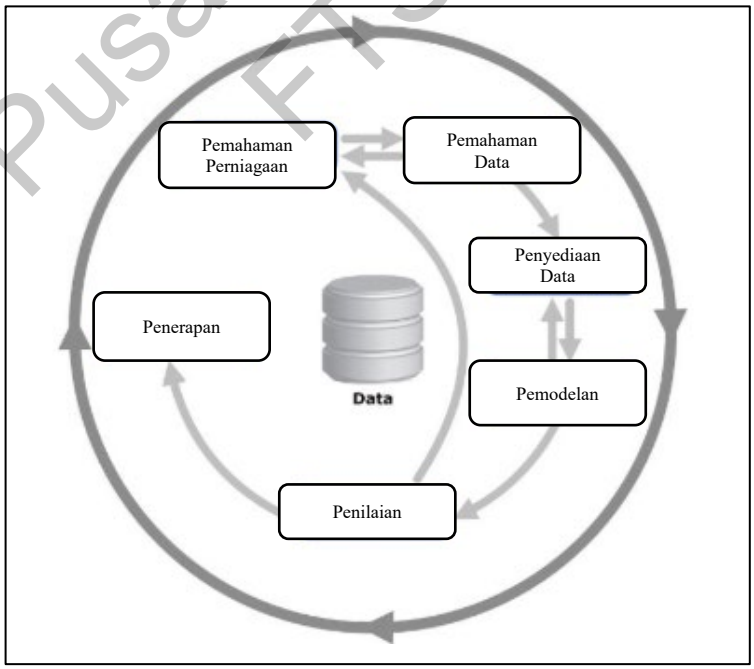
Secara umumnya, kebanyakan kajian meletakkan matlamat untuk membangunkan model ramalan dan mengenalpasti atribut utama yang mempengaruhi kebolehpasaran pelajar. Model ramalan klasifikasi adalah kaedah yang paling kerap digunakan dan memberikan ketepatan ramalan yang baik. Algoritma yang paling kerap digunakan untuk pembangunan model ialah DT, LR, SVM, NB, RF, KNN, ANN dan GBM. Kelompok algoritma pokok seperti DT dan RF pula memberikan hasil ramalan yang lebih baik berbanding algoritma lain. Perkara ini juga selari dengan log kajian yang dibuat oleh Encarnacian dan Cruz pada tahun 2021. Kajian ini juga mendapati beberapa parameter yang kerap digunakan untuk meramal kebolehpasaran pelajar ialah purata nilai gred kumulatif (CGPA), jantina, kemahiran teknikal, komunikasi, kemahiran analitik dan membuat keputusan, kemahiran menyelesaikan masalah, pemikiran kritis, aktiviti kokurikulum, umur, faktor psikomotor seperti tingkah laku dan kehadiran, serta penempatan latihan/praktikal (Encarnacion & Cruz 2021).

BAB III

METODOLOGI

3.1 PENGENALAN

Metodologi kajian ini merangkumi langkah-langkah yang diambil untuk membangunkan dan menilai model ramalan kebolehpasaran pelajar tajaan kerajaan menggunakan kaedah klasifikasi. Kajian ini mengikuti pendekatan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang terdiri daripada enam fasa iaitu pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian, dan penerapan. Model CRISP-DM adalah seperti dalam Rajah 3.1.



Rajah 3.1 Model CRISP-DM

Sumber : Kannengiesser & Gero 2023

Fasa pemahaman perniagaan berfokuskan kepada proses mengenal pasti keperluan kajian, menetapkan matlamat dan menentukan hala tuju kajian. Perkara ini telah pun diterangkan secara terperinci dalam Bab I. Fasa pemahaman data melibatkan pengumpulan, pemilihan dan penerokaan data. Fasa penyediaan data pula merangkumi aktiviti seperti integrasi data, pembersihan data, pembentukan ciri, pemilihan ciri dan transformasi data untuk membina set data akhir daripada data mentah.

Seterusnya ialah fasa pemodelan yang merupakan fasa utama dalam proses pembangunan model ramalan. Dalam fasa ini, data yang telah tersedia digunakan untuk membina dan melatih model klasifikasi menggunakan algoritma pembelajaran mesin yang dikenalpasti. Fasa penilaian bertujuan untuk menentukan sejauh mana model yang dibina dapat memberikan ramalan atau klasifikasi yang tepat dan berguna. Fasa penilaian melibatkan beberapa aktiviti utama, termasuk penggunaan metrik penilaian, analisis hasil model, dan pemilihan model terbaik berdasarkan prestasi. Fasa terakhir ialah fasa penerapan di mana model yang terbaik akan digunakan sebagai hasil kajian untuk membuat ramalan.

3.2 FASA PEMAHAMAN DATA

3.2.1 Pengumpulan Data

Kajian ini menggunakan data dari pangkalan data JPA iaitu Sistem eSILA yang digunakan untuk merekod dan memantau pelajar di bawah tajaan JPA. Data pelajar tajaan yang menamatkan pengajian dari tahun 2016 sehingga 2022 dikumpulkan sebagai kajian kes.

Maklumat pekerjaan pelajar pula diperolehi daripada Laporan Kajian Pengesanan Graduan Tajaan JPA 2017 – 2023. Kajian Pengesanan Graduan Tajaan JPA merupakan kajian dalaman yang dibuat oleh JPA secara atas talian untuk mengetahui status semasa graduan tajaan JPA dalam tempoh setahun selepas menamatkan pengajian. Kajian ini dibuat untuk sebagai laporan kepada pihak pengurusan berkenaan kadar kebolehpasaran graduan tajaan dan pengukuran keberhasilan program penajaan. Data kajian diperolehi daripada soal selidik yang dihantar kepada pelajar secara atas

talian, semak silang dengan Sistem eSILA, Sistem Pengurusan Maklumat Sumber Manusia (HRMIS) dan rekod Program *Scholarship Talent Attraction and Retention* (STAR).

JPA sebagai pemilik data telah memberi keizinan untuk menggunakan data-data tersebut bagi tujuan kajian ini. Jadual 3.1 menunjukkan jumlah data yang telah diperolehi dari Sistem eSILA mengikut tahun tamat pengajian berjumlah 48 952. Secara umumnya, bilangan data pelajar tamat mengikut tahun adalah hampir sama tetapi terdapat peningkatan pada tahun 2022. Ini disebabkan beberapa faktor iaitu peningkatan bilangan penajaan dan pindaan tempoh pengajian kepada pelajar-pelajar yang sepatutnya tamat pada tahun 2020 dan 2021 disebabkan penularan wabak Covid-19.

Jadual 3.1 Bilangan Data Mengikut Tahun Tamat

Tahun Tamat	**Bilangan Data**
2016	6039
2017	6065
2018	6995
2019	6690
2020	6617
2021	6990
2022	9556
Jumlah	**48 952**

Sumber: Sistem eSILA JPA

Data kajian dimuat turun dan disimpan dalam bentuk *Comma Separated Variables* (CSV). Data ini mengandungi maklumat peribadi, maklumat penajaan, maklumat perhubungan, maklumat akademik, maklumat waris dan maklumat kebolehpasaran. Senarai atribut bagi setiap maklumat yang diperolehi adalah seperti di Jadual 3.2.

Jadual 3.2 Senarai Atribut Data Diperolehi

Bil.	Kategori	Atribut
1	Maklumat Peribadi	NoKP, Nama, Jantina, Tarikh Lahir, Negeri Lahir, Agama, Keturunan, Bumiputra, Taraf Kahwin, Emel
2	Maklumat Penajaan	Peringkat, Institusi, Fakulti, Negara, Bidang. Kluster Bidang, Sesi Mula, Sesi Tamat, Tahun Tawar, Tarikh Tawar, Tahun Taja, Tarikh Taja, Tahun Akhir, Tarikh Jangka Tamat, Tarikh Tamat, Tahun Lapor, Tarikh Lapor, Program Tajaan, Status Pengajian, Keterangan Status Pengajian, Tempoh Tajaan
3	Maklumat Perhubungan	AlamatTetap1, AlamatTetap2, PoskodTetap, BandarTetap, NegeriTetap, TelTetap, AlamatSMenyurat1, AlamatSMenyurat2, PoskodSMenyurat, BandarSMenyurat, NegeriSMenyurat, TelSMenyurat, AlamatJPN1, AlamatJPN2, AlamatJPN3, PoskodJPN, BandarJPN, NegeriJPN
4	Maklumat Akademik	CGPA, Keputusan
5	Maklumat Waris	Nama Bapa, No KP Bapa, Telefon Bapa, Hubungan Bapa, Pekerjaan Bapa, Gaji Bapa, Nama Ibu, No KP Ibu, Telefon Ibu, Hubungan Ibu, Pekerjaan Ibu, Gaji Ibu, Nama Penjaga, No KP Penjaga, Telefon Penjaga, Hubungan Penjaga, Pekerjaan Penjaga, Gaji Penjaga, Tanggungan
6	Maklumat Kebolehpasaran	StatusPekerjaan

Sumber: Sistem eSILA JPA

3.2.2 Pemilihan Data

Data mentah yang diperolehi mengandungi 72 atribut. Terdapat maklumat sensitif yang perlu digugurkan bagi melindungi data peribadi pelajar dan juga penaja. Maklumat-maklumat yang terlalu spesifik dan tidak relevan juga perlu digugurkan bagi memastikan model yang dibangunkan boleh digeneralisasikan dalam membuat ramalan. Senarai atribut yang digugurkan adalah seperti Jadual 3.3.

Jadual 3.3 Senarai Atribut Data Mentah Digugur

Bil.	Sebab Digugurkan	Senarai Atribut
1	Mengandungi maklumat peribadi	NoKP, Nama, Tarikh Lahir, Negeri Lahir, Taraf Kahwin, AlamatTetap1, AlamatTetap2, PoskodTetap, BandarTetap, TelTetap, AlamatSMenyurat1, AlamatSMenyurat2, PoskodSMenyurat, BandarSMenyurat, NegeriSMenyurat, TelSMenyurat, AlamatJPN1, AlamatJPN2, AlamatJPN3, PoskodJPN, BandarJPN, NegeriJPN, Nama Bapa, No KP Bapa, Telefon Bapa, Hubungan Bapa, Nama Ibu, No KP Ibu, Telefon Ibu, Hubungan Ibu, Nama Penjaga, No KP Penjaga, Telefon Penjaga, Hubungan Penjaga,
2	Maklumat spesifik	Sesi Mula, Sesi Tamat, Tahun Tawar, Tarikh Tawar, Tahun Taja, Tarikh Taja,
3	Maklumat tidak relevan	Tahun Lapor, Tarikh Lapor, Tempoh Tajaan, Status Pengajian, Keterangan Status Pengajian.

Hanya atribut yang mempunyai maklumat umum demografi, pengajian, penajaan, sosioekonomi keluarga dan kebolehpasaran sahaja yang dikekalkan. Daripada 72 atribut asal, hanya 24 atribut yang akan diproses ke fasa seterusnya. Senarai atribut data mentah adalah seperti di Jadual 3.4.

Jadual 3.4 Senarai Atribut Data Mentah

Bil	Atribut	Deskripsi
1	Jantina	Jantina pelajar
2	TarikhLahir	Tarikh lahir pelajar
3	NegeriLahir	Negeri lahir pelajar
4	Bumiputra	Status Bumiputra
5	Peringkat	Peringkat pengajian
6	Institusi	Institusi pengajian
7	Negara	Negara pengajian
8	Bidang	Bidang pengajian
9	Tarikh Jangka Tamat	Tarikh jangkaan tamat
10	Tarikh Tamat	Tarikh tamat sebenar
11	Program	Program tajaan
12	Tajaan	Jenis tajaan
13	Status Pengajian	Status pengajian semasa
14	Keterangan Status Pengajian	Keterangan status pengajian
15	CGPA	CGPA
16	Keputusan	Jenis keputusan selain CGPA
17	NegeriTetap	Negeri alamat tetap
18	Pekerjaan Bapa	Pekerjaan bapa
19	Gaji Bapa	Pendapatan bapa
20	Pekerjaan Ibu	Pekerjaan ibu
21	Gaji Ibu	Pendapatan ibu
22	Pekerjaan Penjaga	Pekerjaan penjaga
23	Gaji Penjaga	Pendapatan penjaga
24	StatusPekerjaan	Status pekerjaan pelajar

3.3 FASA PENYEDIAAN DATA

Fasa penyediaan data atau pra-pemprosesan data adalah langkah penting dalam analisis data dan pembelajaran mesin. Ia bertujuan untuk memastikan data yang digunakan untuk melatih model adalah bersih, relevan, dan dalam format yang sesuai untuk analisis. Langkah ini penting kerana data mentah sering mengandungi maklumat tidak

konsisten, ralat dan maklumat tidak relevan yang boleh menjejaskan prestasi model ramalan. Matlamat utama pra-pemprosesan data adalah untuk meningkatkan kualiti data, kebolehgunaan, kebolehcapaian dan kebolehgunaan semula (Ye & Wang 2023). Ia melibatkan beberapa tugas seperti integrasi data, pemilihan data, pembersihan data, transformasi data dan pengurangan data.

3.3.1 Integrasi Data

Proses integrasi data dijalankan untuk menggabungkan set data yang diperolehi daripada beberapa sumber. Dalam kajian ini, data yang diperolehi daripada Laporan Kajian Pengesanan Graduan Tajaan JPA dan Sistem eSILA digabungkan bagi mendapatkan satu set data yang besar dan lengkap bagi tujuan analisis dan pembangunan model ramalan kebolehpasaran pelajar tajaan JPA. Nombor kad pengenalan pelajar digunakan sebagai atribut rujukan semasa proses integrasi ini memandangkan setiap data berada pada set yang berbeza mengikut tahun. Data-data ini digabungkan dan disimpan dalam format CSV bagi tujuan analisa dan pembangunan model. Sebanyak 48 952 data graduan yang mengandungi 24 ciri telah diperolehi selepas proses integrasi.

3.3.2 Pembersihan Data

Pembersihan data adalah proses penting dalam kajian berasaskan data yang melibatkan tindakan pengesanan dan pembetulan ralat data. Proses pembersihan ini termasuklah menggantikan data hilang, data tidak konsisten, data tidak tepat, data berulang dan menghapuskan data tidak relevan. Tujuannya adalah untuk meningkatkan kualiti data yang akan mempengaruhi hasil yang diperolehi (Parulian & Ludäscher 2023).

Proses ini memerlukan pengetahuan domain untuk mengenal pasti dan membaiki kesilapan serta mengesahkan pembetulan. Penggunaan kaedah statistik seperti nilai purata, median dan mod untuk menggantikan data ralat boleh meningkatkan kecekapan dan keberkesanan proses pembersihan data (Zou 2022). Laporan kualiti data disediakan untuk mengenalpasti atribut-atribut yang perlu diberi perhatian untuk proses pembersihan data seperti di Jadual 3.5 dan Jadual 3.6.

Jadual 3.5 Laporan Kualiti Data Berterusan (*Continuous*)

Atribut	Bilangan Data	Data Hilang	Purata	Sisihan Piawai	Nilai Min.	25%	50%	75%	Nilai Maks.
Gaji Bapa	47752	1200	2797	3960	0	0	1549	4000	105325
Gaji Ibu	47752	3020	1587	2902	0	0	0	2500	70000
Gaji Penjaga	815	48137	1917	2423	0	0	1100	2681	18484

Jadual 3.5 merupakan laporan kualiti bagi data jenis berterusan. Laporan ini memaparkan maklumat statistik bagi data jenis berterusan di dalam data mentah yang diterima. Perkara yang dilihat ialah atribut, bilangan data, jumlah data hilang, nilai purata keseluruhan data, sisihan piawai, nilai minimum (Nilai Min.), nilai kuartil pertama (25%), nilai kuartil kedua atau median (50%), nilai kuartil ketiga (75%) dan nilai maksimum di dalam data. Berdasarkan laporan kualiti data, data hilang bagi gaji bapa dan gaji ibu digantikan dengan nilai purata manakala rekod data yang ditunjukkan gaji penjaga adalah terlalu sedikit iaitu 815 sahaja. Oleh itu atribut ini akan digugurkan kerana bilangan data yang terlalu sedikit.

Jadual 3.6 Laporan Kualiti Data Kategori

Atribut	Bilangan Data	Data Hilang	Kardinaliti	Nilai Mod	Kekerapan Mod	Peratusan Mod
Jantina	48807	145	2	Perempuan	32264	66.11
TarikhLahir	48807	145	6041	Tiada Rekod	53	0.11
NegeriLahir	48807	145	17	Johor	5889	12.07
Bumiputra	48452	500	2	Bumiputra	40936	84.49
Peringkat	48952	0	2	Ijazah	44188	90.27
Institusi	48952	0	750	Universiti Putra Malaysia	3431	7.01
Negara	48952	0	19	Malaysia	42709	87.25
Bidang	48952	0	2046	Perakaunan	2490	5.09
Tarikh Jangka Tamat	48719	233	498	31/08/2022	6502	13.35
Tarikh Tamat	48719	233	594	31/08/2022	6461	13.26
Program	48952	0	71	Institut Pengajian Tinggi	36007	73.56
Tajaan	48952	0	5	Biasiswa	26457	54.05
Status Pengajian	48697	255	13	Tamat Pengajian Dengan Jaya	35430	72.76
Keterangan Status Pengajian	48697	255	39	Tamat Pengajian Dengan Jaya	35051	71.98

bersambung…

...sambungan

CGPA	42478	6474	1616	Lulus	2330	5.49
Keputusan	2083	46869	59	Tiada rekod	1982	95.15
NegeriTetap	48924	28	17	Selangor	9716	19.86
Pekerjaan Bapa	47752	1200	6	Sendiri	12962	27.14
Pekerjaan Ibu	45932	3020	6	Tidak Bekerja	12647	27.53
Pekerjaan Penjaga	804	48148	10	Swasta	39	4.85

Bagi data jenis kategori pula seperti yang ditunjukkan dalam Jadual 3.6, nilai mod akan digunakan untuk mengisi nilai hilang bagi atribut jantina, tarikh lahir, negeri lahir, bumiputra, tarikh jangka tamat, tarikh tamat, status pengajian, keterangan status pengajian, negeri tetap, pekerjaan bapa dan pekerjaan ibu. Bagi CGPA dan Keputusan pula, atribut ini perlu digabungkan untuk mewujudkan ciri baru kerana kedua-dua data ini mewakili prestasi akademik pelajar.

3.3.3 Transformasi Data

Transformasi data adalah proses mengubah data mentah ke dalam format yang sesuai untuk analisis atau pembelajaran mesin. Proses ini termasuk pelbagai teknik untuk mengubah nilai, struktur atau jenis data agar data tersebut lebih bermakna, relevan dan dapat digunakan oleh model pembelajaran mesin. Transformasi data adalah langkah penting dalam persiapan data kerana ia membantu dalam meningkatkan kualiti data dan prestasi model (Cheng 2022). Teknik transformasi data yang akan dibuat dalam kajian ini ialah pembentukan ciri baru, penggabungan ciri, penukaran kategori dan penskalaan.

Pembentukan ciri baru ialah mencipta ciri baru yang tidak terdapat dalam set data asal tetapi boleh memberikan maklumat tambahan (Maaliw et al. 2022). Dalam kajian ini, ciri Umur Tamat dibentuk dari atribut Tarikh Lahir dan Tarikh Tamat. Seterusnya maklumat umur ini akan dikelompokan mengikut julat yang bersesuaian. Julat umur ini akan menggantikan atribut Tarikh Lahir dan Tarikh Tamat untuk memberi input yang lebih umum dan relevan (Othman et al. 2018).

Seterusnya ialah pembentukan julat pendapatan keluarga. Atribut Gaji Ayah dan Gaji Ibu dicampurkan untuk mendapatkan jumlah pendapatan isi rumah. Jumlah

pendapatan isi rumah ini pula dikategorikan mengikut julat pendapatan isi rumah kategori B40, M40 dan T20 sebagaimana yang dilaporkan oleh Laporan Survei Pendapatan Isi Rumah 2022, Jabatan Perangkaan Malaysia. Julat Pendapatan ini akan menggantikan atribut Gaji Ayah dan Gaji Ibu yang akan memberikan input berkenaan latar belakang sosioekonomi pelajar.

Atribut Tarikh Tamat, Tarikh Jangka Tamat dan Keterangan Status Pengajian pula digunakan untuk membentuk ciri Tamat Mengikut Masa (*Graduate On Time* – GOT). Manakala bagi atribut CGPA dan Keputusan, kedua-dua atribut ini mewakili keputusan akademik pelajar. Atribut ini digabungkan dan dipetakan untuk membentuk ciri Had Kecemerlangan Akademik (HKA) kerana kepelbagaian ukuran prestasi akademik mengikut institusi. Sebagai panduan, JPA juga telah menetapkan HKA yang berbeza-beza mengikut peringkat, institusi, bidang dan negara pengajian. Pelajar yang mendapat keputusan di bawah HKA dikenakan amaran dan boleh ditamatkan penajaan.

Berdasarkan data mentah, terdapat 2046 nama bidang pengajian yang direkodkan. Bagi tujuan penyeragaman dan mengurangkan bilangan dimesti data, nama bidang ini dipetakan semula untuk mewujukan ciri baru iaitu Bidang Pengajian NEC dan Bidang Perincian NEC mengikut senarai di dalam Kod Pendidikan Nasional 2020 (NEC). Pembentukan ciri ini penting bagi tujuan generalisasi memandangkan nama bidang pengajian adalah berbeza-beza mengikut institusi. Ciri baru ini juga akan menggantikan atribut Bidang Pengajian data asal.

Selain itu, atribut Nama Institusi juga dikelompokkan semula mengikut nama umum tanpa mengikut cawangan. Misalnya Universiti Teknologi MARA dikelompokkan kepada UiTM sahaja. Manakala bagi atribut program, nama program dikelompokkan semula mengikut peringkat penajaan untuk penyeragaman memandangkan nama-nama program ini sering berubah mengikut dasar semasa. Senarai ciri baru yang dibentuk dan kepentingan ciri adalah seperti di Jadual 3.7.

Jadual 3.7 Senarai Ciri Baru

Bil	Ciri Baru	Atribut Asal	Kepentingan
1	Julat Umur	TarikhLahir, Tahun Tamat	Faktor kebolehpasaran
2	Nama Institusi	Institusi	Penyeragaman dan pengurangan dimensi data
3	Jenis Institusi	Institusi	Penyeragaman dan pengurangan dimensi data
4	Bidang NEC	Bidang	Penyeragaman dan pengurangan dimensi data
5	Bidang Perincian NEC	Bidang	Penyeragaman dan pengurangan dimensi data
6	Kumpulan Program	Program	Penyeragaman dan pengurangan dimensi data
7	Julat Pendapatan	Gaji Bapa, Gaji Ibu, Gaji Penjaga	Penyeragaman dan pengurangan dimensi data
8	GOT	Tarikh Tamat, Tarikh Jangka Tamat, Keterangan Status Pengajian	Memberikan maklumat status pengajian pelajar
9	HKA	CGPA Semasa, Keputusan	Memberikan maklumat prestasi pelajar

Selepas proses integrasi, pembersihan dan transformasi, bilangan ciri yang tinggal adalah sebanyak 19 atribut daripada 24 atribut asal. Senarai atribut, deskripsi dan taburan data atribut adalah seperti yang disenaraikan dalam Jadual 3.8.

Jadual 3.8 Senarai Atribut, Deskripsi dan Taburan Data

Bil	Atribut	Deskripsi	Taburan Data
1	Jantina	Jantina pelajar	Lelaki (16 688), Perempuan (32 264)
2	NegeriLahir	Negeri lahir pelajar	Johor (5890), Kedah (3855), Kelantan (4746), Lain-lain (288), Melaka (1718), N.Sembilan (1957), Pahang (2903), Perak (4843), Perlis (477), P.Pinang (3037), Sabah (2126), Sarawak (2788), Selangor (5668), Terengganu (2968), K.Lumpur (5581), Labuan (96), Putrajaya (16)
3	Bumiputra	Status Bumiputra	Bumiputra (40 936), Bukan Bumiputra (8016)
4	Peringkat	Peringkat pengajian	Ijazah (44 188), Diploma (4764)

bersambung…

...sambungan

5	Negara	Negara pengajian	Amerika Syarikat (1224), Australia (653), India (345), Indonesia (69), Ireland (17), Jepun (855), Jordan (110), Kanada (125), Korea (413), Malaysia (42 709), Mesir (253), New Zealand (345), Perancis (258), Poland (17), Republik Czech (14), Republik German (162), Republik Rusia (40), Singapura (4), United Kingdom (1339)
6	Tajaan	Jenis tajaan	Biasiswa (26457), Pinjaman (22495)
7	NegeriTetap	Negeri alamat tetap	Johor (5741), Kedah (4179), Kelantan (4301), Lain-lain (4), Melaka (1741), N.Sembilan (2357), Pahang (2821), Perak (4534), Perlis (476), P.Pinang (2750), Sabah (1842), Sarawak (2571), Selangor (9830), Terengganu (3057), K.Lumpur (2402), Labuan (75), Putrajaya (271)
8	Julat Umur	Umur ketika pelajar menamatkan pengajian	19-21 (4455), 22-24 (37691), 25-29(6806)
9	Nama Institusi	Nama umum institusi tanpa cawangan	ADTEC (35), AIMST (38), ALAM (1), APU (2), AUCMS (31), CUCMS (20), CURTIN (19), DRB HICOM (81), HWU (1), IKM (18), IKTBN (16), IMU (311), INTI (1), IPTLN_AUS (653), IPTLN_CZ (14), IPTLN_FR (258), IPTLN_GRM (162), IPTLN_IND (345), IPTLN_INDO (69), IPTLN_IRE (17), IPTLN_JDN (110), IPTLN_JPN (855), IPTLN_KND (125), IPTLN_KOR (413), IPTLN_MSR (253), IPTLN_NZ (345), IPTLN_PLD (17), IPTLN_RS (40), IPTLN_SG (4), IPTLN_UK (1339), IPTLN-USA (1224), IUKL (1), JMTI (9), KDU (1), KKTM (152), Kolej Sunway (1), KUPTM (34), LUCM (9), MAHSA (82), MMMC (63), MMU (328), Monash University (269), MSU (80), Nilai University (3), NOTTINGHAM (109), NUMED (91), PIDC (7), PMC (31), POLITEKNIK (830), RUMC (2), SEGI (19), Sunway University (18), Swinburne University (23), Taylors University (56), UCSI (100), UIAM (2288), UiTM (9123), UKM (3576), UM (3376), UMK (252), UMP (575), UMS (895), UMT (855), UNIKL (76), UNIMAP (342), UNIMAS (1048), UNISZA (1507), UNITEN (699), UP (134), UPM (3502), UPNM (212), UPSI (238), USIM (1281), USM (2405), UTAR (7), UTEM (633), UTHM (1333), UTM (3350), UTP (548), UUM (1562)

bersambung…

...sambungan

10	Bidang NEC	Bidang pengajian mengikut Kod Pendidikan Nasional (NEC)	Sastera dan Kemanusiaan (2451), Sains Sosial, Kewartawanan dan Maklumat (2340), Perniagaan, Pentadbiran dan Perundangan (8445), Sains Semulajadi, Matematik dan Statistik (8307), Teknologi Maklumat dan Komunikasi (2611), Kejuruteraan, Pembuatan dan Pembinaan (14468), Pertanian, Perhutanan, Perikanan dan Veterinar (1218), Kesihatan dan Kebajikan (8681), Perkhidmatan (431)
11	Bidang Perincian NEC	Bidang perincian pengajian mengikut NEC	0210 (1), 0211 (81), 0212 (46), 0213 (67), 0214 (1), 0215 (27), 0216 (10), 0220 (4), 0222 (185), 0223 (1), 0224 (991), 0229 (66), 0231 (8), 0232 (851), 0233 (112), 0300 (1), 0311 (1052), 0312 (308), 0313 (325), 0314 (307), 0321 (3), 0322 (74), 0323 (270), 0411 (3942), 0412 (596), 0413 (414), 0414 (2415), 0415 (181), 0416 (11), 0417 (12), 0421 (874), 0500 (376), 0511 (1594), 0512 (1164), 0521 (327), 0522 (1), 0530 (14), 0531 (1199), 0532 (687), 0533 (449), 0539 (112), 0541 (1186), 0542 (1137), 0588 (61), 0611 (1121), 0612 (520), 0613 (970), 0700 (10), 0710 (926), 0711 (1944), 0712 (1190), 0713 (1624), 0714 (2835), 0715 (445), 0716 (1701), 0717 (73), 0719 (2), 0720 (188), 0721 (771), 0722 (36), 0723 (8), 0731 (645), 0732 (182), 0733 (165), 0734 (839), 0741 (728), 0788 (156), 0811 (669), 0812 (73), 0821 (102), 0831 (146), 0841 (228), 0911 (2511), 0912 (2594), 0913 (20), 0914 (453), 0915 (741), 0916 (1931), 0919 (1), 0923 (430), 1013 (115), 1014 (98), 1022 (205), 1031 (12), 1032 (1)
12	Kumpulan Program	Program tajaan JPA yang dikelompokan mengikut peringkat	Dermasiswa (2592), LSPM (8297), PIDN (38063)
13	Pekerjaan Bapa	Pekerjaan bapa	Kerajaan / GLC (11064), Meninggal Dunia (2922), Pesara (7789), Sendiri (12962), Swasta (11936), Tidak Bekerja (2279)
14	Pekerjaan Ibu	Pekerjaan ibu	Kerajaan / GLC (9895), Meninggal Dunia (577), Pesara (1889), Sendiri (2220), Swasta (4039), Tidak Bekerja (30332)
15	Julat Pendapatan	Julat pendapatan isi rumah	B40 (33828), M40 (10739), T20 (4385)
16	Jenis Institusi	Jenis institusi sama ada awam, swasta atau luar negara	IPTA (39995), IPTLN (6243), IPTS (2714)
17	GOT	Status sama ada tamat pengajian mengikut tempoh tajaan / *Graduate On Time* (GOT) atau tidak.	GOT (46 573), Tidak GOT (2379)

bersambung…

...sambungan

18	HKA	Status sama ada melepasi Had Kecemerlangan Akademik (HKA)yang ditetapkan JPA atau tidak.	Melepasi HKA (47 645), Tidak (1307)
19	StatusPekerjaan	Status pekerjaan pelajar / Kelas	Bekerja (32655), Tidak Bekerja (16297)

Berdasarkan Jadual 3.8, dapat dilihat variasi data yang lebih luas berbanding kajian-kajian terdahulu di mana data ini merangkumi 19 buah negara pengajian, 81 institusi pengajian, 9 bidang pengajian dan 85 bidang perincian pengajian. Data-data ini tidak diasingkan untuk mendapatkan variasi data yang lebih menyeluruh dan meluas terhadap model ramalan yang dibangunkan.

Langkah transformasi data yang seterusnya ialah penukaran kategori. Data-data yang diperolehi adalah jenis data kategori. Data-data ini perlu ditukarkan ke dalam bentuk data nombor supaya ia boleh digunakan oleh algoritma pembelajaran mesin. Teknik yang akan digunakan dalam kajian ini ialah Pengekodan Label (*Label Encoding*). Pengekodan Label menukarkan setiap kategori dalam satu ciri kepada nombor integer unik.

3.3.4 Pemilihan Ciri

Pemilihan ciri digunakan untuk menyingkirkan ciri yang tidak relevan dalam membina sebuah model. Ia membantu memilih ciri terbaik dan berguna dalam membina model. Dengan menggunakan ciri yang berkaitan boleh meningkatkan ketepatan model klasifikasi, memendekkan tempoh kajian dan juga membentuk konsep yang ringkas. Tujuan proses pemilihan ciri adalah untuk memilih ciri penting dan berguna untuk meningkatkan peratusan ketepatan model. Dengan cara ini, model yang dibina akan lebih tepat dan berkesan dalam menerangkan data serta menjalankan fungsi ramalan dengan lebih baik (Othman et al. 2018).

Dalam kajian ini, teknik Maklumat Keuntungan (*Information Gain*) digunakan untuk menilai dan memilih ciri-ciri penting daripada set data ini. Teknik ini mengukur ketergantungan antara setiap ciri dan sasaran, membolehkan ciri-ciri yang paling

bermaklumat untuk tujuan ramalan dikenal pasti (Maaliw et al. 2022). Ia berasaskan konsep entropi dari teori maklumat. Entropi mengukur ketidakpastian dalam satu set data. Maklumat Keuntungan mengukur penurunan entropi apabila data dibahagikan berdasarkan nilai sesuatu ciri. Pengiraan entropi *H* untuk satu set data *S* dengan *n* kelas ditunjukkan dalam Persamaan 3.1

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad \ldots(3.1)$$

di mana p_i adalah kebarangkalian bagi kelas *i*.

$$Maklumat\ Keuntungan\ (A) = H(S) - \sum_{v \in NilaiA} \frac{|S_v|}{|S|} H(S_v) \quad \ldots(3.2)$$

di mana:

- Nilai (A) adalah set semua nilai unik untuk ciri *A*.
- S_v adalah subset dari S di mana ciri *A* mempunyai nilai *v*.
- |S| adalah jumlah contoh dalam set data S.
- $|S_v|$ adalah jumlah contoh dalam set data S_v.

Teknik ini akan memberikan nilai skor kepada setiap ciri, ciri-ciri yang mempunyai nilai skor yang tinggi adalah lebih informatif dan relevan berbanding ciri-ciri yang lain. Pemilihan ciri dibuat berdasarkan nilai skor dan kepentingan kajian.

3.4 FASA PEMODELAN

Objektif utama kajian ini adalah untuk membangunan model ramalan kebolehpasaran pelajar tajaan kerajaan menggunakan kaedah klasifikasi. Model berkenaan akan dibangunkan menggunakan 48 952 data graduan dengan ciri pilihan yang telah dikenalpasti. Atribut status pekerjaan yang mengandungi maklumat bekerja atau tidak bekerja menjadi label kelas.

Algoritma pembelajaran mesin yang digunakan untuk pembangunan model ialah LR, NB, DT, RF dan GBM. Algoritma ini dipilih mengambil kira konteks kajian dan sifat data yang digunakan. LR merupakan algoritma yang ringkas dan mudah difahami. Koefisien dalam LR boleh digunakan untuk menjelaskan pengaruh setiap ciri terhadap kebarangkalian hasil.

NB boleh mengendalikan data jenis kategori dengan baik. Ia mudah dilaksanakan dan tidak memerlukan banyak sumber pengkomputeran. NB adalah cepat untuk dilatih dan diuji menjadikannya sesuai untuk set data kompleks serta mampu mengendalikan data dengan kelas yang tidak seimbang dengan baik.

DT juga mampu mengendali data jenis kategori dengan baik. Hasil model ini mudah difahami dan dijelaskan kepada pihak berkepentingan dalam bentuk visualisasi disamping dapat mengenal pasti interaksi antara ciri-ciri yang berbeza.

RF dan GBM pula merupakan algoritma gabungan yang lebih kompleks boleh memberikan ketepatan ramalan yang lebih tinggi dan mengurangkan masalah terlebih penyesuaian berbanding algoritma tunggal. RF dan GBM juga lebih berdaya tahan terhadap variasi di dalam set data dan boleh memberikan prestasi yang konsisten. Algoritma ini juga sesuai untuk menguruskan data yang besar dan kompleks.

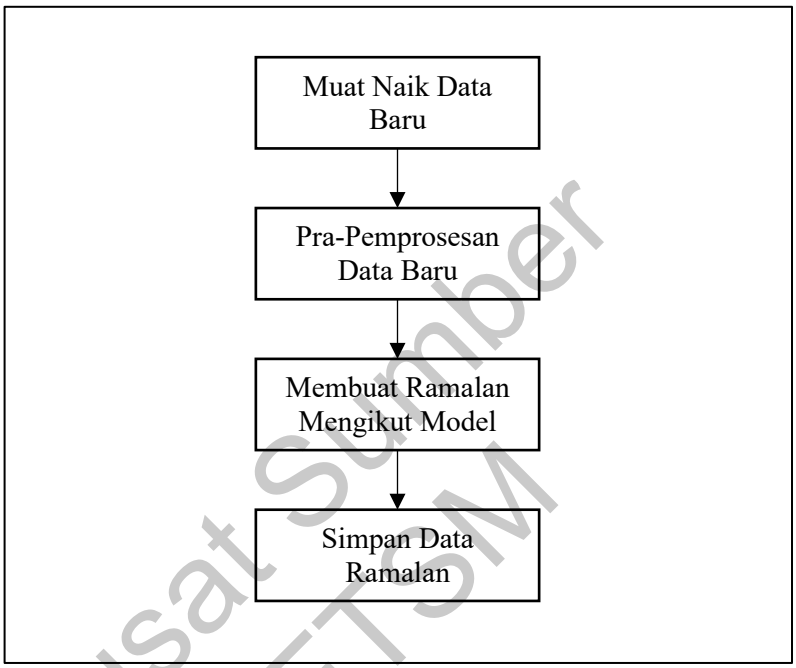
Dalam proses pemodelan, kaedah pengesahan silang lipatan digunakan bagi melatih dan menguji data. Ia dapat memastikan setiap data digunakan sebagai data latihan dan ujian pada satu masa dan memaksimumkan penggunaan set data. Nilai K-lipatan yang ditetapkan adalah 5.

3.5 FASA PENILAIAN

Prestasi model ramalan yang dibangunkan dinilai menggunakan matriks kekeliruan dan matriks penilaian. Matriks penilaian yang digunakan untuk menilai prestasi model ialah Ketepatan, Kejituan, Sensitiviti, Skor F1 dan AUC-ROC. Perincian berkaitan matriks ini telah dinyatakan dalam Bab 2.4.3. Prestasi model-model kemudiannya dibandingkan antara satu sama lain untuk mengenal pasti yang memberikan prestasi terbaik untuk meramal kebolehpasaran pelajar tajaan kerajaan.

3.6 FASA PENERAPAN

Selepas penilaian model dibuat dan dibandingkan, model terbaik disimpan untuk membuat ramalan kepada set data baru. Set data baru yang akan digunakan adalah senarai pelajar yang akan bergraduat pada tahun 2024. Carta alir penerapan model adalah seperti Rajah 3.2.



Rajah 3.2 Penerapan Model Ramalan

Set data baru dimuat naik ke dalam persekitaran analisis. Set data melalui pra-pemprosesan iaitu pembersihan dan transformasi data. Seterusnya proses ramalan dibuat untuk melabelkan data menggunakan model latihan yang dibangunkan. Data ini dilabel sama ada “Bekerja” atau “Tidak Bekerja” sebagai ramalan kebolehpasaran pelajar. Data yang telah dilabel disimpan untuk analisis deskriptif dan cadangan tindakan.

3.7 RUMUSAN

Dalam bahagian ini, data mentah yang diperolehi daripada Sistem eSILA dan Kajian Pengesanan Graduan Tajaan JPA dikumpulkan. Data ini kemudiannya melalui pra-pemprosesan untuk membersihkan data dan mengenalpasti pasti atribut yang bersesuaian. Jumlah data yang diperolehi adalah sebanyak 48,952. Data ini

kemudiannya akan digunakan untuk membangunkan model ramalan klasifikasi. Sebanyak 5 model menggunakan algoritma pembelajaran mesin yang berbeza akan dibangunkan iaitu Regresi Logistik (LR), Naïve Bayes (NB), Perhutanan Rawak (RF), Pohon Keputusan (DT) dan Gradient Boosting (GBM) . Prestasi model-model ini akan dinilai menggunakan matriks kekeliruan dan matriks prestasi iaitu Ketepatan, Kejituan, Sensitiviti, Skor F1 dan AUC-ROC. Prestasi model akan dibandingkan untuk mengenalpasti model yang terbaik untuk meramal kebolehpasaran pelajar tajaan kerajaan. Model terbaik seterusnya akan digunakan untuk membuat ramalan kebolehpasaran pelajar terhadap set data baru iaitu pelajar yang akan bergraduat pada tahun 2024.

BAB IV

ANALISA KAJIAN

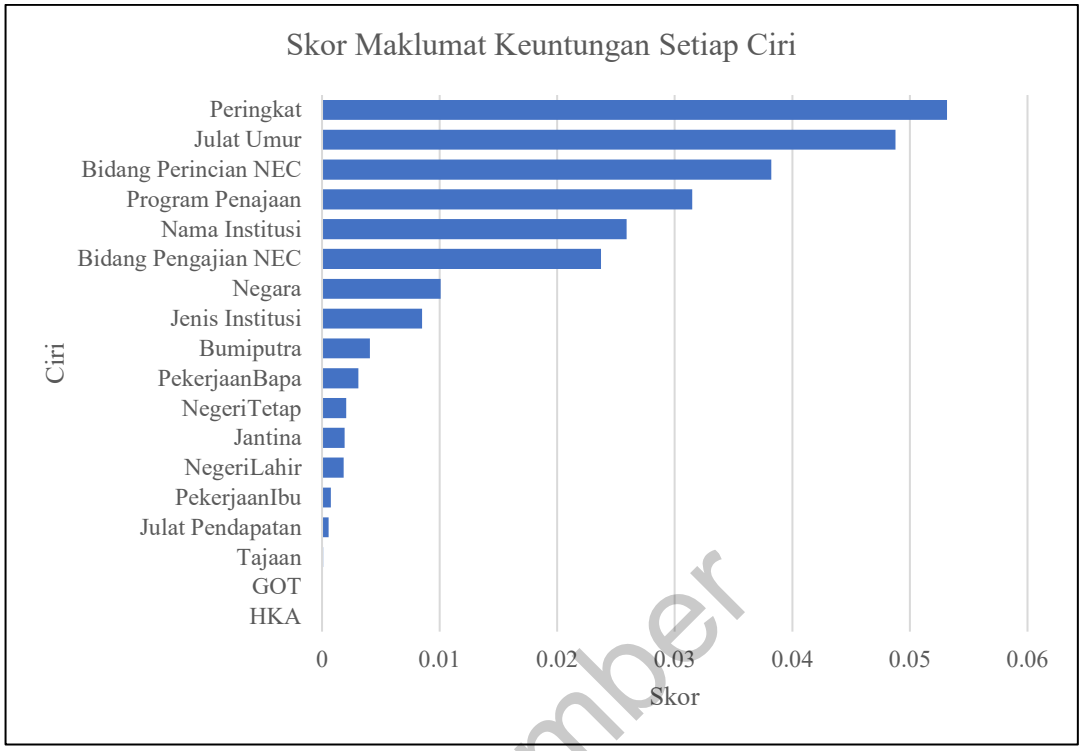
4.1 PENGENALAN

Bab ini akan menerangkan analisa kajian yang dilaksanakan terhadap data yang telah diperolehi berdasarkan objektif dan metodologi yang telah ditetapkan. Analisis ini merangkumi analisis pemilihan ciri, penilaian prestasi model ramalan, perbandingan model, faktor penentu kebolehpasaran pelajar dan ramalan kebolehpasaran.

4.2 ANALISIS PEMILIHAN CIRI

Analisis yang dijalankan menggunakan Maklumat Keuntungan (*Information Gain*) telah mengenal pasti ciri-ciri utama yang mempengaruhi kelas sasaran iaitu kebolehpasaran pelajar. Maklumat Keuntungan adalah ukuran yang digunakan dalam teori maklumat untuk mengukur pengaruh ciri terhadap atribut sasaran. Semakin tinggi skor Maklumat Keuntungan, semakin besar pengaruh ciri tersebut terhadap sasaran. Skor Maklumat Keuntungan setiap ciri ditunjukkan dalam Rajah 4.1.

Berdasarkan Rajah 4.1, ciri-ciri yang paling signifikan adalah Peringkat dengan skor tertinggi 0.053, diikuti oleh Julat Umur dengan skor 0.049. Ini menunjukkan bahawa peringkat pendidikan dan umur graduan memainkan peranan penting dalam menentukan kebolehpasaran mereka. Berdasarkan data diperolehi, kadar kebolehpasaran graduan di peringkat ijazah adalah 72% berbanding kadar kebolehpasaran graduan diploma iaitu 19%. Ini juga menjelaskan mengapa julat umur yang lebih tinggi mempunyai kebolehpasaran yang lebih tinggi.



Rajah 4.1 Skor Maklumat Keuntungan Setiap Ciri

Ciri Bidang Perincian NEC, Program Penajaan dan Bidang Pengajian NEC juga menunjukkan kepentingan yang ketara, masing-masing dengan skor 0.038, 0.031 dan 0.024. Bidang pengajian dan jenis program penajaan mempengaruhi status pekerjaan secara signifikan, di mana bidang-bidang tertentu seperti kesihatan dan teknologi maklumat menawarkan peluang pekerjaan yang lebih tinggi. Kedua-dua ciri ini adalah sangat berkaitan di mana Program LSPM ialah program penajaan kepada pelajar cemerlang lepasan SPM dalam bidang-bidang perkhususan tertentu seperti perubatan dan kejuruteraan. Ia mencerminkan kepentingan program penajaan khusus dalam menentukan kejayaan graduan di pasaran kerja.

Nama Institusi (0.026), Negara (0.010) dan Jenis Institusi (0.009) menonjolkan peranan institusi pendidikan dan negara pengajian dalam mempengaruhi kebolehpasaran. Institusi yang lebih berprestij memberikan kelebihan kepada graduan dalam mendapatkan pekerjaan.

Faktor-faktor seperti Status Bumiputra, Pekerjaan Bapa, Negeri Tetap, Jantina, Negeri Lahir, Pekerjaan Ibu dan Julat Pendapatan mempunyai skor yang lebih rendah, menunjukkan pengaruh yang lebih kecil terhadap status pekerjaan. Ini mencadangkan