PENDEKATAN PEMBELAJARAN MESIN BAGI MERAMAL KEBOLEHPASARAN GRADUAN TVET

MOHD HASHIM BIN ASHAARI

UNIVERSITI KEBANGSAAN MALAYSIA

PENDEKATAN PEMBELAJARAN MESIN BAGI MERAMAL KEBOLEHPASARAN GRADUAN TVET

MOHD HASHIM BIN ASHAARI

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

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2021

PENGAKUAN

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

18 Oktober 2021

MOHD HASHIM BIN ASHAARI
P101448

PENGHARGAAN

Dengan nama Allah yang Maha Pemurah lagi Maha Penyayang , Selawat dan salam buat junjungan besar Nabi Muhammad S.A.W serta para sahabat dan kaum keluarga baginda. Segala puji bagi Allah S.W.T kerana dengan berkat dan limpah kurnianya dapatlah saya menyiapkan tesis sarjana ini dengan jayanya.

Setinggi-tinggi penghargaan dan terima kasih saya ucapkan kepada penyelia sarjana saya, Prof. Madya Dr. Shahnorbanun binti Sahran atas segala nasihat dan tunjuk ajar serta sokongan yang tidak berbelah bagi sepanjang proses penerbitan tesis ini. Semoga Allah merahmati dan memberkati segala usaha dan tunjuk ajar yang telah beliau berikan kepada saya. Tidak lupa juga kepada penyelaras program Sarjana Sains Data, Dr. Mohd Ridzwan bin Yaakub dan semua tenaga pengajar yang terlibat membimbing dan mencurahkan ilmu dalam memastikan saya berjaya menamatkan sarjana saya ini.

Penghargaan juga diucapkan kepada Jabatan Pendidikan Politeknik dan Kolej Komuniti yang sentiasa memberi sokongan dan bantuan yang diperlukan bagi menjayakan penerbitan tesis ini. Penghormatan juga kepada kedua ibu bapa, isteri dan seluruh ahli keluarga saya yang sentiasa mendoakan dan memberi sokongan moral yang tidak terhingga untuk saya menyiapkan tesis ini.

Tidak dilupakan ungkapan jutaan terima kasih kepada rakan-rakan seperjuangan sekalian yang telah bersama-sama bertungkus lumus dan saling bantu membantu secara langsung dan tidak langsung sepanjang tempoh pengajian ini. Jasa dan bantuan yang kalian berikan hanya Allah sahaja yang boleh membalasnya.

Akhir kata, hasil penyelidikan yang telah saya laksanakan ini dapat memberi manfaat dan mencetus idea baru kepada generasi akan datang dalam melaksanakan penyelidikan dalam bidang sains data.

ABSTRAK

Kebolehpasaran graduan TVET memainkan peranan yang sangat penting dalam memastikan keberkesanan program TVET di Malaysia. Isu kebolehpasaran graduan TVET ini telah menjadi perdebatan dan perhatian yang mendalam oleh ahli akademik, pengamal TVET, industri dan pentadbir organisasi TVET. Menurut laporan kajian Pengesanan Graduan TVET menunjukkan bahawa kadar kebolehpasaran graduan TVET adalah 88.9% dan 93.4% pada tahun 2018 dan 2019. Walaupun kadar kebolehpasaran graduan TVET adalah tinggi namun laporan ILMIA menyatakan ramai graduan TVET bekerja di luar bidang. Kajian lepas telah membangunkan model kebolehpasaran graduan menggunakan pendekatan tunggal tetapi mempunyai keputusan ketepatan yang rendah (59.01%), dan kajian lepas juga mendapati bahawa kaedah model gabungan menghasilkan model ketepatan yang lebih baik. Oleh itu tujuan kajian ini adalah untuk menghasilkan model kebolehpasaran yang tinggi dengan dua kaedah. Pertama mengenalpasti atribut penting menggunakan chi square dan Naive Bayes. Kedua ialah membangunkan model kebolehpasaran menggunakan kaedah pengabungan antara algorithma RF, DT, KNN, LR dan NB berdasarkan 54 atribut dari rekod data tahun 2018 dan 2019. Experimen pembangunan model menggunakan kaedah 10-kali lipatan. 10 model gabungan dibangunkan antara Hutan Rawak (RF) dan *K-Nearest Neighbors* (KNN) dengan algoritma Naive Bayes (NB), Pokok Keputusan (DT) dan Regresi Logistik (LR). Hasil kajian mendapati 10 atribut penting mempengaruhi kebolehpasaran graduan iaitu syarikat berkerja sama LI, program pengajian bantu pekerjaan, jantina, sektor ekonomi, sub sektor ekonomi, taraf pekerjaan, pendapatan bulanan, cadangan belajar di institusi, status latihan industri dan kod kursus. Model penggabungan antara algoritma Hutan Rawak (RF) dan *K-Nearest Neighbors* (KNN) dengan algoritma Naive Bayes (NB), Pokok Keputusan (DT) dan Regresi Logistik (LR) menghasilkan prestasi model yang hampir sama iaitu 77.30% bagi model gabungan RF+KNN+DT, 77.11% bagi model gabungan RF+KNN+LR dan 77.07% bagi model gabungan RF+KNN+NB. Namun, hasil ujian Cochran's Q juga mendapati tiada perbezaan yang signifikan diantara ketiga-tiga model gabungan yang dicadangkan. Hasil keputusan menunjukkan model gabungan terbaik adalah RF+KNN+DT dengan ketepatan 77.3% berbanding kaedah model tunggal peroleh ketepatan terbaik 76.96%.

ABSTRACT

TVET graduates' employability plays a very important role in ensuring the effectiveness of TVET programs in Malaysia. The issue of TVET graduates' employability has been the subject of intense debate and attention by academics, TVET practitioners, industry and administrators of TVET organizations. According to the TVET Graduates Tracer Study report report shows that the employbility rate of TVET graduates is 88.9% and 93.4% in 2018 and 2019. Although the employbility rate of TVET graduates is high, ILMIA report states that many TVET graduates work outside the field. Past studies have developed graduate employbility models using a single approach but have low accuracy results (59.01%), and also found that the combined model method produces better accuracy models. Therefore the purpose of this study is to produce a high employbility model with two methods. First identify important attributes using chi square and Naive Bayes. The second is to develop a employbility model using the ensemble method between RF, DT, KNN, LR and NB algorithms based on 54 attributes from the 2018 and 2019 data records. Model development experiments using the 10-fold method. 10 combined models were developed between Random Forest (RF) and K-Nearest Neighbors (KNN) with Naive Bayes (NB), Decision Tree (DT) and Logistic Regression (LR) algorithms. The results of the study have identified 10 main attributes that influence the employability of graduates which is companies working with LI, employment assistance study program, gender, economic sector, economic sub -sector, employment status, monthly income, study proposal at institution, LI status and course code. The ensemble model between the Random Forest (RF) and K- Nearest Neighbors (KNN) algorithms with the Naive Bayes (NB), Decision Tree (DT) and Logistic Regression (LR) algorithms produced almost the same model performance of 77.30% for the RF+KNN combined model +DT, 77.11% for the RF+KNN+LR combined model and 77.07% for the RF+KNN+NB combined model in term of accuracy. However, the results of Cochran's Q test found there is no significant differences between the three proposed combined models. The results show that the best combined model is RF+KNN+DT with an accuracy of 77.3% compared to the single model method which obtained the best accuracy of 76.96%.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI RAJAH

SENARAI SINGKATAN

ANN	Rangkaian Neaural Buatan
CCMS	Sistem Pengurusan Maklumat Kolej Komuniti
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
COPTPA	*Code of Practice for TVET Programme Accreditation*
DT	Pokok Keputusan
FN	Salah Negatif
FP	Salah Positif
H_0	Hipotesis Null
H_1	Hipotesis Alternatif
IKBN	Institut Kemahiran Belia Negara
ILMIA	Institut Maklumat dan Analisis Pasaran Buruh
ILP	Institut Latihan Perindustrian
JPK	Jabatan Pembangunan Kemahiran
JPPKK	Jabatan Pendidikan Politeknik dan Kolej Komuniti
KBS	Kementerian Belia dan Sukan
KKR	Kementerian Kerja Raya
KNN	K-Nearest Neighbours
KPLB	Kementerian Pembangunan Luar Bandar
KPM	Kementerian Pendidikan Malaysia
KPT	Kementerian Pengajian Tinggi
KSM	Kementerian Sumber Manusia
LI	Latihan Industi
LR	Regresi Logistik
MAFI	Kementerian Pertanian dan Industri Makanan

MAMPU	Unit Pemodenan Tadbiran dan Perancangan Pengurusan Malaysia
MEA	Kementerian Hal Ehwal Ekonomi
MCA	Master of Computer Applications
MLP	*Multilayer Perceptron*
MQA	Agensi Kelayakan Malaysia
NB	Naïve Bayes
OEEU	Spanish Observatory for Employability and Employment
OKU	Orang Kurang Upaya
PKB	Politeknik Kota Bharu
POLIMAS	Politeknik Sultan Abdul Halim Muadzam Shah
POLISAS	Politeknik Sultan Haji Ahmad Shah
PSP	Politeknik Seberang Perai
PTS	Politeknik Tawau Sabah
PwC	*PriceWaterhouseCoopers*
RF	Hutan Rawak
RMKe-11	Rancangan Malaysia Kesebelas
ROI	Pulangan Keatas Pelaburan
SKPG	Sistem Kajian Pengesanan Graduan
SMO	*Sequential Minimal Optimization*
SPM	Sijil Pelajaran Malaysia
SPMP	Sistem Pengurusan Maklumat Politeknik
SVM	Mesin Sokongan Vektor
TP	Benar Positif
TN	Benar Negatif
TVET	*Technical and Vocational Education and Training*

UNESCO	*United Nations Organisation for Education, Science and Culture*
UKM	Universiti Kebangsaan Malaysia
WBL	Pembelajaran Berasaskan Tempat Kerja

BAB I

PENGENALAN

1.1 PENDAHULUAN

Pembangunan bakat modal insan merupakan faktor kritikal bagi menjana dan mengekalkan pertumbuhan ekonomi Malaysia. Rancangan Malaysia Kesebelas (RMKe-11) 2016 – 2020 meneruskan usaha untuk melahirkan bakat modal insan berpengetahuan, berkemahiran dan memiliki sikap positif untuk terus maju dalam ekonomi global. TVET telah dilihat berperanan bukan sahaja sebagai satu medium untuk memproses dan menghasilkan tenaga mahir tempatan, malahan juga sebagai nadi pembangunan negara. Perkembangan TVET negara adalah penting untuk meletakkan negara terus relevan dan tidak ketinggalan dalam bersaing dengan negara-negara maju yang lain. Dalam tempoh RMKe-11dijangka 1.5 juta pekerjaan diwujudkan menjelang tahun 2020 dan sebanyak 60 peratus pekerjaan baharu tersebut adalah dalam bidang *Technical and Vocational Education and Training* (TVET).

TVET merupakan proses pendidikan dan latihan yang menjurus ke arah pekerjaan dan memberi penekanan kepada amalan industri dalam pelbagai bidang berkaitan sains dan teknologi (MEA, 2019). TVET juga merupakan pembelajaran sepanjang hayat yang meliputi pembelajaran di peringkat menengah, pasca menengah dan pembelajaran berasaskan tempat kerja (WBL) (UNESCO(GC), 2015). *United Nations Organisation for Education, Science and Culture (UNESCO)* mendefinisikan TVET sebagai aspek proses pendidikan selain pendidikan umum yang melibatkan pembelajaran dalam bidang teknologi dan sains berkaitan, dan juga latihan kemahiran praktikal, sikap, pemahaman, dan pengetahuan tentang pekerjaan dalam pelbagai sektor ekonomi dan kehidupan sosial (KPM, 2015). Merujuk *Code of Practice for TVET Programme Accreditation* (COPTPA) yang telah dikeluarkan oleh Agensi Kelayakan Malaysia (MQA) bersama Jabatan Pembangunan Kemahiran (JPK), TVET diistilahkan

sebagai proses pendidikan dan latihan ke arah pekerjaan dengan memberi penekanan utama terhadap amalan industri dimana ianya bertujuan untuk melahirkan tenaga kerja yang kompeten di sesebuah negara (MQA & JPK, 2019).

Pada tahun 1964, dua institusi TVET awam telah ditubuhkan untuk menyediakan Pendidikan dan latihan kemahiran kepada belia, iaitu Institut Kemahiran Belia Negara (IKBN) Dusun Tua dan Institut Latihan Perindustrian (ILP) Kuala Lumpur. Sehingga kini, terdapat 1,248 institusi TVET yang diuruskan oleh Kerajaan dan swasta yang mana telah menawarkan pelbagai program TVET bagi semua peringkat pendidikan. Berdasarkan dokumen Rancangan Malaysia Kesebelas (RMKe-11), terdapat enam (6) kementerian utama yang terlibat dalam menguruskan institusi TVET iaitu Kementerian Sumber Manusia (KSM), Kementerian Pembangunan Luar Bandar (KPLB), Kementerian Kerja Raya (KKR),Kementerian Belia dan Sukan (KBS), Kementerian Pertanian dan Industri Makanan (MAFI), Kementerian Pendidikan Malaysia (KPM) dan Kementerian Pengajian Tinggi (KPT) dimana dokumen RMKe-11 telah mengelaskan KPM dan KPT sebagai sebuah kementerian iaitu di bawah KPM.

Justeru, kebolehpasaran graduan TVET memainkan peranan yang sangat penting dalam memastikan keberkesanan program TVET di Malaysia. Isu kebolehpasaran graduan TVET ini telah menjadi perdebatan dan perhatian yang mendalam oleh ahli akademik, pengamal-pengamal TVET, industri dan pentadbir organisasi TVET. Bagi mengenal pasti kebolehpasaran graduan TVET, semua kementerian TVET telah melaksanakan kajian kebolehpasaran graduan masing-masing semasa berlangsungnya majlis konvokesyen di institut masing-masing. Pada tahun 2018, KPM dan KPT telah membangunkan Sistem Kajian Pengesanan Graduan (SKPG) TVET secara atas talian dimana sistem ini turut digunakan oleh Kementerian penyedia TVET yang lain untuk mengkaji kebolehpasaran graduan TVET. Definisi kebolehpasaran graduan yang diguna pakai dalam SKPG TVET adalah graduan yang bekerja/bekerja sendiri, melanjutkan pelajaran, meningkatkan kemahiran dan menunggu penempatan pekerjaan pada tahun konvokesyen.

Berdasarkan Statistik Pengajian Tinggi 2019 yang telah dikeluarkan oleh Kementerian Pengajian Tinggi (KPT), purata kebolehpasaran graduan TVET adalah

89.72 peratus dimana institusi TVET di bawah KPT dan KPM iaitu politeknik, kolej komuniti dan kolej vokasional telah mencatatkan kebolehpasaran graduan yang tinggi iaitu melebihi 90 peratus dalam tempoh 6 bulan selepas bergraduat. Walaubagaimanapun, laporan Kajian Pembangunan Pelan Induk Kebangsaan Latihan Teknikal Dan Vokasional (TVET) Ke Arah Negara Maju yang telah dikeluarkan oleh Institut Maklumat dan Analisis Pasaran Buruh (ILMIA) pada 12 Oktober 2018 dimana kajian tersebut dijalankan oleh *PriceWaterhouseCoopers (PwC)* telah melaporkan bahawa 31 peratus graduan TVET bekerja, melanjutkan pengajian atau menjadi usahawan bukan dalam yang mereka pelajari semasa berada di institusi. Tambahan pula, dapatan kajian yang telah dijalankan telah melaporkan bahawa 72 peratus graduan TVET yang bekerja menerima bayaran atau gaji yang setimpal dengan kelayakan yang dimiliki iaitu di bawah RM 1,500 sebulan (ILMIA, 2018). Ini telah menyebabkan graduan TVET dilihat sebagai graduan kelas kedua di Malaysia.

Oleh itu, amat penting bagi sesebuah institusi TVET memastikan graduan yang mereka hasilkan dapat dipasarkan di industri dengan menerima bayaran yang setimpal dengan kelayakan dan bekerja sesuai dengan bidang yang dipelajari. Pelbagai kaedah dan inisiatif telah dilaksanakan oleh institusi TVET khususnya institusi TVET awam bagi memastikan graduan mereka dapat dipasarkan dalam tempoh 6 bulan selepas bergraduat. Walaubagaimanapun, masih belum ada inisiatif yang boleh diambil bagi mengenal pasti dan meramal kebolehpasaran graduan sebelum pelajar-pelajar tersebut bergraduat. Melalui Pelan Strategik ICT Sektor Awam 2016 – 2020 yang telah dibangunkan oleh Unit Pemodenan Tadbiran dan Perancangan Pengurusan Malaysia (MAMPU), kerajaan melalui Teras Keduanya iaitu Kerajaan Berpacukan Data telah menggariskan supaya agensi kerajaan mengurus dan membuat keputusan berpacukan data agar hasil yang optimum dapat dicapai. Sejajar dengan itu, kajian ini akan berfokuskan kepada penghasilan satu model bagi meramal kebolehpasaran graduan TVET menggunakan aplikasi perlombongan data dan teknik pembelajaran mesin bagi mengenal pasti kadar kebolehpasaran graduan TVET sebelum pelajar tersebut bergraduat.

1.2 PERMASALAHAN KAJIAN

Dalam Rancangan Malaysia ke-11 (RMKe-11) salah satu agenda ekonomi yang digariskan adalah untuk mewujudkan 1.5 juta pekerjaan menjelang tahun 2020 dalam usaha kerajaan untuk mengurangkan kebergantungan terhadap pekerja asing dalam sektor berkaitan TVET. Menerusi RMKe-11 kerajaan meramalkan bahawa 60 peratus daripada pekerjaan yang dijangka wujud memerlukan kelayakan berkaitan TVET. TVET juga telah dilihat sebagai *game changer* dalam menghasilkan tenaga kerja yang berkemahiran di Malaysia. Bagi mencapai hasrat tersebut , program TVET perlu ditransformasi bagi memenuhi kehendak industri.

Menurut laporan Kajian Pengesanan Graduan TVET yang dikeluarkan oleh KPT bermula tahun 2018 yang telah dilaksanakan secara atas talian telah menunjukkan bahawa kadar kebolehpasaran graduan TVET adalah 88.9 peratus pada tahun 2018 dan 93.4% pada tahun 2019 dimana politeknik dan kolej komuniti merupakan institusi TVET yang diselia dibawah Jabatan Pendidikan Politeknik dan Kolej Komuniti (JPPKK) memperolehi kadar kebolehpasaran yang tertinggi berbanding institusi TVET yang lain.

Pada tahun 2018, sebanyak 73.5 peratus graduan memperolehi pekerjaan dalam tempoh enam bulan selepas bergraduat, 12.9 peratus melanjutkan pengajian, 0.4 peratus meningkatkan kemahiran, 2.1 peratus menunggu penempatan pekerjaan dan 11.1 peratus belum mendapat pekerjaan. Pada tahun 2019 pula, 73.3 peratus graduan memperolehi pekerjaan dalam tempoh enam bulan selepas bergraduat, 15.9 peratus melanjutkan pengajian, 0.6 peratus meningkatkan kemahiran, 3.5 peratus menunggu penempatan pekerjaan dan 6.6 peratus belum mendapat pekerjaan.

Walaupun kadar kebolehpasaran graduan TVET adalah tinggi masih wujud permasalahan dimana terdapat graduan TVET bekerja di luar bidang yang diceburi selari dengan laporan yang telah dikeluarkan oleh ILMIA. Situasi ini menyebabkan keperluan tenaga kerja dalam sektor atau bidang tertentu tidak dapat diisi dan hasilnya pelaburan Kerajaan sejumlah RM10,000 setahun dalam menanggung anggaran kos per kepala pelajar TVET dilihat tidak memberi pulangan seperti yang diharapkan.

Melihat kepada permasalahan tersebut , penyedia-penyedia TVET perlu menjalankan kajian dan analisis untuk mencari jalan penyelesaian terbaik bagi mengatasi dan merancang hala tuju institusi pada masa akan datang. Kajian-kajian berkaitan kebolehpasaran yang dilaksanakan oleh penyelidik-penyelidik terdahulu tidak menyentuh isu berkenaan status pekerjaan graduan TVET sama ada di dalam atau luar bidang yang diceburi. Justeru, bagi meramal kebolehpasaran graduan TVET di Malaysia, kajian ini mencadangkan untuk membina satu model bagi meramalkan kebolehpasaran graduan TVET samada bekerja dalam bidang atau di luar bidang dengan menggunakan data graduan politeknik dan kolej komuniti.

Beberapa kajian telah dijalankan oleh penyelidik seluruh dunia berkaitan peramalan kadar kebolehpasaran graduan menggunakan kaedah perlombongan data dan pembelajaran mesin. Namun masih belum terdapat kajian yang menggunakan pendekatan model penggabungan bagi meramal kebolehpasaran graduan. Di Malaysia khususnya, pada tahun 2019 Tan et al. telah menggunakan model tunggal untuk membangunkan model bagi mengenal pasti faktor penting kebolehpasaran graduan IPT di Malaysia dengan keputusan ketepatan model 59.01 peratus.

1.3 PERSOALAN KAJIAN

Berdasarkan latar belakang kajian dan pernyataan masalah terdapat keperluan untuk menyelesaikan persoalan berikut :

a) Apakah teknik atau model pengelasan berasakan pembelajaran mesin yang terbaik untuk meramal kebolehpasaran graduan TVET samada bekerja dalam bidang atau luar bidang pengajian?

b) Apakah atribut yang mempengaruhi graduan TVET bekerja dalam dan luar bidang dan hubungan antara atribut-atribut tersebut?

Bagi menjawab persoalan di atas, kajian ini adalah dilaksanakan untuk membina dan menguji model bagi menilai data set graduan politeknik dan kolej komuniti yang diperolehi daripada Sistem Kajian Pengesanan Graduan-TVET.

1.4 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah:

a) Mengenal pasti atribut-atribut yang mempengaruhi status pekerjaan samada di dalam atau luar bidang pengajian graduan politeknik dan kolej komuniti bagi membantu pihak JPPKK dalam menangani isu pelajar bekerja di luar bidang dengan menggunakan kaedah statistik chi square.

b) Mengenal pasti model atau algoritma pembelajaran mesin yang terbaik bagi membuat ramalan kadar kebolehpasaran graduan politeknik dan kolej komuniti sama ada bekerja dalam bidang atau di luar bidang pengajian.

1.5 SKOP KAJIAN

Dalam melaksanakan kajian ini, beberapa batasan kajian telah dikenal pasti iaitu ;

i. Kajian ini terbatas kepada data pelajar politeknik seluruh Malaysia yang diperolehi daripada Sistem Kajian Pengesanan Graduan TVET (SKPG-TVET).

ii. Jumlah kohort data yang digunakan dalam kajian ini melibatkan pelajar politeknik yang bergraduat pada tahun 2018 dan 2019 dan tidak mengambil kira data pelajar institusi TVET yang lain.

1.6 KEPENTINGAN KAJIAN

Isu kebolehpasaran graduan memainkan peranan yang signifikan dalam usaha kerajaan untuk membangunkan modal insan yang diharap akan membantu pertumbuhan ekonomi serta memberi pulangan ke atas pelaburan (ROI) kepada negara melalui program TVET. Dalam RMKe-11 kerajaan telah memperuntukkan sejumlah peruntukan yang besar untuk meningkat program TVET bagi menyediakan tenaga kerja berkemahiran tinggi. Justeru, kebolehpasaran graduan TVET amat dititik beratkan oleh

kerajaan dalam memastikan graduan TVET dalam menyumbang kepada pertumbuhan ekonomi negara.

Oleh itu, melalui kajian ini diharapkan dapat membantu institusi TVET khususnya politeknik dan kolej komuniti mengenal pasti atribut-atribut yang boleh mempengaruhi kebolehpasaran graduan TVET samada bekerja dalam bidang atau luar bidang pengajian dan seterusnya dapat membantu pihak pengurusan di JPPKK untuk mengenal pasti ciri-ciri graduan yang perlu diberi pertolongan serta membantu merangka dan merancang pelan-pelan tindakan yang boleh membantu graduan tersebut meningkatkan peluang mendapatkan pekerjaan dalam bidang yang diceburi.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Perlombongan data merupakan proses penting dalam mencari dan mengenal pasti corak dan pengetahuan dari data yang besar. Perlombongan data telah banyak diaplikasikan dan terkenal di kalangan syarikat-syarikat besar seperti Facebook, Microsoft dan Google. Perlombongan 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 *(outliers)*, analisis trend, regresi dan sebagainya (Rahman et al., 2017; Sani et al., 2018).

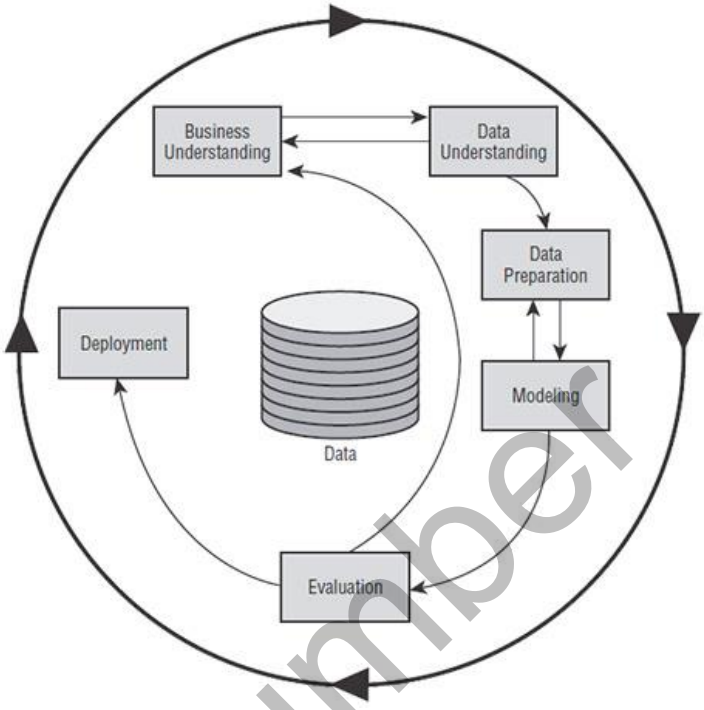
Pembelajaran mesin adalah hasil daripada kemajuan program kecerdasan buatan yang membolehkan sesebuah program belajar tanpa campur tangan dan bantuan manusia atau arahan yang jelas menggunakan algoritma khusus untuk menganalisis, mempelajari dan mengenali corak set data tertentu. Dengan mempelajari data yang telah diberikan kepada sesebuah program, pembelajaran mereka dari masa ke masa akan ditingkatkan. Sasaran ramalan adalah untuk mengelaskan hasil dengan tepat berdasarkan hubungan yang masuk akal dan boleh diterima dari data yang telah disediakan (Osisanwo et al., 2017).

Pembelajaran yang diselia dan tidak diselia adalah dua kaedah algoritma pembelajaran yang digunakan dalam pembelajaran mesin. Dalam pembelajaran yang diselia, mesin atau program tersebut dilatih menggunakan data yang mempunyai label atau data yang diketahui. Mesin akan belajar dari data latihan dan menghasilkan corak daripada data tersebut. Kemudian data baru atau data yang tidak pernah dilihat

dibekalkan kepada model yang telah dibina untuk meramalkan hasil bagi data baru tersebut. Masalah pengelasan biasanya menggunakan kaedah pembelajaran yang diselia untuk menyelesaikan masalah tersebut. Berlainan dengan pembelajaran tidak diselia, mesin atau program akan berfungsi dengan sendirinya untuk mencari maklumat penting daripada data tersebut. Pembelajaran tidak diselia biasanya digunakan bagi data yang tidak berlabel atau data yang tidak diketahui (Jaffar et al, 2019).

Bagi tujuan mendapatkan model tersebut, model proses CRISP-DM digunakan. Terdapat enam fasa model proses CRISP-DM iaitu pemahaman perniagaan, pemahaman data, penyediaan data, pemodelan, penilaian dan penyebaran data seperti yang ditunjukkan pada Rajah 2.1. Fasa pemahaman perniagaan berfokuskan kepada proses mengenal pasti objektif dan keperluan sesuatu projek dan kemudiannya mengubahnya menjadi penyataan masalah kepada proses perlombongan data. Fasa pemahaman data dimulai dengan proses pengumpulan data awal dan memahami data untuk mengenal pasti masalah kualiti data, penemuan awal berkenaan data atau mengesan subset yang menarik berkaitan data untuk membentuk hipotesis bagi maklumat yang tersembunyi.

Fasa penyediaan data merangkumi semua aktiviti-aktiviti untuk membina data set akhir daripada data mentah. Fasa pemodelan merupakan fasa dimana teknik pemodelan bagi data tersebut dipilih dan diaplikasikan. Pemilihan model adalah berdasarkan keperluan khusus berdasarkan data set dan tujuan analisis dijalankan. Setelah model-model dibangunkan, model-model tersebut seterusnya akan diuji untuk memastikan model-model tersebut boleh *generalize* data yang belum dilihat bagi memilih model yang terbaik bagi tujuan penyebaran (Rahman et al., 2017).



Rajah 2.1 Model CRISP-DM (Sumber: Rahman et al, 2017)

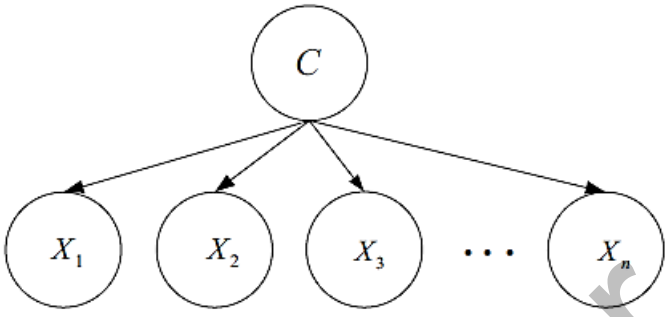
2.2 PEMODELAN PEMBELAJARAN MESIN

Kajian ini dijalankan dengan menggunakan tujuh teknik pembelajaran mesin iaitu Naïve Bayes (NB), Mesin Sokongan Vektor (SVM), *K-Nearest Neighbours* (KNN), Rangkaian Neural Buatan (ANN), Regresi Logistik (LR), Hutan Rawak (RF) dan Pokok Keputusan (DT) bagi mencapai objektif kajian. Penjelasan mengenai model-model pengelasan yang akan digunakan dalam kajian ini adalah seperti berikut:

2.2.1 Naïve Bayes (NB)

Naive Bayes adalah algoritma kebarangkalian yang mengaplikasikan teorem Bayes yang terdiri daripada grafik *acyclic* terarah dengan andaian semua atribut adalah tidak saling bersandar antara satu sama lain (Osisanwo et al., 2017). Algoritma dinamakan sebagai model kebarangkalian berdasarkan anggaran dan keadaan *naïve* (Jaffar et al., 2019). Algoritma ini sering digunakan dalam kaedah pembelajaran diselia dengan menganggap atribut kelas tidak mempunyai kaitan dengan atribut yang lain. Naive Bayes model boleh dilatih dengan ketepatan yang tinggi bergantung kepada

model kebarangkalian yang betul (Berend et al., 2015). Rajah 2.2 menunjukkan Model Naïve Bayes.

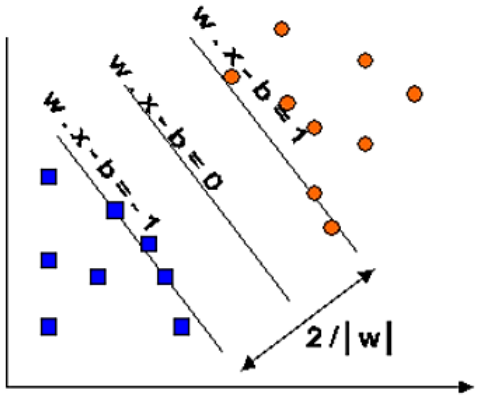


Rajah 2.2 Model Naïve Bayes (Sumber: Taheri et al., 2013)

Teorem Bayes: $\mathrm{P}(c|x) = \frac{P(x|c)P(c)}{P(x)}$

2.2.2 Mesin Sokongan Vektor (SVM)

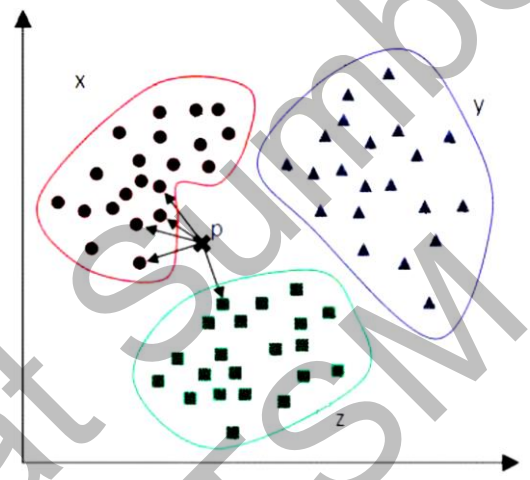
Mesin Sokongan Vektor (SVM) merupakan teknik pembelajaran mesin diselia yang terkini dimana kebanyakannya digunakan dalam analisis masalah berkaitan pengelasan dan regresi (Osisanwo et al, 2017). Model ini hampir sama dengan model rangkaian neural khususnya *Multilayer Perceptron* (MLP). Model SVM membuat satu atau lebih *hyperplan* yang bertujuan untuk memisahkan kelas data dalam analisis pengelasan. Bagi mengurangkan ralat dalam analisis pengelasan menggunakan model SVM, jarak dan margin antara *hyperplan* dan *instances* di kedua-dua sisinya mesti dalam keadaan maksimum (Durgesh et al, 2010). Konsep Model SVM dapat ditunjukkan dalam Rajah 2.3.



Rajah 2.3 Model Mesin Sokongan Vektor (SVM) (Sumber: Durgesh et al., 2010)

2.2.3 K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) merupakan algoritma pembelajaran yang diselia bagi tujuan menjalankan analisis pengelasan. Algoritma pengelasan ini berfungsi dengan meramalkan jiran terdekat bagi sampel ujian berdasarkan sampel latihan K dengan menentukan kategori kebarangkalian terbesar bagi sampel tersebut (Suguna et al., 2010). KNN telah membuktikan prestasi yang sangat baik dalam analisis pengelasan bagi kumpulan data yang berbeza. Konsep model KNN dapat ditunjukkan dalam Rajah 2.4.

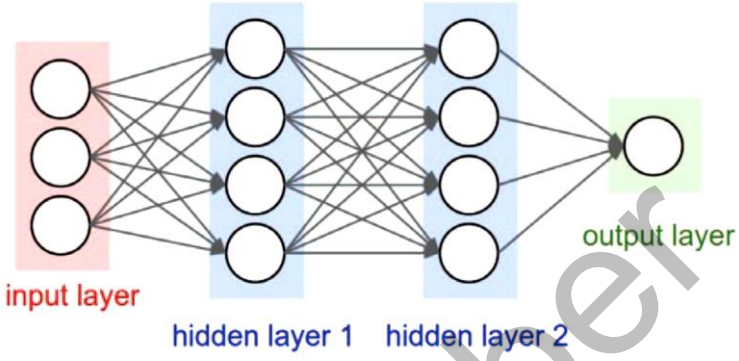


Rajah 2.4 Model *K-Nearest Neighbours* (KNN) (Sumber:Suguna et al., 2010)

2.2.4 Rangkaian Neural Buatan (ANN)

Rangkaian Neural Buatan (ANN) merupakan teknik pembelajaran mesin yang bermodelkan otak manusia dimana ianya terdiri daripada sejumlah neuron buatan. Neuron dalam ANN cenderung mempunyai rangkaian yang lebih banyak berbanding neuron biologi. Model ANN biasanya merangkumi tiga jenis lapisan, iaitu lapisan input, lapisan tersembunyi dan lapisan keluaran. Prosedur pembelajaran dalam ANN merangkumi dua fasa, iaitu penyebaran kata kunci bagi tujuan pemindahan maklumat dan penyebaran balik bagi tujuan mengubah suai ralat. Dalam penyebaran kata kunci, maklumat input dipindahkan dari lapisan input ke lapisan ke keluaran. Semasa proses ini, pembelajaran beralih kepada prosedur penyebaran belakang apabila hasil keluaran yang diinginkan tidak dapat diperolehi. Oleh itu, semua pemberat rangkaian neuron diubah bagi mengurangkan ralat. Proses ini diulang sehingga ralat kurang atau sama

daripada nilai ambang yang telah ditentukan pada awalnya (Liu et al., 2019). Model rangkaian ANN ditunjukkan dalam Rajah 2.5.



Rajah 2.5 Model Rangkaian Neural Buatan (ANN) (Sumber: Liu et al., 2019).

2.2.5 Regresi Logistik (LR)

Regresi logistik adalah satu kaedah klasik dalam pembelajaran mesin yang mengaplikasikan pembelajaran mesin yang diselia. Regresi logistik menggunakan Fungsi Sigmoid untuk meningkatkan generalisasi algoritma. Regresi logistik dikenali sebagai algoritma regresi dan algoritma pengelasan. Biasanya, ia digunakan sebagai algoritma pengelasan dan hanya dapat menyelesaikan masalah pengelasan binari dengan menggantikan regresi linear multivariat kepada fungsi mampatan seperti fungsi sigmoid, perataan dan non-linear dengan menggunakan nilai kebarangkalian untuk mengelaskan data diskrit yang tidak linear [22]. Bentuk fungsi sigmoid adalah seperti dibawah.

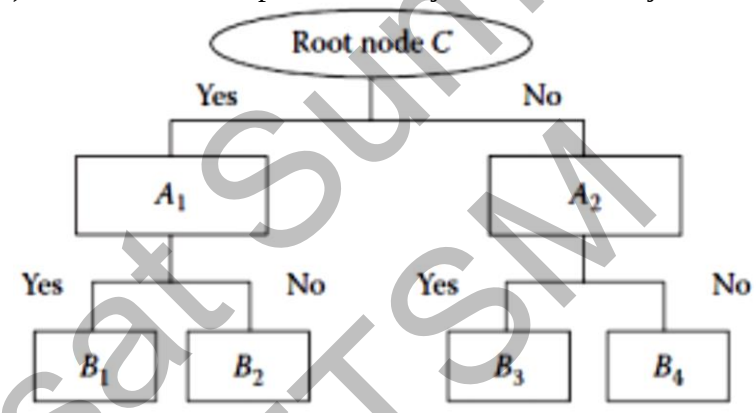
$$\text{Fungsi Sigmoid}, \sigma\,(t) = \frac{1}{1 + e^{-t}}$$

dimana $\sigma\,(t)$ mempunyai julat antara [0,1].

2.2.6 Pokok Keputusan (DT)

Pokok keputusan merupakan kaedah pengelasan yang terkenal disebabkan kecepatan dan mudah diproses dan mudah difahami. Supianto et al. mentakrifkan pokok keputusan sebagai kaedah pengelasan yang menggunakan mewakili struktur pokok di mana nod-nod mewakili atribut tertentu dan daun mewakili kelas dengan nod

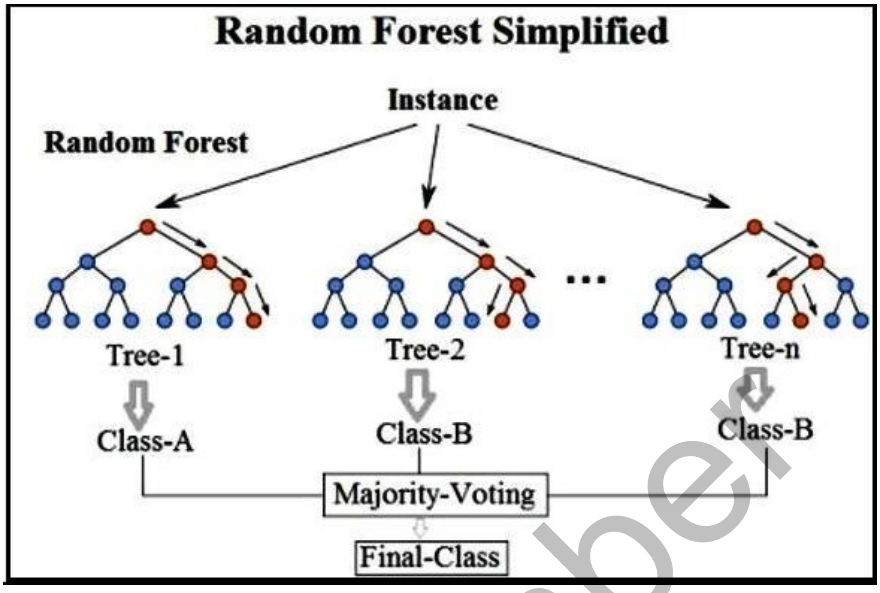
paling atas merupakan akar (Supianto et al., 2018). Nod akar merupakan nod teratas dari pokok keputusan yang tidak mempunyai input dan tidak mempunyai lebih daripada satu keluaran. Nod dalaman pula merupakan sebarang cabang nod dari nod akar dan mempunyai satu input dan sekurang-kurangnya dua output. Nod daun adalah sebarang nod yang mempunyai input tunggal dan tidak ada mempunyai keluaran ataupun dikelani sebagai nod terminal. Dalam pohon keputusan, kelas yang diramalkan diwakili dalam bentuk simpul daun yang dilukis dalam bulatan, hasil ujian diwakili dengan cabang yang dilukis dengan anak panah dengan label dan arah, sementara ujian diwakili dalam bentuk kotak. Pokok Keputusan boleh digunakan bagi menyelesaikan masalah seperti prestasi pelajar (Bhaskaran et al., 2015), pengelasan keciciran pelajar universiti (Abu-Oda et al., 2015), dan meramalkan kadar tamat pengajian pelajar di sesebuah institusi (Jaman, 2016). Model Pokok Keputusan ditunjukkan dalam Rajah 2.6.



Rajah 2.6 Contoh Pokok Keputusan (Sumber: Bhaskaran et al., 2015)

2.2.7 Hutan Rawak (RF)

Hutan rawak merupakan teknik pembelajaran mesin yang terselia berasaskan gabungan pokok-pokok keputusan. Teknik pembelajaran mesin ini lazimnya digunakan untuk pengelasan dan regresi., Algoritma hutan rawak membina hutan dengan pokok keputusan yang banyak dan mengambil purata ramalan daripada semua pokok keputusan. Pengelasan Hutan Rawak menggunakan kaedah *bagging* dan *random subspace* dalam membina setiap pokok untuk membuat hutan pokok yang tidak mempunyai korelasi. Keputusan atau ramalan akhir dibuat berdasarkan majoriti undian dari setiap node pokok keputusan. Hutan Rawak menggunakan *bootstrap aggregating* atau *bagging* untuk mengurangkan risiko *overfitting* dan masa latihan yang diperlukan (Apao e tal., 2020). Model Hutan Rawak ditunjukkan dalam Rajah 2.7.

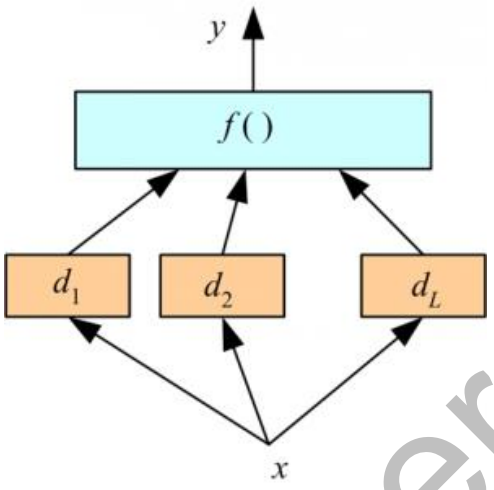


Rajah 2.7 Contoh Hutan Rawak (Sumber: Chari et al., 2019)

2.3 PEMBELAJARAN ENSEMBEL (PENGGABUNGAN)

Pembelajaran Ensembel atau penggabungan merupakan kaedah pengelasan dimana satu set kumpulan pengelasan yang hasil ramalannya digabungkan bagi membina satu kumpulan pengelasan untuk mengklasifikasikan kes baru. Kaedah Model Penggabungan ini telah digunakan dalam pelbagai bidang. Ketepatan yang diperolehi hasil model penggabungan biasanya lebih tinggi daripada kaedah pengelasan secara tunggal. Kaedah ini membina dan menggabungkan satu set hipotesis untuk memperbaiki kelemahan data latihan yang menggunakan pendekatan pembelajaran individu. Kaedah ini terdiri daripada beberapa pendekatan yang biasa digunakan dalam pengelasan untuk membina model seperti pendekatan *bagging*, *boosting* dan *voting* (Pierola et al. 2016). Dalam kajian ini pendekatan *voting* digunakan bagi tujuan pembelajaran penggabungan.

Voting merupakan algoritma penggabungan termudah, dan selalunya sangat berkesan. Ia dapat digunakan untuk masalah pengelasan atau regresi. *Voting* berfungsi dengan membina dua atau lebih sub-model. Setiap sub-model membuat ramalan yang akan digabungkan dalam beberapa cara seperti dengan mengambil purata atau mod ramalan dimana setiap sub-model dibenarkan untuk memilih apa hasilnya (Kabari et al, 2019). Contoh model penggabungan secara *voting* ditunjukkan dalam Rajah 2.8.



Rajah 2.8 Contoh Model Penggabungan *Voting* (Sumber: Kabari et al, 2019)

Rojarath et al dalam kajian mereka telah mengaplikasikan model penggabungan *Voting* ke atas lima (5) dataset yang diperolehi *UCI Machine Learning Repository* iaitu dataset cpu, b-scale, hepatitis, heart_de dan lymp. Enam (6) model yang dipilih untuk tujuan penggabungan pokok keputusan, mesin sokongan vektor, bayesian, Naive Bayes, MLP dan KNN (Rojarath et al., 2020). Jukic et al pula dalam kajian mereka telah mendapati model penggabungan menghasilkan keputusan ketepatan yang lebih baik berbanding model tunggal dengan keputusan ketepatan sebanyak 95.25% berbanding 92.55% untuk model tunggal bagi algoritma KNN (Jukic et al., 2020).

2.4 KAJIAN LEPAS BERKAITAN KEBOLEHPASARAN GRADUAN

Seiring dengan perkembangan dan peredaran dunia ke arah pendigitalan, proses pendidikan terutama TVET telah berubah ke arah pendidikan yang berorientasikan pekerjaan dimana reputasi dan pencapaian sesebuah institusi bergantung kepada kebolehpasaran graduan pelajar yang dihasilkan oleh institusi tersebut (Mishra et al, 2017). Masalah kebolehpasaran atau kebolehdapatan pekerjaan boleh memberi kesan yang negatif kepada individu, masyarakat dan negara terutama di kalangan graduan yang baru menamatkan pengajian di institusi pendidikan.

Penghasilan graduan yang semakin bertambah setiap tahun telah menyebabkan kebolehpasaran graduan institusi pendidikan tinggi terutama TVET telah menjadi sentiasa menjadi isu nasional di Malaysia. Secara umumnya, Jabatan Perangkaan

Malaysia (DOSM) telah melaporkan pada suku ketiga tahun 2020 kadar pengangguran di Malaysia berapa pada tahap 4.7 peratus. Menyedari masalah tersebut, model ramalan yang baik amat penting untuk mengenal pasti dan menentukan intervensi yang sesuai diberikan kepada graduan tertentu supaya ramalan yang tepat boleh dilaksanakan. Oleh itu, model ramalan yang boleh menunjukkan graduan atau pelajar mana yang akan dan tidak akan mendapat pekerjaan setelah menamatkan pengajian akan boleh membantu institusi mengenal pasti graduan yang memerlukan bantuan untuk mendapat pekerjaan.

Beberapa kajian telah dijalankan oleh penyelidik seluruh dunia berkaitan peramalan kadar kebolehpasaran graduan menggunakan kaedah perlombongan data dan pembelajaran mesin. Tan et al. menggunakan kaedah perlombongan data untuk mengenal pasti faktor utama yang menyumbang kepada kebolehpasaran graduan baru menamatkan pengajian dengan membuat perbandingan enam algoritma perlombongan data iaitu Regresi Logistik, Pokok Keputusan, Naïve Bayes, *K-Nearest Neighbor*, Mesin Sokongan Vektor dan Rangkaian Neural dengan mengaplikasikan 70-30 sebagai nisbah pecahan validasi. Hasil dapatan daripada kajian yang penyelidik tersebut laksanakan, algoritma Rangkaian Neural merupakan algoritma pengelasan yang paling baik diantara 5 algoritma yang lain dimana algoritma tersebut telah mengenal pasti enam atribut utama yang mempengaruhi kebolehpasaran iaitu kesediaan graduan menghadapi cabaran dunia luar dan alam pekerjaan, kebolehan berkomunikasi secara efektif, bidang teknikal, waktu konvokesyen dan jantina graduan tersebut (Tan et al., 2019).

Deepak et al. dalam kajian mereka telah menggunakan kaedah perlombongan data dengan menggunakan data dari sebuah kolej di India dimana data tersebut terdiri daripada 31 atribut daripada lapan faktor utama iaitu kualiti pengajaran, profil fakulti, penilaian pembelajaran, menjaga hubungan, kualiti organisasi dan hasil bimbingan dan kaunseling, tadbir urus, penyelidikan dan pembangunan. Penyelidik-penyelidik telah menggunakan algoritma Mesin Sokongan Vektor untuk menjalankan analisis pengelasan dan regresi (Deepak et al., 2016). Shafei et al. pula meramalkan trend pengangguran berdasarkan maklumat sesawang dengan menggunakan Novel Rangkaian Neural (NN) dimana pengkaji mengenal pasti enam atribut utama berkaitan kebolehpasaran iaitu sifat peribadi, bekerja dalam kumpulan, pengurusan kendiri,

kemahiran komunikasi, kebolehan mempelajari sendiri dan keusahawanan (Shafie et al., 2010). Manakala Aziz et al. pula menggunakan Naïve Bayes, Regresi Logistik, *Multilayer Perceptron, K-Nearest Neighbor* dan Pokok Keputusan untuk mengenal pasti model pengelasan yang paling sesuai bagi tujuan pengelasan kebolehpasaran graduan Kolej Profesional MARA (KPM) di Malaysia berdasarkan atribut jantina, program pengajian, bilangan semester, pencapaian ko-kurikulum dan pencapaian akademik pelajar tersebut (Aziz et al., 2016).

Selain itu, terdapat beberapa kajian luar negara yang berkaitan dengan kebolehpasaran graduan seperti Jantawan et al. Yang telah menjalankan kajian dan penyelidikan melalui penghasilan model kebolehpasaran graduan untuk meramalkan sama ada graduan akan bekerja atau menganggur melalui perbandingan antara algoritma Pokok Keputusan dan Bayesian menggunakan data *Maejo University* di Chiang Mai Thailand dan mendapati atribut utama yang menyumbang kepada graduan yang kekal tidak bekerja adalah atribut *ReasonNotWork* (Jantawan et al., 2013).

Peñalvo et al. pula telah menjalankan kajian kebolehpasaran graduan menggunakan pendekatan pembelajaran mesin untuk mengenal pasti bagaimana graduan boleh diterima bekerja. Penyelidik membangunkan model ramalan menggunakan algoritma pembelajaran mesin iaitu Hutan Rawak bagi tujuan mengenal pasti faktor-faktor yang releven berkaitan kebolehpasaran.

Hasil kajian mendapati antara atribut yang menyumbang kepada kebolehpasaran graduan adalah universiti dimana graduan menyambung pengajian, jantina graduan, peribadi graduan, program pengajian, bidang dan jawatan yang dimohon dan kesanggupan bekerja diluar tempat bermastautin (Peñalvoet al., 2018).

Misra et al. pula menggunakan enam teknik pembelajaran mesin iaitu *Bayesian, Multilayer Perceptron, Sequential Minimal Optimization* (SMO) dan Pokok Keputusan, Hutan Rawak, Pokok Rawak dan Naive Bayes bagi meramal kebolehpasaran graduan. Hasil kajian mendapati Hutan Rawak merupakan algoritma terbaik dengan ketepatan 71.304% dan atribut projek pelajar, silibus pengajian, jumlah jam kredit pengajian dan sifat empati pelajar kepada pelajar lain menyumbang kepada kebolehpasaran graduan tersebut (Mishra et al., 2016).

Jadual 2.1 menunjukkan ringkasan kajian literatur kebolehpasaran graduan melalui pembelajaran mesin yang pernah dijalankan oleh pengkaji yang lepas.

Jadual 2.1 Ringkasan Kajian Literatur Kebolehpasaran Graduan Melalui Pembelajaran Mesin

Pengkaji	Objektif Kajian	Algoritma Kajian	Sumber Data	Algoritma Terbaik	Keputusan (%)
KianLam Tan, Nor Azziaty Abdul Rahman, ChenKim Lim, 2019	Mengenal pasti faktor penting kebolehpasaran di kalangan graduan baharu	LR,DT,KNN,NB,SVM ,ANN	SKPG,KPT	ANN	59.01
Aziz M & Yusof Y, 2016	Meramal pengelasan graduan MARA Professional College samada bekerja, tidak bekerja atau menyambung pengajian.	LR,DT,KNN,NB,ANN	MARA Professional College, KPM	LR	92.47
Jantawan B, Tsai C, 2013	Membangunkan model kadar kebolehpasaran graduan menggunakan kaedah pengkelasan	DT,NB	Maejo University Thailand	NB	99.77
García-Peñalvo,FCruz-Benito, JMartín-González, M et al., 2018	Membina model ramalan menggunakan algoritma pembelajaran mesin bagi mengekstrak faktor relevan kebolehpasaran graduan bagi menjalankan analisis lanjutan	RF	Spanish Observatory for Employability and Employment (OEEU),	RF	71.00
Kumar D, Jambheshwar G, 2016	Meramal kebolehpasaran pelajar Master of Computer Applications (MCA)	DT,RF,RT,SMO,MLP, NB	Survey Pelbagai Kolej di India	RF	71.30
Rahman N,Tan K,Lim C, 2017	Mencadangkan model pengkelasan yang sesuai dalam membuat ramalan penilaian atribut pelajar untuk memenuhi kriteria pemilihan kerja yang diperlukan oleh industri	LR,DT,KNN,NB,SVM , ANN	SKPG,KPT	-	-

2.5 TEKNIK PEMILIHAN ATRIBUT

Pemilihan atribut memainkan peranan yang penting dalam menentukan prestasi pengelasan dengan menentukan samada perlu memasukan atau mengecualikan atribut tersebut dalam proses pemodelan. Pemilihan atribut akan memberi kesan kepada prestasi pemodelan samada semakin meningkat atau menurun. Atribut yang relevan adalah penting bagi tujuan proses latihan pemodelan disebabkan ianya mempunyai aspek maklumat yang boleh meningkatkan prestasi pengelasan dan sebaliknya atribut yang tidak relevan akan memberi kesan kepada penurunan prestasi pengelasan (Al-Harbi, O. 2019).

Antara teknik pemilihan atribut yang boleh digunakan bagi tujuan memilih atribut yang relevan adalah perolehan maklumat (*Information Gain)*, Ujian *Chi Square*, Skor Fisher dan ujian korelasi. Teknik perolehan maklumat mengira pengurangan entropi dari transformasi set data. Ia boleh digunakan bagi tujuan pemilihan atribut dengan menilai perolehan maklumat bagi setiap atribut dalam konteks atribut sasaran. Ujian Chi Square pula digunakan bagi set data berciri kategori data. Teknik ini mengira chi-square antara setiap atribut dan sasaran dan memilih bilangan atribut yang diinginkan berpandukan skor *chi-square* yang terbaik. Teknik Skor Fisher merupakan kaedah pemilihan atribut yang terselia dimana algoritma yang digunakan akan menghasilkan keputusan kedudukan bagi setiap atribut berdasarkan skor fisher dalam urutan menurun. Manakala teknik ujian korelasi pula akan mengukur hubungan linear antara 2 atau lebih atribut bagi memilih atribut yang mempunyai kolerasi yang tinggi dengan atribut sasaran. Melalui teknik ini, ramalan terhadap atribut lain boleh dilaksanakan (Jantawan et al., 2014).

2.6 TEKNIK PENALAAN PARAMETER

Penalaan parameter dalam algoritma pembelajaran mesin adalah proses penting untuk mendapatkan nilai parameter yang optimum bagi algoritma pembelajaran mesin. Beberapa kajian telah membuktikan proses penalaan parameter menghasilkan model pengelasan yang mempunyai ketepatan yang tinggi. Proses penalaan parameter bergantung kepada keputusan eksperimen dan bukan hasil daripada keputusan teori (Siji et al., 2020).

Salah satu kaedah dalam proses penalaan parameter adalah menggunakan kaedah carian grid. Kaedah carian grid merupakan kaedah alternatif untuk mencari parameter terbaik bagi sesuatu model. Kaedah ini dikategorikan sebagai kaedah lengkap bagi mencari nilai parameter yang terbaik dengan menetapkan nilai awal parameter tersebut. Seterusnya kaedah ini akan menunjukkan nilai terbaik bagi setiap parameter yang mana akan dipilih sebagai parameter model tersebut. Kaedah ini juga dapat digunakan dengan mengenal pasti batas teratas dan terbawah bagi setiap pemboleh ubah bebas (Ramadhan et al., 2017).

2.7 RUMUSAN

Berdasarkan kepada pengetahuan yang telah diperolehi melalui kajian literatur, kebanyakan kajian yang dijalankan untuk meramal kadar kebolehpasaran graduan adalah bersifat umum dimana tiada kajian yang spesifik dijalankan untuk melihat kadar kebolehpasaran graduan samada bekerja dalam bidang atau diluar bidang. Bagi kajian lepas yang menggunakan data set yang diperolehi daripada Sistem Kajian Pengesanan Graduan (SKPG) yang dilaksanakan oleh Kementerian Pengajian Tinggi didapati bahawa peratus ketepatan model pengelasan adalah di bawah 60 peratus.

Antara algoritma yang telah digunakan oleh pengkaji terdahulu untuk membangunkan model pengelasan berkaitan kebolehpasaran graduan berdasarkan kajian literatur terdahulu adalah LR ,DT, KNN, NB, SVM dan ANN. Hasil kajian literatur juga mendapati bahawa NB selalu diguna pakai sebagai algoritma pembelajaran asas bagi meramal kadar kebolehpasaran graduan disebabkan oleh NB kebiasaannya diguna pakai untuk meramal pengelasan bagi data berbentuk kategori. Namun berdasarkan kajian literatur yang telah dilaksanakan, masih belum ada pengkaji yang menggunakan model penggabungan bagi meramal kadar kebolehpasaran graduan. Sehubungan itu, kajian ini akan menggunakan algoritma pembelajaran penggabungan bagi mengisi jurang pembelajaran mesin tersebut dalam usaha untuk meningkatkan peratus ketepatan model bagi data yang diperolehi daripada SKPG. Selain itu, Kajian ini juga menggunakan kaedah carian grid dalam proses penalaan parameter dan *chi-square* dalam melaksanakan proses pemilihan atribut yang mempengaruhi kebolehpasaran graduan samaada bekerja dalam bidang atau luar bidang pengajian

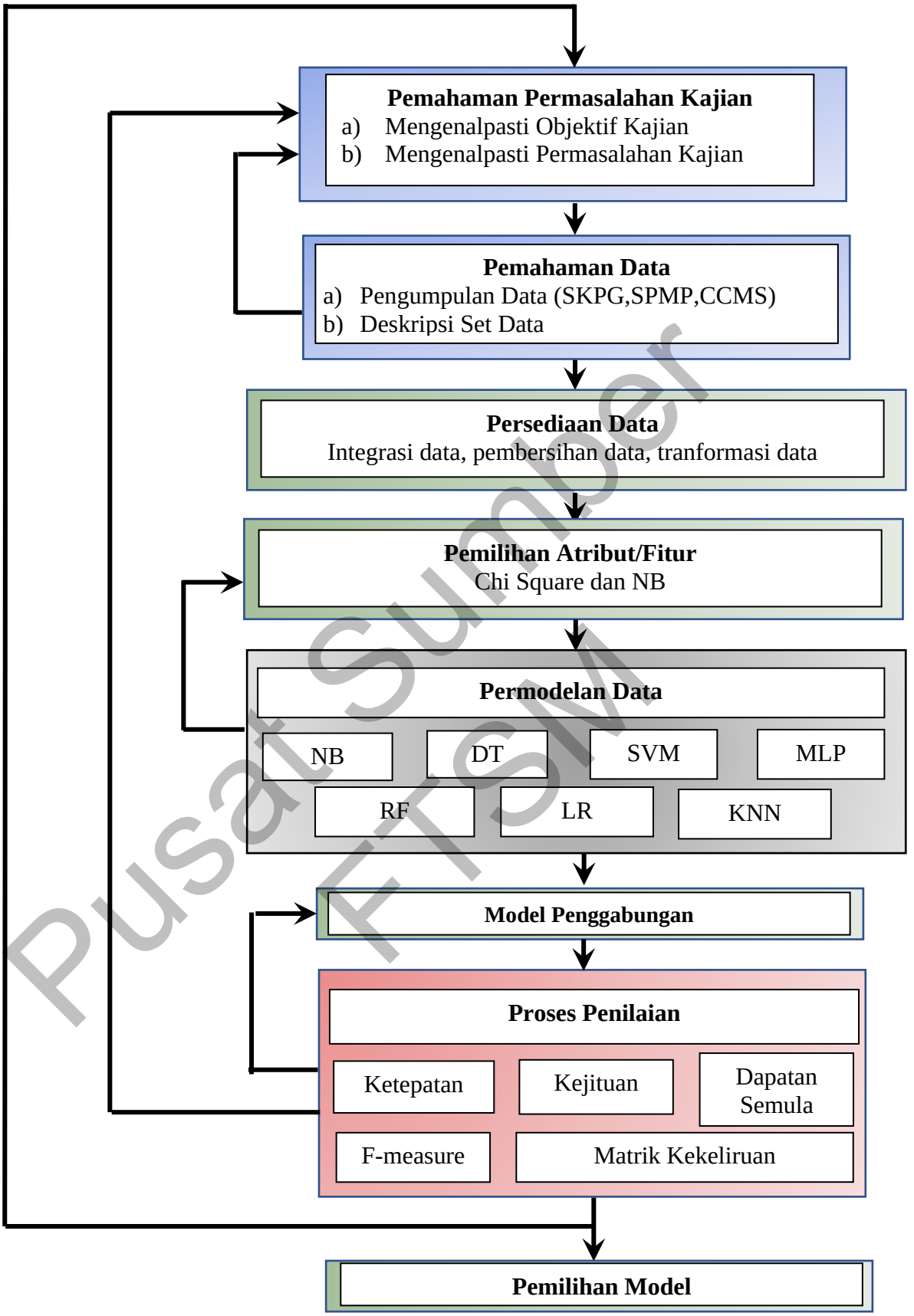
BAB III

KAEDAH KAJIAN

3.1 PENGENALAN

Metodologi kajian ini terbahagi kepada tiga peringkat utama. Peringkat pertama dimulakan dengan memahami pemasalahan kajian, memahami data dan menganalisis kuantiti, perincian, kelas dan atribut data tersebut. Peringkat kedua melibatkan aktiviti-aktiviti pra-pemprosesan bagi tujuan menyiapkan data untuk menjalankan proses pemodelan data. Antara aktiviti-aktiviti pra-pemprosesan yang telah dilakukan dalam kajian ini meliputi proses transformasi data, integrasi data, pembersihan data dan pengurangan atribut atau fitur. dan.

Data yang telah diproses pra-proses kemudian digunakan pada tahap ketiga untuk mengenal pasti teknik pembelajaran mesin yang terbaik dengan membangunkan analisis perbandingan antara teknik-teknik yang dipilih. Di akhir proses analisis, analisis ketepatan, kejituan, dapatan semula, *F-measure* dan Matriks Kekeliruan bagi model pengelasan yang dipilih dibandingkan bagi menentukan model atau algoritma yang terbaik bagi meramal kebolehpasaran graduan TVET samada bekerja dalam bidang atau luar bidang. Gambaran keseluruhan metodologi kajian ini adalah seperti di Rajah 3.



Rajah 3.1 Kerangka Kajian Analisis Ramalan Kebolehpasaran Graduan TVET

3.2 PEMAHAMAN PERMASALAHAN KAJIAN

Pemahaman permasalahan kajian melibatkan dua (2) komponen utama iaitu mengenal pasti objektif dan permasalahan kajian. Dalam kajian ini, permasalahan utama adalah mengenal pasti apakah faktor-faktor utama yang menyumbang dan mempengaruhi kebolehpasaran graduan TVET samada bekerja di dalam atau luar bidang yang diceburi. Faktor-faktor yang diberi tumpuan dalam kajian ini adalah berkaitan demografi dan bidang pengajian serta faktor-faktor lain yang disoal dalam kajian kebolehpasaran graduan yang dilaksanakan oleh KPT yang boleh mempengaruhi kebolehpasaran graduan samada bekerja di dalam atau luar bidang.

3.3 PEMAHAMAN DATA

Data yang digunakan dalam kajian ini diperolehi daripada pangkalan data tiga sistem utama iaitu Sistem Kajian Pengesan Graduan (SKPG), Sistem Pengurusan Maklumat Politeknik (SPMP) dan Sistem Pengurusan Maklumat Kolej Komuniti (CCMS). Sistem SKPG mengandungi data kebolehpasaran graduan yang telah di isi oleh graduan semasa menghadiri konvokesyen di politeknik dan kolej komuniti. Sistem SPMP pula mengandungi data-data pelajar politeknik bermula dari pelajar tersebut mendaftar di politeknik sehingga pelajar tersebut menamatkan pengajian. Manakala Sistem CCMS pula mengandungi data-data pelajar kolej komuniti bermula dari pelajar tersebut mendaftar di kolej komuniti sehingga pelajar tersebut menamatkan pengajian. Bilangan sampel bagi kajian ini adalah 42,807 responden yang merupakan graduan TVET bagi tahun 2018 dan 2019. Manakala kaedah pensampelan yang digunakan adalah pensampelan rawak berstrata dimana responden yang dikaji merupakan graduan politeknik yang telah menamatkan pengajian pada tahun tersebut.

JPPKK sebagai pemilik data telah memberi keizinan bagi menggunakan data-data tersebut bagi tujuan kajian ini disebabkan data tersebut dilindungi di bawah Akta Perlindungan Data Peribadi 2010. Bagi menjaga kerahsiaan data, maklumat peribadi graduan tidak akan didedahkan dimana hanya maklumat umum berkaitan butiran diri dan maklumat akademik sahaja akan dikaji. Atribut-atribut bagi data mentah yang akan dikaji dalam kajian ini adalah seperti di Jadual 3.1. Perincian bagi setiap atribut data set kajian ini adalah seperti di Lampiran B.

Jadual 3.1 Kategori dan Atribut bagi Data Kebolehpasaran Graduan TVET (Politeknik dan Kolej Komuniti)

Bil	Kategori	Atribut
1	Maklumat peribadi	No pendaftaran, Jantina, Bangsa, Agama, Alamat, Daerah, Negeri, Poskod, No.Telefon, Tarikh Lahir, Parlimen, Dun, Status OKU, Jenis kecacatan
2	Maklumat Akademik	Program Pengajian, Bidang pengajian, Peringkat Pengajian, Asrama, Penajaan, Kelayakan Masuk, HPNM, Tahun Mula Pengajian, Tahun Tamat Pengajian,
3	Maklumat Institusi Pengajian	Nama Institusi, Kategori Institusi, alamat, Daerah, Negeri, Poskod, No. Telefon, Status Pengiktirafan
4	Maklumat Kebolehpasaran	Status Latihan Industi (LI), Nama Syarikat LI, Alamat Syarikat, Poskod Syarikat, Negeri Syarikat, Status Kebolehpasaran. Taraf pekerjaan, taraf jawatan, sektor pekerjaan, jawatan, pendapatan bulanan, sektor ekonomi, tempoh mendapat pekerjaan, bilangan temu duga, status pekerjaan (dalam atau luar bidang pengajian), sebab utama tidak bekerja, kaedah mendapatkan maklumat kekosongan pekerjaan
5	Maklumat lain	Kesesuaian kandungan pengajian, Program latihan industri/praktikum, Mata pelajaran wajib institusi, penyediaan pelajar untuk menghadapi dunia pekerjaan, maklumat peluang pekerjaan dan kerjaya, bantuan mendapatkan pekerjaan, interaksi tenaga pengajar dengan pelajar, penyampaian kuliah dan kualiti pengajaran, kemudahan prasarana di institusi pengajian.

3.4 PRA-PEMPROSESAN DATA

Pra-pemprosesan data adalah teknik penting untuk mengubah set data mentah menjadi bentuk atau format data yang dapat difahami oleh mesin atau alat bagi tujuan pemprosesan selanjutnya. Kebanyakan data mentah yang diperolehi selalunya tidak konsisten, tidak mencukupi, tidak lengkap dan terdapat banyak kesilapan, *noisy* dan data terasing. *Noisy* data merujuk kepada data-data yang tidak memberi makna dan tidak boleh digunakan untuk menjalankan analisis. Manakala data terasing pula merujuk kepada data yang berada diluar kebiasaan data bagi sesuatu sampel data ataupun dikenali sebagai data asing. Oleh itu, untuk memastikan data dalam format yang boleh difahami dan dalam bentuk yang sesuai bagi tujuan menjalankan pemprosesan selanjutnya, pra-pemprosesan data mesti dilakukan bagi mengatasi masalah-masalah berkaitan kualiti data tersebut (Sani et al., 2018).

Antara aktiviti pra-pemprosesan yang telah dilakukan dalam kajian ini meliputi proses pembersihan data, integrasi data, transformasi data dan pengurangan dan pemilihan atribut atau fitur, teknik kejuruteraan fitur dan pengurangan dimensi. Pada masa kini, terdapat banyak alat dan perisian perlombongan data yang dapat digunakan bagi tujuan pemprosesan data. Dalam kajian ini, perisian 'PyCharm' digunakan sebagai perisian untuk tujuan menjalankan proses pra-pemprosesan data dan membina model pengelasan. PyCharm adalah perisian pembelajaran mesin yang menggunakan bahasa Python yang telah dibangunkan oleh organisasi dan komuniti bukan berasaskan kepada keuntungan. Kedua-dua perisian ini mengandungi pelbagai algoritma pembelajaran mesin dan visualisasi bagi tujuan menjalankan proses pra-pemprosesan data, analisis data dan membangunkan model ramalan untuk kajian ini.

3.4.1 Pengintegrasian Data

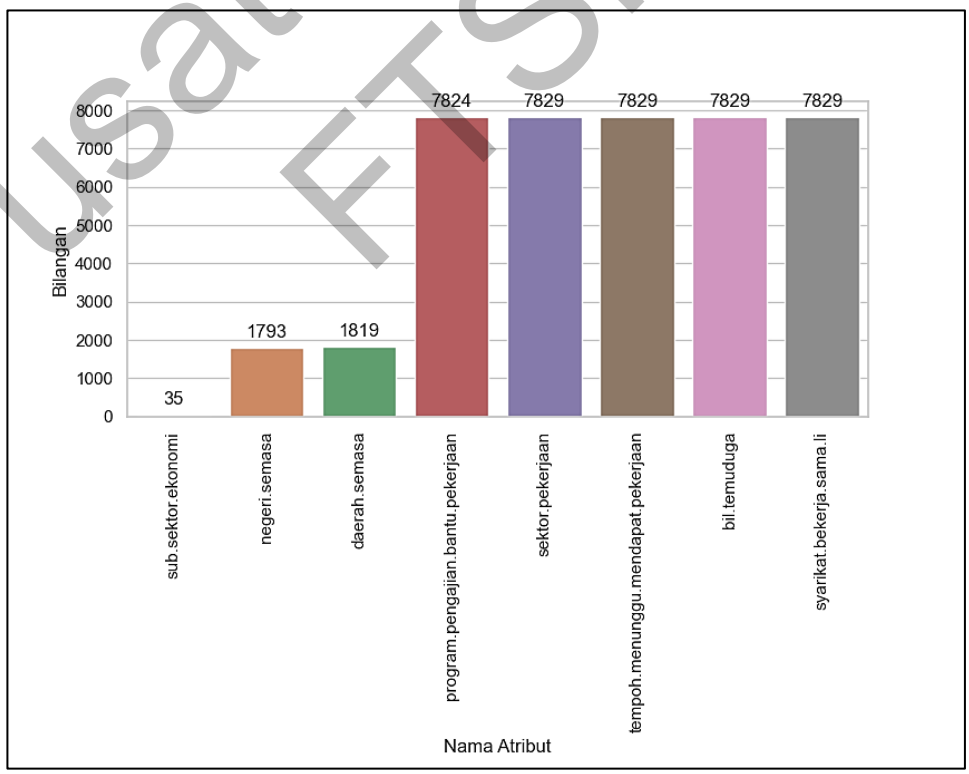
Proses integrasi data dijalankan bagi menggabungkan set data yang diperolehi daripada pelbagai sumber pangkalan data. Dalam kajian ini data yang diperolehi dari pangkalan data sistem SKPG, SPMP dan CCMS digabungkan bagi mendapatkan satu data set yang besar dan lengkap bagi tujuan menjalankan analisis dan membangunkan model ramalan kebolehpasaran graduan TVET. Proses integrasi data ini adalah penting bagi mengelakkan dan mengurangkan berlakunya data yang tidak konsisten dan pertindihan

serta memastikan semua atribut berkaitan analisis kebolehpasaran graduan TVET tidak tertinggal selain membantu meningkatkan ketepatan dan masa untuk memproses data tersebut.

3.4.2 Pembersihan Data

Proses pembersihan data perlu bagi mengendalikan data kurang berkualiti, menyelesaikan masalah data yang tidak konsisten, menggantikan data yang hilang, mengenal pasti dan membuat keputusan berkaitan data asing atau *outlier* samada perlu menyimpan atau memadam data tersebut dan memperbaiki data yang rosak. Tujuan utama pembersihan data mentah ini adalah untuk memastikan dan meningkatkan keputusan bagi pembangunan dan pengujian model pembelajaran mesin dan seterusnya meningkatkan tahap kepercayaan terhadap keputusan analisis yang dijalankan.

Dalam kajian ini, data yang telah digabungkan diproses untuk menggantikan data yang hilang. Data-data yang hilang digantikan menggunakan algoritma KNN. Atribut dan bilangan data-data yang hilang adalah seperti di Lampiran A dan Rajah 3.2.



Rajah 3.2 Atribut dan Bilangan Data Hilang

3.4.3 Transformasi Data

Transformasi Data adalah proses mengubah format, struktur, atau nilai sesuatu data. Format, struktur atau nilai data mentah yang tidak dapat difahami oleh algoritma atau mesin perlu ditransformasi kepada format yang boleh difahami oleh algoritma atau mesin supaya ianya dapat diproses dan dianalisis dengan lebih tepat. Antara proses-proses yang digunakan dalam transformasi data adalah penormalan data. Dalam proses penormalan, data diubah sehingga berada di bawah julat tertentu. Apabila atribut berada pada skala yang berbeza, proses pemodelan dan perlombongan data mungkin menjadi sukar. Justeru, proses penormalan data akan membantu dalam proses perlombongan data dan mengekstrak data dengan lebih cepat.

Set data yang telah melalui proses pra-pemprosesan atau dikenali dengan set data bersih akan diproses untuk tujuan latihan dan ujian bagi membangunkan model ramalan kebolehpasaran graduan menggunakan beberapa algoritma model pembelajaran yang telah dipilih.

3.4.4 Pemilihan dan Pengurangan Fitur

Pemilihan dan pengurangan fitur adalah proses pengurangan atribut semasa membangunkan model ramalan dengan mengenal pasti atribut yang mempengaruhi keputusan atribut sasaran dari set data mentah. Jaffar et al. dalam kajian mereka mendapati pemilihan dan pengurangan fitur boleh mengurangkan masa pemprosesan dan pemodelan dan juga boleh meningkatkan prestasi model yang dibangunkan serta boleh mengelakkan berlakunya *over-fitting* semasa proses pemodelan (Jaffar et al., 2019). Pemilihan dan pengurangan fitur boleh dijalankan melalui pelbagai algoritma atau metodologi seperti Pokok Keputusan, Hutan Rawak, ujian korelasi, Regresi Linear, ujian Chi Square dan sebagainya. Berdasarkan data mentah yang diperolehi jumlah keseluruhan atribut adalah sebanyak 54 atribut. Walaubagaimanapun tidak semua atribut digunakan sebagai atribut untuk tujuan pemodelan. Berdasarkan kajian literatur berkaitan teknik pemilihan atribut, ujian Chi Square merupakan salah satu teknik yang sesuai bagi pemilihan atribut bagi jenis datasset yang bersifat kategori data. Hasil ujian chi square 10 atribut yang mempunyai skor tertinggi dipilih bagi tujuan pemodelan pembelajaran mesin berdasarkan hasil keputusan ketepatan yang tinggi

semasa proses pemodelan awal dijalankan dengan menggunakan algoritma NB. Skor dan kedudukan atribut adalah seperti di Jadual 3.2 dan keputusan ketepatan model berdasarkan bilangan atribut adalah seperti di Jadual 3.3.

Jadual 3.2 Skor dan Kedudukan Atribut

Atribut	***Chi Square* Skor**	**Kedudukan**
syarikat.bekerja.sama.li	1067.00	1
program.pengajian.bantu.pekerjaan	326.10	2
Jantina	173.00	3
sektor.ekonomi	147.20	4
sub.sektor.ekonomi	143.40	5
taraf.pekerjaan	131.00	6
pendapatan.bulanan	101.60	7
cadangan.belajar.di.institusi	61.65	8
status.li	36.59	9
kod.kursus	36.15	10
tempoh.menunggu.mendapat.pekerjaan	28.94	11
bantuan.mendapatkan.pekerjaan	24.74	12
manfaat.LI	24.54	13
kesesuaian.kandungan.pengajian	19.89	14
bil.temuduga	17.45	15
maklumat.peluang.pekerjaan.kerjaya	15.93	16
pendapatan.bulanan.keluarga	15.60	17
persediaa.pelajar.menghadapi.dunia.pekerjaan	14.90	18
kod.nec	11.85	19
keturunan	10.94	20
fb	10.82	21
program.LI	8.04	22
peringkat.pengajian	7.87	23
sektor.pekerjaan	6.41	24
mata.pelajaran.wajib.institusi	6.36	25
interaksi.dengan.pelajar	4.33	26
bulan.konvo	3.73	27
ig	3.57	28
asrama	3.27	29
ict	3.26	30

bersambung

sambungan..		
penyampaian.kuliah.kualiti.pengajaran	2.55	31
kafeteria	2.42	32
cgpa	2.01	33
makmal.bengkel	1.71	34
dewan.kuliah.bilik.tutorial	1.64	35
No.Matrik	1.57	36
daftar.lanjut.pengajian	1.55	37
bulan.tamat.pengajian	1.34	38
umur	1.11	39
poskod.tetap	0.73	40
twitter	0.45	41
status.pelajar.semasa	0.41	42
tahun.mula.pengajian	0.31	43
negeri.tetap	0.21	44
daerah.tetap	0.20	45
penaja.pengajian	0.19	46
negeri.semasa	0.12	47
tahun.tamat.pengajian	0.12	48
daerah.semasa	0.10	49
nama.institusi	0.08	50
bulan.mula.pengajian	0.01	51
status.oku	0.00	52

Jadual 3.3 Bilangan Atribut dan Keputusan Ketepatan Model NB

Jumlah Atribut Terpilih	**Ketepatan (%)**
8	68.84
9	69.04
10	**69.26**
11	69.14
12	69.09
13	69.11
14	68.84
15	68.78
16	68.47

3.4.5 Laporan Kualiti Data

Dalam kajian ini, laporan kualiti data dihasilkan bagi mendapatkan gambaran awal bentuk dan kualiti data yang telah dikumpulkan. Ia juga bertujuan untuk melihat dan mengenal pasti atribut yang mengandungi data hilang, *outliers* dan sisihan piawai bagi data set tersebut. Selain itu, ia juga bertujuan untuk melihat kardinaliti bagi setiap atribut yang boleh digunakan semasa proses pemilihan atribut bagi digunakan semasa penghasilan model pembelajaran mesin. Laporan penuh kualiti data bagi kajian ini adalah seperti di jadual 3.4 dan jadual

Jadual 3.4 Laporan Kualiti Data (Bukan Kategori Data)

Nama Atribut	Bilangan	Bilangan Data Hilang (%)	Kardinaliti	Minimum	Kuartil 1	Purata	Median	Kuartil 3	Maximum	Sisihan Piawai
Umur	42807	0	28	19	26	22.819	33	40	47	2.059
cgpa	42807	0	200	2	2.5	2.942	3	3.5	4.00	0.388

Jadual 3.5 Laporan Kualiti Data (Kategori Data)

Nama Atribut	Bilangan	Bilangan Data Hilang (%)	Kardinaliti	Mod	Kekerapan Mod	% Mod	Mod Kedua	Kekerapan Mod Kedua	% Mod Kedua
Jantina	42807	0	2	2	22399	52.33%	1	20408	47.67%
keturunan	42807	0	16	1	34216	79.93%	3	3006	7.02%
negeri.semasa	42807	1793 (4.18%)	17	10	6910	16.14%	1	4309	10.07%
daerah.semasa	42807	1819 (4.25%)	128	102	1969	4.60%	1400	1834	4.28%
poskod.tetap	42807	0	1305	8000	723	1.69%	93050	589	1.38%

bersambung...

sambungan...

negeri.tetap	42807	0	16	10	6480	15.14%	2	4608	10.76%
daerah.tetap	42807	0	128	102	1693	3.95%	1008	1546	3.61%
fb	42807	0	2	2	31879	74.47%	1	10928	25.53%
twitter	42807	0	2	2	40846	95.42%	1	1961	4.58%
ig	42807	0	2	2	34013	79.46%	1	8794	20.54%
status.oku	42807	0	2	2	42600	99.52%	1	207	0.48%
nama.institusi	42807	0	34	93	2752	6.43%	99	2677	6.25%
peringkat.pengajian	42807	0	5	1	42389	99.02%	41	267	0.62%
kod.kursus	42807	0	85	DA001	3958	9.25%	DM001	3436	8.03%
kod.nec	42807	0	27	523	6077	14.20%	526	4957	11.58%
bulan.mula.pengajian	42807	0	4	6	27434	64.09%	12	15155	35.40%
tahun.mula.pengajian	42807	0	6	2015	19828	46.32%	2014	19058	44.52%
bulan.tamat.pengajian	42807	0	8	6	21478	50.17%	12	19877	46.43%
tahun.tamat.pengajian	42807	0	5	2017	19875	46.43%	2018	17610	41.14%
bulan.konvo	42807	0	7	8	13809	32.26%	7	11253	26.29%
status.li	42807	0	2	1	42646	99.62%	4	161	0.38%
penaja.pengajian	42807	0	14	2	28911	67.54%	5	13129	30.67%
pendapatan.bulanan.keluarga	42807	0	10	3	9962	23.27%	4	7360	17.19%
kesesuaian.kandungan.pengajian	42807	0	5	5	19076	44.56%	4	18779	43.87%
program.LI	42807	0	6	5	20448	47.77%	4	16627	38.84%
mata.pelajaran.wajib.institusi	42807	0	6	5	19485	45.52%	4	17783	41.54%

bersambung...

sambungan..

persediaa.pelajar.menghadapi.dunia.pekerjaan	42807	0	5	5	22449	52.44%	4	16038	37.47%
manfaat.LI	42807	0	6	5	22158	51.76%	4	14581	34.06%
maklumat.peluang.pekerjaan. kerjaya	42807	0	6	5	19600	45.79%	4	16663	38.93%
bantuan.mendapatkan.pekerjaan	42807	0	6	5	18179	42.47%	4	16358	38.21%
interaksi.dengan.pelajar	42807	0	5	5	24382	56.96%	4	15213	35.54%
penyampaian.kuliah.kualiti. pengajaran	42807	0	6	5	22686	53.00%	4	15665	36.59%
makmal.bengkel	42807	0	6	5	20178	47.14%	4	15916	37.18%
dewan.kuliah.bilik.tutorial	42807	0	6	5	20572	48.06%	4	16252	37.97%
kafeteria	42807	0	6	5	18760	43.82%	4	16424	38.37%
asrama	42807	0	6	5	18253	42.64%	4	15527	36.27%
ict	42807	0	6	5	18045	42.15%	4	15468	36.13%
cadangan.belajar.di.institusi	42807	0	2	1	41179	96.20%	2	1628	3.80%
daftar.lanjut.pengajian	42807	0	2	2	39627	92.57%	1	3180	7.43%
status.pelajar.semasa	42807	0	2	5	39604	92.52%	2	3203	7.48%
taraf.pekerjaan	42807	0	5	8	32813	76.65%	4	3997	9.34%
pendapatan.bulanan	42807	0	11	3	17750	41.47%	4	10191	23.81%
sektor.pekerjaan	42807	7829 (18.29%)	7	4	17546	40.99%	8	9245	21.60%
sektor.ekonomi	42807	0	21	19	7923	18.51%	3	7076	16.53%

bersambung...

sambungan...									
sub.sektor.ekonomi	42807	35 (0.08%)	88	1996	6098	14.25%	956	2409	5.63%
program.pengajian.bantu.pekerjaan	42807	7824 (18.27%)	5	4	14753	34.46%	5	11186	26.13%
tempoh.menunggu.mendapat.pekerjaan	42807	7829 (18.29%)	14	2	14413	33.67%	1	5054	11.81%
bil.temuduga	42807	7829 (18.29%)	3	1	33185	77.52%	2	1364	3.19%
statu.pekerajaan.dalam.bidang	42807	0	2	1	22513	52.59%	2	20294	47.41%
syarikat.bekerja.sama.li	42807	7829 (18.29%)	2	2	25365	59.25%	1	9613	22.46%

3.5 PEMODELAN PEMBELAJARAN MESIN

Kajian ini dijalankan dengan menggunakan tujuh teknik pembelajaran mesin iaitu Naïve Bayes, Mesin Sokongan Vektor (SVM), *K-Nearest Neighbours* (KNN), Rangkaian Neural Buatan (ANN), Regresi Logistik (LR) Hutan Rawak (RF) dan Pokok Keputusan (DT) bagi mencapai objektif kajian. Kajian ini menggunakan kaedah ramalan berdasarkan pengelasan data dimana model-model pengelasan akan dilatih dan diuji dengan set data yang berlainan bagi membangunkan model tersebut. Data latihan dan ujian dipecahkan kepada 70 peratus bagi tujuan latihan dan 30 peratus lagi bagi tujuan ujian. Dalam proses pemodelan pembelajaran mesin kaedah pengesahan silang lipatan (*cross-validatian fold*) digunakan bagi sampel data latihan dan nilai K-lipatan (*K-Fold*) bagi kajian ini adalah k=10. 10 kali uji kaji dilaksanakan bagi setiap model yang dibangunkan dan purata prestasi dan sisihan piawai bagi setiap uji kaji direkodkan bagi melihat kestabilan model yang dibangunkan. Proses pemodelan ini dijalankan menggunakan perisian PyCharm.

3.6 PROSES PENILAIAN

Proses penilaian bagi kajian ini berpandukan kepada prestasi setiap model yang dibangunkan berdasarkan kepada ketepatan, kejituan, dapatan semula, *F-Measure* dan matriks kekeliruan. Dalam kajian ini, hasil utama dapatan dinilai berdasarkan peratus ketepatan ramalan dan sisihan piawai model-model tersebut. Peratus ketepatan ramalan dikira berdasarkan jumlah data yang diramalkan betul daripada jumlah bilangan set data keseluruhan. Semakin tinggi peratus ketepatan ramalan model, semakin tinggi prestasi model ramalan tersebut. Nilai sisihan piawai pula dinilai bagi melihat kestabilan model yang dibina.

Fasa pertama proses pemilihan model dalam kajian ini adalah dengan membuat penilaian ketepatan setiap model yang diuji. Model yang menghasilkan ketetapan yang tinggi dipilih untuk ke fasa kedua pemilihan iaitu dengan membina model penggabungan dengan menggabungkan beberapa model tersebut bagi menghasilkan satu model yang mempunyai ketepatan yang lebih tinggi. Peratus ketepatan ramalan yang tertinggi hasil daripada model penggabungan ini akan dinilai dan dipilih sebagai

model yang digunakan bagi meramal kebolehpasaran graduan politeknik dan kolej komuniti samada bekerja dalam bidang atau di luar bidang.

Analisis bagi ketepatan, kejituan, dapatan semula dan *F-measure* dikira menggunakan persamaan seperti dibawah:

$$Ketepatan = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.1)$$

$$Kejituan = \frac{TP}{TP + FP} \quad (3.2)$$

$$Dapatan\ Semula = \frac{TP}{TP + FN} \quad (3.3)$$

$$F - measure = 2\frac{(Dapatan\ Semula\ x\ Kejituan)}{(Dapatan\ Semula + Kejituan)} \quad (3.4)$$

di mana TP is a benar positif, TN is a benar negatif, FP is a salah positif, and FN is a salah negatif.

Prestasi model turut dinilai berdasarkan matriks kekeliruan yang dihasilkan oleh model-model yang dikaji. Matriks kekeliruan bagi kajian ini adalah seperti di jadual 3.6.

Jadual 3.6 Matriks Kekeliruan Kajian Kebolehpasaran Graduan TVET

		RAMALAN	
		Positif (Dalam Bidang)	*Negatif (Luar Bidang)*
SEBENAR	*Kelas Positif (Dalam Bidang)*	Positif Sebenar (TP)	Negatif Palsu (FN)
	Kelas Negatif (Luar Bidang)	Positif Palsu (FP)	Negatif Sebenar (TN)

Pentaksiran bagi matriks kekeliruan bagi kajian ini adalah seperti di jadual 3.7. Pentafsiran bagi matriks kekeliruan ini adalah berdasarkan objektif utama kajian iaitu untuk mengenal pasti faktor yang mempengaruhi graduan TVET samada bekerja dalam bidang atau luar bidang yang diceburi.

Jadual 3.7 Pentafsiran Matriks Kekeliruan Berdasarkan Objektif Kajian

Matriks Kekeliruan	**Pentafsiran**
Positif Sebenar (TP)	Graduan yang diramalkan betul bekerja di dalam bidang
Positif Palsu (FP)	Graduan yang diramalkan dengan salah bekerja di dalam bidang (Diramal bekerja di dalam bidang tetapi sebenarnya adalah bekerja di luar bidang)
Negatif Sebenar (TN)	Graduan yang diramalkan betul bekerja luar bidang
Negatif Palsu (FN)	Graduan yang diramalkan dengan salah bekerja di luar bidang (Diramal bekerja di luar bidang tetapi sebenarnya adalah bekerja di dalam bidang)

Seterusnya, ujian Cochran's Q dilaksanakan bagi membandingkan prestasi pelbagai model pengelasan. Ujian Cochran's Q merupakan versi umum ujian McNemar yang digunakan untuk menilai beberapa model pengelasan. Hipotesis null (H_0) yang ditetapkan bagi tujuan ujian ini adalah dengan membuktikan bahawa tiada perbezaan yang signifikan diantara tiga (3) model penggabungan yang mempunyai peratus ketetapan yang tertinggi iaitu RF+KNN+DT, RF+KNN+LR dan RF+KNN+NB bersama dengan dua (2) model tunggal iaitu RF dan KNN dengan menetapkan pengukuran nilai p atau p-value < 0.05 (atau 5% tahap signifikan) dimana hipotesis null (H_0) akan ditolak jika nilai p lebih kecil daripada 0.05. Formula bagi ujian Cochran's Q adalah seperti berikut.

$$Qc = (L-1)\frac{L\sum_{i=1}^{L} G_i^2 - T^2}{LT - \sum_{j=1}^{N_{ts}} (L_j)^2} \qquad (3.5)$$

dimana

G_i= bilangan sampel yang betul dikelaskan oleh model

L_i= bilangan model pengelasan

T = jumlah bilangan undi yang betul di antara model pengelasan L dan

N_{ts}= bilangan sampel ujian dataset

3.7 RUMUSAN

Dalam kajian ini, data mentah yang diperolehi daripada sistem SKPG, SPMP dan CCSM telah dibersihkan dalam fail excel menggunakan teknik pra-pemprosesan data sebelum ditukarkan kepada fail csv dan dimuat naik ke dalam perisian Pycharm untuk diproses bagi tujuan membangunkan model ramalan kebolehpasaran graduan politeknik dan kolej komuniti tersebut. Model ramalan yang terhasil daripada beberapa teknik pembelajaran mesin seperti Naïve Bayes, Mesin Sokongan Vektor (SVM), *K-Nearest Neighbours* (KNN), Rangkaian Neural Buatan (ANN), Regresi Logistik (LR) Hutan Rawak (RF) dan Pokok Keputusan (DT) diukur peratusan ketepatan ramalannya bagi pemilihan beberapa model terbaik untuk digunakan dalam penghasilan model penggabungan yang menghasilkan peratus ketepatan yang lebih tinggi. Seterusnya hasil keputusan daripada model pembelajaran penggabungan yang terbaik dipilih untuk meramal kebolehpasaran graduan politeknik dan kolej komuniti samada bekerja dalam bidang atau luar bidang pengajian

BAB VII

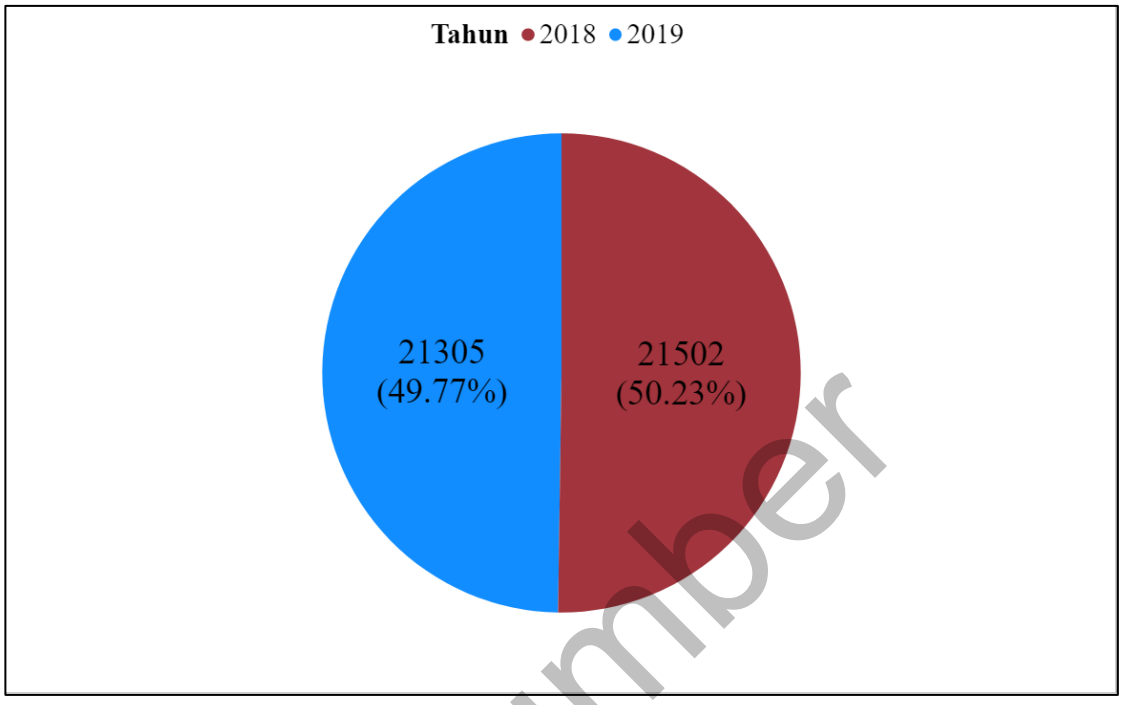
DAPATAN KAJIAN

4.1 PENGENALAN

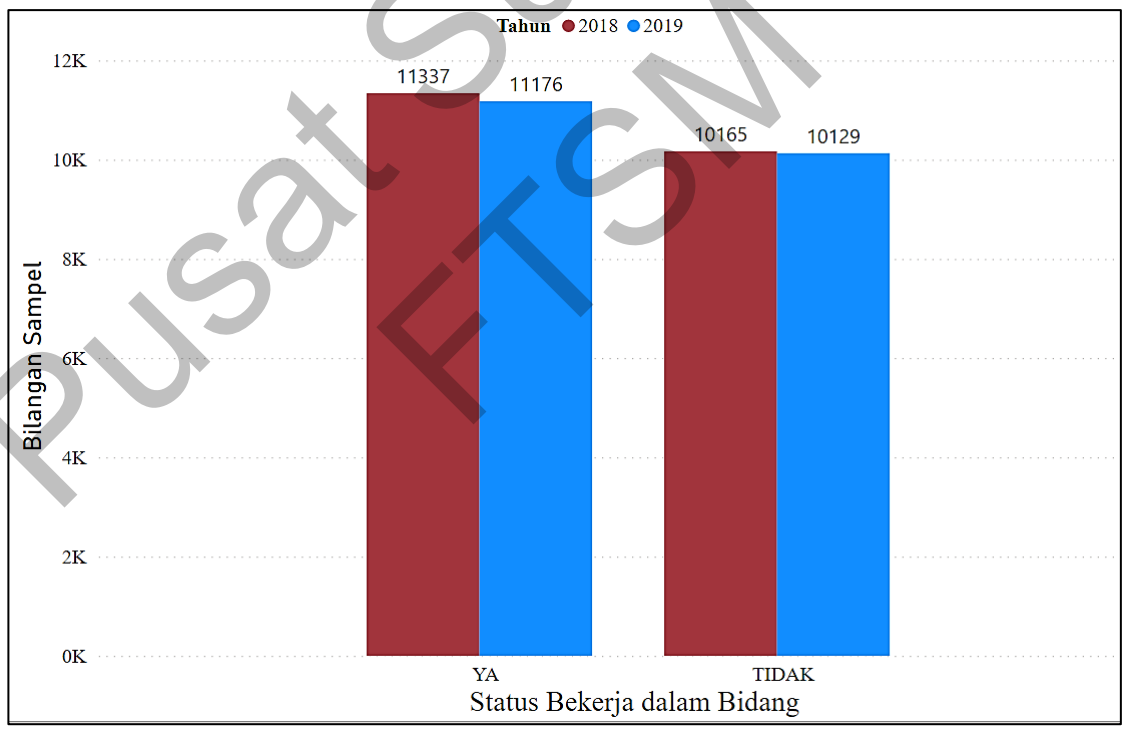
Bab ini menerangkan tentang dapatan kajian yang dilaksanakan terhadap data yang diperolehi daripada Sistem Kajian Pengesanan Graduan – TVET (SKPG-TVET) yang telah dibangunkan dan diguna pakai oleh KPT bermula pada tahun 2018 bagi menghasilkan kajian dan laporan kebolehpasaran graduan TVET yang merangkumi semua kementerian penyedia TVET. Dapatan kajian ini merangkumi analisis berkaitan kebolehpasaran graduan TVET samada ada bekerja dalam bidang atau luar bidang berdasarkan objektif dan metodologi kajian yang telah digariskan. Analisis dapatan kajian tersebut merangkumi analisis deskriptif, analisis pemilihan fitur dan analisis hasil model pengelasan yang telah dilaksanakan.

4.2 ANALSIS DESKRIPTIF DATA SKPG-TVET

Dalam kajian ini, analisis deskriptif set data SKPG-TVET dilaksanakan sebelum proses pembersihan dan pemodelan data dijalankan. Analisis ini dilaksanakan bertujuan untuk melihat dan memahami taburan sampel data yang digunakan dalam kajian ini. Data-data ini dianalisis secara visual melalui kaedah perwakilan data dengan menggunakan sampel data yang diperolehi daripada sistem SKPG-TVET bagi tahun 2018 dan 2019. Seperti yang ditunjukkan pada Rajah 4.1, jumlah sampel data yang digunakan dalam kajian ini adalah sebanyak 42,807 baris sampel data yang melibatkan 21,502 (50.23%) merupakan data bagi tahun 2018 dan 21,305 (49.77%) adalah data bagi tahun 2019.



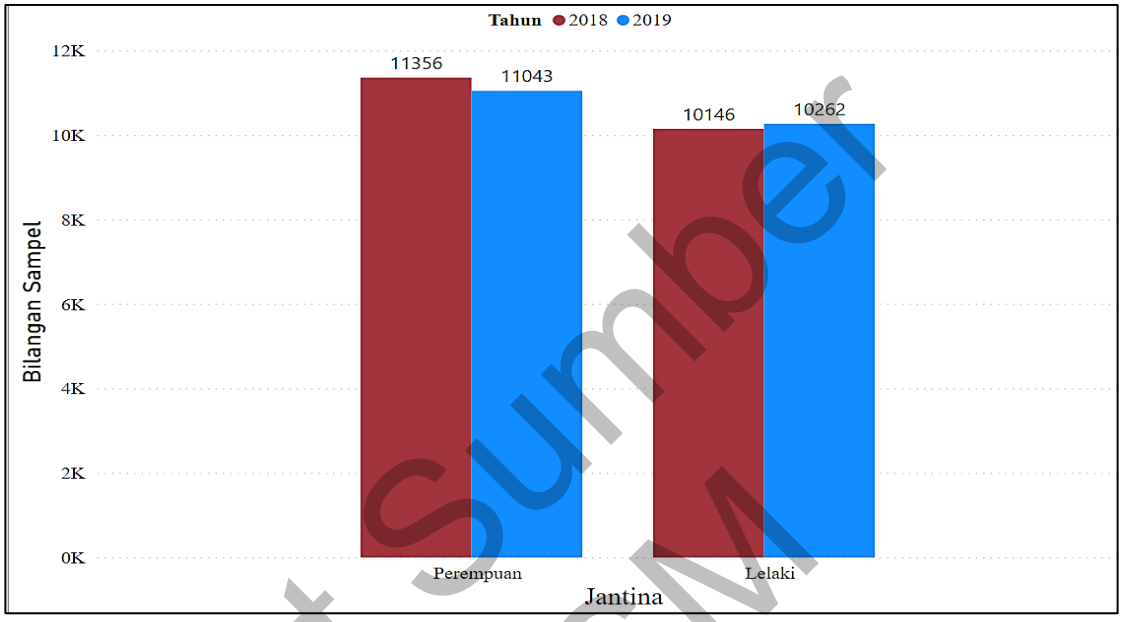
Rajah 4.1 Bilangan Sampel Data Kajian bagi Tahun 2018 dan 2019



Rajah 4.2 Bilangan Sampel Data Kajian Mengikut Kelas Kajian

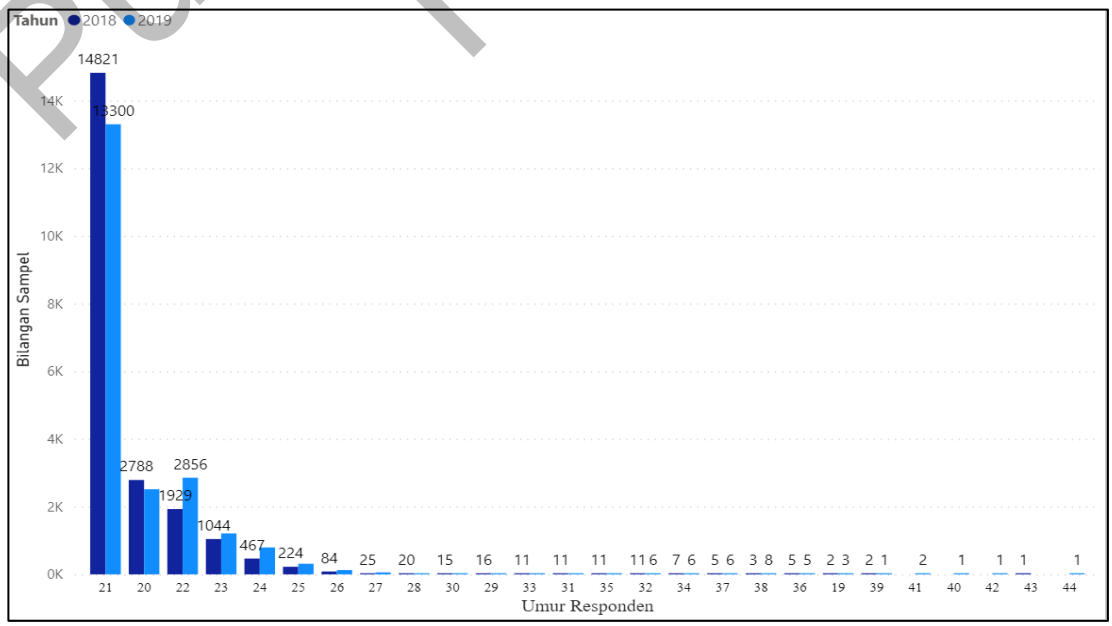
Berdasarkan Rajah 4.2, pembahagian kelas kajian adalah seimbang bagi 2 set data yang digunakan dalam kajian ini dimana bagi tahun 2018, sejumlah 11,337 (52.73%) sampel kelas adalah bekerja dalam bidang dan 10,165 (47.27%) sampel kelas adalah bekerja diluar bidang. Manakala bagi tahun 2019, sejumlah 11,176 (52.46%)

sampel kelas adalah bekerja dalam bidang dan 10,129 (47.54%) adalah sebaliknya. Perbezaan yang kecil iaitu 0.27 % diantara sampel kelas bagi tahun 2018 dan tahun 2019 telah menunjukkan bahawa data tersebut terdapat persamaan permasalahan yang wujud setiap tahun berkaitan status pekerjaan graduan TVET samada bekerja dalam atau luar bidang.



Rajah 4.3 Bilangan Sampel Data Kajian Mengikut Jantina Responden

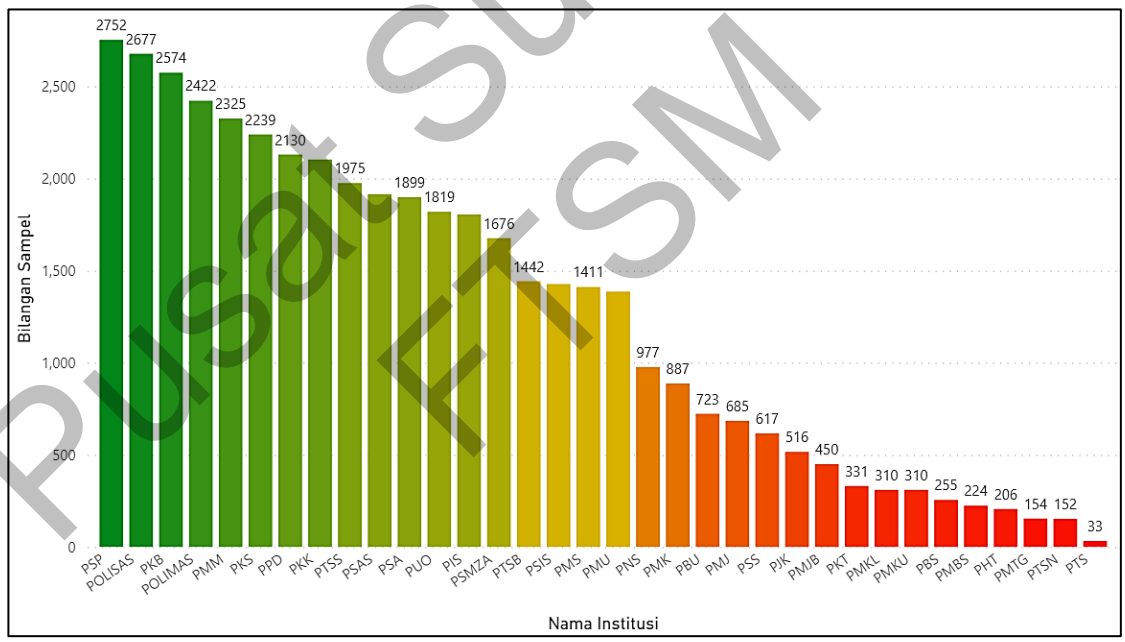
Manakala Rajah 4.3 menunjukkan jantina responden mengikut data tahun kajian. Berdasarkan rajah 4.3 dapat dilihat bahawa bilangan sampel perempuan adalah sedikit melebih sampel lelaki 5.63% bagi tahun 2018 dan 3.67% bagi tahun 2019.



Rajah 4.4 Bilangan Sampel Data Kajian Mengikut Umur bagi Tahun 2018 dan 2019

Berdasarkan kepada Rajah 4.4, majoriti umur responden adalah 21 tahun. Ini disebabkan oleh, data bagi sampel kajian ini merupakan graduan politeknik seluruh Malaysia yang diperolehi daripada Jabatan Pendidikan Politeknik dan Kolej Komuniti. Selari dengan matlamat penubuhan politeknik iaitu menawarkan program pengajian di peringkat sarjana muda, diploma dan sijil yang mana sebahagian besar pelajar lepasan SPM yang melanjutkan pengajian di politeknik akan mengambil ditawarkan program pengajian di peringkat diploma.

Justeru, sebahagian besar graduan politeknik adalah berumur 21 tahun iaitu graduan diploma. Berdasarkan data yang ditunjukkan dalam Rajah 4.4 juga mendapati bahawa wujudnya data terasing bagi atribut umur iaitu terdapat sebilangan kecil responden berumur lebih daripada 30 tahun dan berumur 19 tahun. Data-data ini boleh mempengaruhi keputusan model yang dibangunkan.



Rajah 4.5 Bilangan Sampel Data Kajian Mengikut Politeknik

Manakala Rajah 4.5 pula menunjukkan taburan data mengikut politeknik. Berdasarkan Rajah 4.5 terdapat perbezaan yang ketara antara jumlah sampel data daripada Politeknik Seberang Perai (PSP), Politeknik Sultan Haji Ahmad Shah (POLISAS), Politeknik Kota Bharu (PKB), Politeknik Sultan Abdul Halim Muadzam Shah (POLIMAS) dengan politeknik lain terutama Politeknik Tawau Sabah (PTS) dimana perbezaan jumlah sampel data bagi politeknik tersebut dengan PTS adalah amat

besar. Ini disebabkan oleh PTS merupakan politeknik yang baru beroperasi bermula tahun 2017 dengan jumlah ambilan pelajar yang sedikit.

4.3 PEMODELAN PENGELASAN

4.3.1 Pemodelan Pengelasan Model Tunggal

Model pengelasan dibangunkan berdasarkan kepada perancangan yang telah digariskan di dalam Bab 3. Model-model pengelasan dibangunkan dengan menggunakan atribut-atribut yang telah dipilih hasil daripada proses pemilihan fitur yang telah dilaksanakan. Model-model pengelasan ini dibangunkan menggunakan set data latihan dan pengujian yang sama bagi ketujuh-tujuh algoritma pembelajaran mesin yang telah ditentukan dalam kajian ini berdasarkan kajian kesusasteraan yang telah dilaksanakan. Tujuh (7) algoritma pembelajaran mesin yang digunakan dalam kajian ini adalah LR, NB, SVM, MLP, RF, DT dan KNN. Seterusnya hasil keputusan bagi setiap model dibuat perbandingan bagi memilih lima (5) model yang menghasilkan keputusan terbaik dipilih bagi tujuan penalaan parameter untuk meningkatkan lagi keputusan uji kaji. Hasil keputusan dan sisihan piawai bagi uji kaji ke atas setiap model pengelasan yang telah dilaksanakan boleh dilihat pada Jadual 4.1 sehingga Jadual 4.7.

Jadual 4.1 Prestasi Algoritma LR

Algoritma Regresi Logistik (LR)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
2	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
3	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
4	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
5	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
6	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
7	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
8	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
9	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
10	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
Purata	**69.50 (0.003328)**	**0.71 (0.009587)**	**0.71 (0.011379)**	**0.71 (0.004910)**

SD = Sisihan Piawai

Jadual 4.2 Prestasi Algoritma NB

Algoritma Naive Bayes (NB)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
2	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
3	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
4	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
5	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
6	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
7	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
8	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
9	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
10	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
Purata	**69.26 (0.004954)**	**0.73 (0.009704)**	**0.66 (0.010299)**	**0.69 (0.005326)**

SD = Sisihan Piawai

Jadual 4.3 Prestasi Algoritma SVM

Algoritma Mesin Sokongan Vektor (SVM)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
2	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
3	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
4	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
5	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
6	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
7	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
8	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
9	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
10	58.80 (0.009828)	0.58 (0.010485)	0.79 (0.010372)	0.67 (0.009213)
Purata	**58.80 (0.009828)**	**0.58 (0.010485)**	**0.79 (0.010372)**	**0.67 (0.009213)**

SD = Sisihan Piawai

Jadual 4.4 Prestasi Algoritma MLP

Algoritma *Multilayer Perceptron* (MLP)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	66.27 (0.044773)	0.68 (0.078202)	0.72 (0.152947)	0.66 (0.103202)
2	64.78 (0.052500)	0.70 (0.081742)	0.70 (0.130683)	0.69 (0.058114)
3	68.15 (0.023604)	0.68 (0.102123)	0.69 (0.185580)	0.66 (0.057384)
4	67.96 (0.018840)	0.67 (0.064901)	0.79 (0.134833)	0.71 (0.026372)
5	66.88 (0.031202)	0.67 (0.058066)	0.77 (0.126139)	0.68 (0.068767)
6	66.46 (0.041171)	0.72 (0.073726)	0.67 (0.126529)	0.70 (0.031944)
7	67.32 (0.042989)	0.72 (0.088798)	0.70 (0.171987)	0.70 (0.059330)
8	65.11 (0.045325)	0.65 (0.081366)	0.79 (0.143399)	0.69 (0.056500)
9	67.90 (0.031627)	0.68 (0.061450)	0.66 (0.191987)	0.64 (0.098664)
10	67.66 (0.033309)	0.66 (0.061489)	0.72 (0.131825)	0.68 (0.059721)
Purata	**66.85 (0.036534)**	**0.68 (0.075186)**	**0.72 (0.149591)**	**0.68 (0.062000)**

SD = Sisihan Piawai

Jadual 4.5 Prestasi Algoritma RF

Algoritma Hutan Rawak (RF)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	74.43 (0.004695)	0.75 (0.008402)	0.77 (0.015137)	0.76 (0.006614)
2	74.46 (0.004637)	0.75 (0.008665)	0.77 (0.012144)	0.76 (0.006244)
3	74.47 (0.004534)	0.75 (0.008185)	0.77 (0.012076)	0.76 (0.006669)
4	74.52 (0.003342)	0.75 (0.009667)	0.77 (0.012414)	0.76 (0.005575)
5	74.52 (0.004382)	0.75 (0.008411)	0.77 (0.012512)	0.76 (0.005526)
6	74.47 (0.004325)	0.75 (0.009630)	0.77 (0.011921)	0.76 (0.006169)
7	74.52 (0.003890)	0.75 (0.009068)	0.77 (0.010200)	0.76 (0.005278)
8	74.48 (0.003554)	0.75 (0.007711)	0.77 (0.014015)	0.76 (0.006650)
9	74.61 (0.003287)	0.75 (0.009339)	0.77 (0.011749)	0.76 (0.004705)
10	74.50 (0.005817)	0.75 (0.009181)	0.77 (0.012576)	0.76 (0.006107)
Purata	**74.50 (0.004246)**	**0.75 (0.008826)**	**0.77 (0.012474)**	**0.76 (0.005954)**

SD = Sisihan Piawai

Jadual 4.6 Prestasi Algoritma DT

Algoritma Pokok Keputusan (DT)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	72.00 (0.004511)	0.72 (0.008796)	0.77 (0.010913)	0.74 (0.006605)
2	72.12 (0.005496)	0.72 (0.007885)	0.77 (0.008228)	0.74 (0.006336)
3	72.05 (0.005122)	0.72 (0.008341)	0.77 (0.010580)	0.74 (0.006262)
4	72.10 (0.005788)	0.72 (0.007823)	0.77 (0.009432)	0.74 (0.005579)
5	72.06 (0.005442)	0.72 (0.008308)	0.77 (0.009717)	0.74 (0.006034)
6	72.02 (0.004890)	0.72 (0.009213)	0.77 (0.011045)	0.74 (0.005981)
7	72.11 (0.005799)	0.72 (0.007886)	0.77 (0.010777)	0.74 (0.004898)
8	72.06 (0.005384)	0.72 (0.009001)	0.77 (0.010616)	0.74 (0.005901)
9	72.03 (0.005723)	0.72 (0.007886)	0.77 (0.010274)	0.74 (0.005745)
10	72.03 (0.004253)	0.72 (0.008544)	0.77 (0.010196)	0.74 (0.006443)
Purata	**72.06 (0.005241)**	**0.72 (0.008368)**	**0.77 (0.010178)**	**0.74 (0.005978)**

SD = Sisihan Piawai

Jadual 4.7 Prestasi Algoritma KNN

Algoritma K-Nearest Neighbors (KNN)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
2	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
3	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
4	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
5	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
6	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
7	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
8	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
9	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
10	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
Purata	**74.96 (0.004743)**	**0.76 (0.009225)**	**0.77 (0.008063)**	**0.76 (0.005037)**

SD = Sisihan Piawai

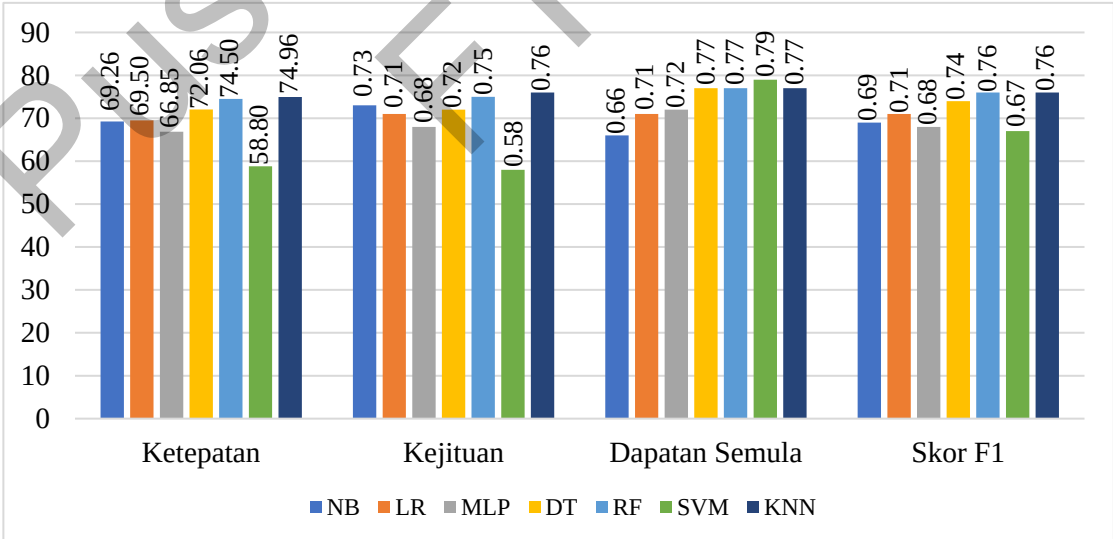
Jadual 4.1 sehingga Jadual 4.7 menunjukkan keputusan bagi purata setiap bilangan uji kaji yang telah dilaksanakan menggunakan kaedah pengesahan silang (10-lipatan). Bacaan setiap uji kaji diambil secara purata bagi nilai ketepatan, kejituan, dapatan semula dan *F-measure* bagi melihat kebolehan ketujuh-tujuh model untuk

meramal pengelasan hasil positif dan negatif yang merujuk kepada objektif kajian iaitu untuk meramal graduan samada bekerja dalam bidang atau luar bidang. Nilai purata bagi kesemua model bagi setiap algoritma dijadikan asas untuk proses penambahbaikan prestasi model melalui kaedah penalaan parameter. 5 algoritma yang menghasilkan prestasi yang terbaik dipilih untuk dilaksanakan proses penalaan parameter. 5 algoritma yang terpilih tersebut adalah Hutan Rawak (RF), Pokok Keputusan (DT), *K-Nearest Neighbors* (KNN), Regresi Logistik (LR) dan Naive Bayes (NB) berdasarkan analisis prestasi seperti yang ditunjukkan di dalam Jadual 4.8.

Jadual 4.8 Analisis Prestasi Awal Algoritma

Al	AC	SAC	P	SP	R	SR	F	SF	JS	K
NB	69.26	3	0.73	5	0.66	1	0.69	3	12	**5**
LR	69.50	4	0.71	3	0.71	2	0.71	4	13	**4**
MLP	66.85	2	0.68	2	0.72	3	0.68	2	9	**7**
DT	72.06	5	0.72	4	0.77	5	0.74	5	19	**3**
RF	74.50	6	0.75	6	0.77	5	0.76	6.5	23.5	**1**
SVM	58.80	1	0.58	1	0.79	7	0.67	1	10	**6**
KNN	74.96	7	0.76	7	0.77	5	0.76	3.5	22.5	**2**

Petunjuk: Al = Algoritma, AC= Ketepatan, P= Kejituan, R= Dapatan Semula, F= Skor F-measure, JS= Jumlah Skor, K= Kedudukan Prestasi, SAC= Skor Ketepatan, SP= Skor Kejituan, SR= Skor Dapatan Semula, SF= Skor F-measure



Rajah 4.6 Perbandingan Prestasi Awal Algoritma

Berdasarkan Jadual 4.8 dan Rajah 4.6, penilaian kedudukan prestasi dilaksanakan dengan menilai jumlah skor yang diperolehi oleh setiap model dengan

mengambil kira semua skor bagi setiap penilaian terhadap model iaitu skor bagi ketepatan, kejituan, dapatan semula dan *F-measure*. Hasil penilaian mendapati RF memperolehi jumlah skor yang tertinggi diikuti oleh KNN kedudukan kedua, DT tempat ketiga, LR kedudukan keempat dan NB kedudukan kelima. Hasil pemodelan awal juga mendapati MLP mendapat skor yang terendah dikalangan kesemua algoritma pembelajaran mesin bagi kajian ini.

4.3.2 Pemodelan Pengelasan Tunggal (Penalaan Parameter)

Lima (5) algoritma telah dipilih untuk dilaksanakan proses penalaan parameter bagi meningkatkan prestasi model tersebut. Kaedah penalaan telah dilaksanakan dengan menggunakan kaedah carian grid (*GridSearchCV*) dimana algoritma tersebut akan mengenal pasti parameter yang paling sesuai bagi setiap algoritma bagi menambah baik dan meningkatkan keputusan uji kaji. Keputusan penalaan parameter menggunakan kaedah carian grid ditunjukkan di Jadual 4.9 manakala keputusan hasil proses penalaan yang telah dilaksanakan bagi kelima-lima model tersebut ditunjukkan dalam Jadual 4.10 hingga Jadual 4.14.

Jadual 4.9 Keputusan Penalaan Parameter Menggunakan Kaedah Carian Grid

BIl	**Model**	**Parameter Penalaan**
1	NB	var_smoothing=1.232846739442066e-07
2	LR	C=10000, penalty='l2'
3	RF	n_estimators=1400,bootstrap=True,max_depth=100,min_samples_split=10, min_samples_leaf=3, max_features='sqrt'
4	KNN	n_neighbors=17, leaf_size=14, metric='manhattan',p=1, weights='uniform'
5	DT	max_depth=9, criterion='entropy', min_samples_split=9, min_samples_leaf=2

Jadual 4.10 Prestasi Penalaan Parameter Algoritma RF

Algoritma Hutan Rawak (RF)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	76.94 (0.005021)	0.77 (0.008835)	0.80 (0.007654)	0.78 (0.005781)
2	76.97 (0.004795)	0.77 (0.008700)	0.80 (0.008511)	0.78 (0.005051)
3	76.96 (0.005177)	0.77 (0.008549)	0.79 (0.008877)	0.78 (0.005199)
4	76.92 (0.004793)	0.77 (0.008418)	0.79 (0.008889)	0.78 (0.005898)
5	76.97 (0.005325)	0.77 (0.008550)	0.80 (0.009101)	0.78 (0.005201)
6	76.95 (0.004618)	0.77 (0.008321)	0.80 (0.008860)	0.78 (0.005401)
7	76.93 (0.005010)	0.77 (0.008365)	0.79 (0.008629)	0.78 (0.005125)
8	76.96 (0.005585)	0.77 (0.008549)	0.79 (0.008455)	0.78 (0.005978)
9	77.00 (0.004728)	0.77 (0.007890)	0.80 (0.008774)	0.78 (0.005465)
10	76.97 (0.005577)	0.77 (0.008378)	0.79 (0.008763)	0.78 (0.005995)
Purata	**76.96 (0.005063)**	**0.77 (0.008456)**	**0.80 (0.008651)**	**0.78 (0.005509)**

SD = Sisihan Piawai

Jadual 4.11 Prestasi Penalaan Parameter Algoritma KNN

Algoritma K-Nearest Neighbors (KNN)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
2	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
3	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
4	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
5	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
6	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
7	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
8	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
9	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
10	75.80 (0.004534)	0.76 (0.009846)	0.79 (0.009366)	0.78 (0.005982)
Purata	**75.80 (0.004534)**	**0.76 (0.009846)**	**0.79 (0.009366)**	**0.78 (0.005982)**

SD = Sisihan Piawai

Jadual 4.12 Prestasi Penalaan Parameter Algoritma DT

Algoritma Pokok Keputusan (DT)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	73.71 (0.005494)	0.74 (0.011454)	0.77 (0.015172)	0.75 (0.005278)
2	73.67 (0.005786)	0.74 (0.011445)	0.77 (0.015011)	0.75 (0.005270)
3	73.69 (0.005626)	0.74 (0.011653)	0.77 (0.015156)	0.75 (0.005467)
4	73.67 (0.005808)	0.74 (0.011560)	0.77 (0.014974)	0.75 (0.005114)
5	73.71 (0.005375)	0.74 (0.011698)	0.77 (0.015449)	0.75 (0.005362)
6	73.69 (0.005489)	0.74 (0.011316)	0.77 (0.015138)	0.75 (0.005226)
7	73.68 (0.005757)	0.74 (0.011375)	0.77 (0.015413)	0.75 (0.005145)
8	73.68 (0.005772)	0.74 (0.011550)	0.77 (0.014837)	0.75 (0.005109)
9	73.69 (0.005486)	0.74 (0.011463)	0.77 (0.015090)	0.75 (0.005388)
10	73.68 (0.005691)	0.74 (0.011452)	0.77 (0.014914)	0.75 (0.005301)
Purata	**73.69 (0.005628)**	**0.74 (0.011497)**	**0.77 (0.015115)**	**0.75 (0.005266)**

SD = Sisihan Piawai

Jadual 4.13 Prestasi Penalaan Parameter Algoritma LR

Algoritma Regresi Logistik (LR)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
2	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
3	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
4	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
5	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
6	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
7	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
8	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
9	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
10	69.55 (0.002522)	0.71 (0.008664)	0.71 (0.009735)	0.71 (0.004254)
Purata	**69.55 (0.002522)**	**0.71 (0.008664)**	**0.71 (0.009735)**	**0.71 (0.004254)**

SD = Sisihan Piawai

Jadual 4.14 Prestasi Penalaan Parameter Algoritma NB

Algoritma Naive Bayes (NB)				
Bil. Uji Kaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
2	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
3	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
4	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
5	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
6	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
7	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
8	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
9	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
10	69.98 (0.006632)	0.72 (0.011161)	0.71 (0.010658)	0.71 (0.008103)
Purata	**69.98 (0.006632)**	**0.72 (0.011161)**	**0.71 (0.010658)**	**0.71 (0.008103)**

SD = Sisihan Piawai

Jadual 4.15 Analisis Perbandingan Sebelum dan Selepas Penalaan Parameter Algoritma

Al	Status	AC (SD)	P(SD)	R(SD)	F(SD)
RF	Sebelum	74.50 (0.004246)	0.75 (0.008826)	0.77 (0.012474)	0.76 (0.005954)
	Selepas	**76.96 (0.005063)**	**0.77 (0.008456)**	**0.80 (0.008651)**	**0.78 (0.005509)**
KNN	Sebelum	74.96 (0.004743)	0.76 (0.009225)	0.77 (0.008063)	0.76 (0.005037)
	Selepas	**75.80 (0.004534)**	**0.76 (0.009846)**	**0.79 (0.009366)**	**0.78 (0.005982)**
DT	Sebelum	72.06 (0.005241)	0.72 (0.008368)	0.77 (0.010178)	0.74 (0.005978)
	Selepas	**73.69 (0.005628)**	**0.74 (0.011497)**	**0.77 (0.015115)**	**0.75 (0.005266)**
LR	Sebelum	69.50 (0.003328)	0.71 (0.009587)	0.71 (0.011379)	0.71 (0.004910)
	Selepas	**69.55 (0.002522)**	**0.71 (0.008664)**	**0.71 (0.009735)**	**0.71 (0.004254)**
NB	Sebelum	69.26 (0.004954)	0.73 (0.009704)	0.66 (0.010299)	0.69 (0.005326)
	Selepas	**69.98 (0.006632)**	**0.72 (0.011161)**	**0.71 (0.010658)**	**0.71 (0.008103)**

Petunjuk: Al = Algoritma, AC= Ketepatan, P= Kejituan, R= Dapatan Semula, F= *F-measure*,
SD = Sisihan Piawai

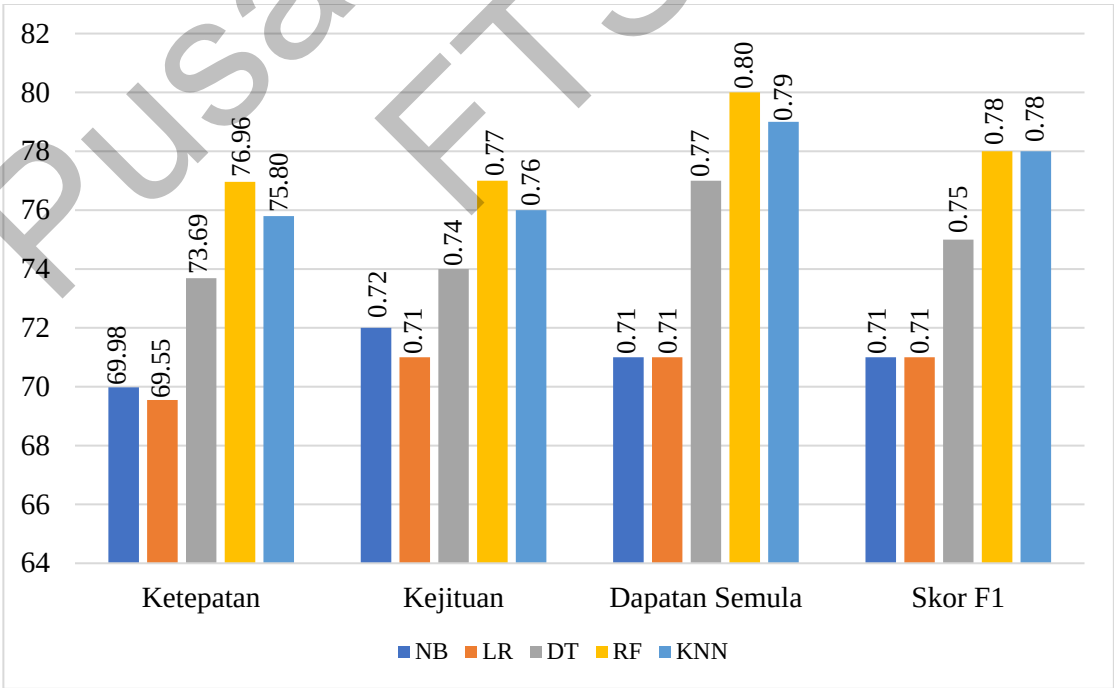
Jadual 4.15 menunjukkan analisis perbandingan hasil sebelum dan selepas proses penalaan paramater dilaksanakan bagi kelima-lima algoritma pembelajaran mesin menggunakan kaedah carian grid (*GridSearchCV*). Dapatan uji kaji mendapati hasil keputusan yang dihasilkan oleh algoritma yang telah dilaksanakan penalaan

paramater menunjukkan peningkatan dari segi keputusan bagi penilaia ketepatan, kejituan, dapatan semula dan *F-measure*. Kadar peningkatan keputusan ketepatan adalah diantara 0.05 peratus sehingga 2.46 peratus. Algoritma RF menunjukkan peningkatan ketepatan tertinggi iaitu sebanyak 2.46 peratus diikuti oleh DT sebanyak 1.63 peratus, KNN sebanyak 0.84 peratus, NB sebanyak 0.72 peratus dan LR sebanyak 0.05 peratus. Seterusnya penilaian prestasi bagi kelima-lima model dilaksanakan seperti yang ditunjukkan di Jadual 4.16.

Jadual 4.16 Analisis Prestasi Penalaan Parameter Algoritma

Al	AC	SAC	P	SP	R	S3	F	S4	JS	K
NB	69.98	2	0.72	2	0.71	1.5	0.71	1.5	7	**4**
LR	69.55	1	0.71	1	0.71	1.5	0.71	1.5	5	**5**
DT	73.69	3	0.74	3	0.77	3	0.75	3	12	**3**
RF	76.96	5	0.77	5	0.80	5	0.78	4.5	19.5	**1**
KNN	75.80	4	0.76	4	0.79	4	0.78	4.5	16.5	**2**

Petunjuk: Al = Algoritma, AC= Ketepatan, P= Kejituan, R= Dapatan Semula, F= Skor F-measure, JS= Jumlah Skor, K= Kedudukan Prestasi, SAC= Skor Ketepatan, SP= Skor Kejituan, SR= Skor Dapatan Semula, SF= Skor F-measure



Rajah 4.7 Perbandingan Prestasi Model Selepas Proses Penalaan Parameter

Jadual 4.16 dan Rajah 4.7 menunjukkan prestasi dan kedudukan setiap model yang telah dilaksanakan proses penalaan parameter. Dapatan telah menunjukkan bahawa selepas proses penalaan kedudukan prestasi bagi setiap algoritma tetap tidak berubah dimana RF tetap berada dikedudukan pertama dan diikuti oleh KNN, DT,NB dan LR. Justeru penalaan paramater adalah penting untuk dilaksanakan bagi meningkatkan hasil keputusan model-model pembelajaran mesin.

4.3.3 Pemodelan Pengelasan Model Penggabungan

Berdasarkan kedudukan yang telah ditunjukkan dalam Jadual 4.16, tiga (3) model yang berada di kedudukan tertinggi dipilih untuk dijalankan kajian menggunakan kaedah pembelajaran penggabungan model bagi meningkatkan lagi keputusan prestasi model tersebut. Model-model yang dipilih adalah RF, KNN dan DT sebagai model asas pembelajaran penggabungan. Antara kombinasi penggabungan model yang dilaksanakan adalah seperti di Jadual 4.17.

Jadual 4.17 Kombinasi Model Penggabungan

Model Ensembel	Kombinasi Model
1	RF+KNN+DT
2	RF+KNN+LR
3	RF+KNN+NB
4	RF+DT+LR
5	RF+DT+NB
6	KNN+DT+LR
7	KNN+DT+NB
8	RF+KNN
9	RF+DT
10	KNN+DT

Keputusan hasil penggabungan yang telah dilaksanakan bagi kelima-lima model tersebut ditunjukkan dalam Jadual 4.18 hingga Jadual 4.27.

Jadual 4.18 Prestasi Penggabungan Algoritma RF+KNN+DT

Algoritma RF+ KNN+DT				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	77.29 (0.004230)	0.77 (0.007323)	0.81 (0.008416)	0.79 (0.004872)
2	77.30 (0.004358)	0.77 (0.007923)	0.81 (0.008213)	0.79 (0.004847)
3	77.29 (0.004268)	0.77 (0.008087)	0.81 (0.008387)	0.79 (0.004773)
4	77.29 (0.004588)	0.77 (0.007862)	0.81 (0.008811)	0.79 (0.004973)
5	77.34 (0.004254)	0.77 (0.007642)	0.81 (0.008166)	0.79 (0.004740)
6	77.29 (0.004864)	0.77 (0.008116)	0.81 (0.008903)	0.79 (0.005002)
7	77.31 (0.004424)	0.77 (0.007737)	0.81 (0.008012)	0.79 (0.004552)
8	77.29 (0.004238)	0.77 (0.007414)	0.81 (0.008669)	0.79 (0.004763)
9	77.33 (0.004495)	0.77 (0.007776)	0.81 (0.008718)	0.79 (0.005240)
10	77.29 (0.005023)	0.77 (0.007860)	0.81 (0.008545)	0.79 (0.005333)
Purata	**77.30 (0.004474)**	**0.77 (0.007774)**	**0.81 (0.008484)**	**0.79 (0.004910)**

SD = Sisihan Piawai

Jadual 4.19 Prestasi Penggabungan Algoritma RF+KNN+LR

Algoritma RF+KNN+LR				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	77.10 (0.003656)	0.77 (0.008203)	0.81 (0.005074)	0.79 (0.004091)
2	77.14 (0.003695)	0.77 (0.008278)	0.81 (0.005323)	0.79 (0.004407)
3	77.16 (0.003549)	0.77 (0.008542)	0.81 (0.005300)	0.79 (0.004108)
4	77.10 (0.003092)	0.77 (0.008076)	0.80 (0.005979)	0.79 (0.004081)
5	77.10 (0.003709)	0.77 (0.007857)	0.80 (0.005896)	0.79 (0.004015)
6	77.07 (0.003645)	0.77 (0.007977)	0.81 (0.005454)	0.79 (0.004277)
7	77.10 (0.003568)	0.77 (0.008430)	0.81 (0.005957)	0.79 (0.004459)
8	77.10 (0.003548)	0.77 (0.008203)	0.81 (0.005051)	0.79 (0.004082)
9	77.13 (0.004171)	0.77 (0.008172)	0.81 (0.005427)	0.79 (0.003456)
10	77.09 (0.003439)	0.77 (0.008347)	0.81 (0.004995)	0.79 (0.004011)
Purata	**77.11 (0.003607)**	**0.77 (0.008209)**	**0.81 (0.005446)**	**0.79 (0.004099)**

SD = Sisihan Piawai

Jadual 4.20 Prestasi Penggabungan Algoritma RF+KNN+NB

Algoritma RF+KNN+NB				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	77.06 (0.004048)	0.77 (0.007856)	0.80 (0.005502)	0.79 (0.004072)
2	77.03 (0.004065)	0.77 (0.007415)	0.80 (0.005510)	0.79 (0.003490)
3	77.06 (0.003754)	0.77 (0.007460)	0.80 (0.006160)	0.79 (0.004438)
4	77.11 (0.003716)	0.77 (0.007600)	0.80 (0.005106)	0.79 (0.004056)
5	77.07 (0.003849)	0.77 (0.007184)	0.80 (0.004842)	0.79 (0.004046)
6	77.10 (0.003835)	0.77 (0.007479)	0.80 (0.006242)	0.79 (0.003194)
7	77.08 (0.003457)	0.77 (0.007407)	0.80 (0.005387)	0.79 (0.003956)
8	77.10 (0.003474)	0.77 (0.007557)	0.80 (0.005890)	0.79 (0.003674)
9	77.05 (0.003931)	0.77 (0.007615)	0.80 (0.005270)	0.79 (0.003828)
10	77.04 (0.003338)	0.77 (0.007528)	0.80 (0.005819)	0.79 (0.003744)
Purata	**77.07 (0.003747)**	**0.77 (0.007510)**	**0.80 (0.005573)**	**0.79 (0.003850)**

SD = Sisihan Piawai

Jadual 4.21 Prestasi Penggabungan Algoritma RF+DT+LR

Algoritma RF+DT+LR				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.26 (0.003916)	0.76 (0.007127)	0.78 (0.010793)	0.77 (0.005212)
2	75.28 (0.004347)	0.76 (0.006780)	0.78 (0.011264)	0.77 (0.004970)
3	75.28 (0.004606)	0.76 (0.007334)	0.78 (0.009916)	0.77 (0.004769)
4	75.29 (0.004486)	0.76 (0.007349)	0.78 (0.010794)	0.77 (0.005204)
5	75.26 (0.004442)	0.76 (0.007610)	0.78 (0.010298)	0.77 (0.005068)
6	75.25 (0.004331)	0.76 (0.007431)	0.78 (0.010581)	0.77 (0.004952)
7	75.28 (0.004626)	0.76 (0.006640)	0.78 (0.010259)	0.77 (0.005499)
8	75.30 (0.004403)	0.76 (0.007669)	0.78 (0.010152)	0.77 (0.005118)
9	75.30 (0.004686)	0.76 (0.007428)	0.78 (0.010431)	0.77 (0.005682)
10	75.28 (0.004633)	0.76 (0.007213)	0.78 (0.010506)	0.77 (0.005170)
Purata	**75.28 (0.004448)**	**0.76 (0.007258)**	**0.78 (0.010499)**	**0.77 (0.005164)**

SD = Sisihan Piawai

Jadual 4.22 Prestasi Penggabungan Algoritma RF+DT+NB

Algoritma RF+DT+NB				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.38 (0.003465)	0.76 (0.008153)	0.78 (0.009338)	0.77 (0.003363)
2	75.38 (0.003477)	0.76 (0.007646)	0.78 (0.009632)	0.77 (0.003411)
3	75.40 (0.003562)	0.76 (0.007723)	0.78 (0.008885)	0.77 (0.003222)
4	75.39 (0.003754)	0.76 (0.007685)	0.78 (0.009845)	0.77 (0.003642)
5	75.43 (0.003743)	0.76 (0.007896)	0.78 (0.009418)	0.77 (0.003461)
6	75.40 (0.003638)	0.76 (0.008039)	0.78 (0.009272)	0.77 (0.003825)
7	75.35 (0.003692)	0.76 (0.008371)	0.78 (0.009054)	0.77 (0.003894)
8	75.37 (0.003214)	0.76 (0.008346)	0.78 (0.010132)	0.77 (0.003746)
9	75.40 (0.002886)	0.76 (0.008274)	0.78 (0.010169)	0.77 (0.003768)
10	75.37 (0.003173)	0.76 (0.008360)	0.78 (0.009215)	0.77 (0.003475)
Purata	75.39 (0.003460)	0.76 (0.008049)	0.78 (0.009496)	0.77 (0.003581)

SD = Sisihan Piawai

Jadual 4.23 Prestasi Penggabungan Algoritma KNN+DT+LR

Algoritma KNN+DT+LR				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.51 (0.004289)	0.75 (0.007710)	0.79 (0.009864)	0.77 (0.005616)
2	75.53 (0.004283)	0.75 (0.007763)	0.79 (0.009864)	0.77 (0.005618)
3	75.51 (0.004336)	0.75 (0.007704)	0.79 (0.009603)	0.77 (0.005676)
4	75.52 (0.004363)	0.75 (0.007615)	0.79 (0.009661)	0.77 (0.005697)
5	75.52 (0.004239)	0.75 (0.007828)	0.79 (0.009603)	0.77 (0.005623)
6	75.52 (0.004346)	0.75 (0.007604)	0.79 (0.009603)	0.77 (0.005648)
7	75.51 (0.004352)	0.75 (0.007722)	0.79 (0.009661)	0.77 (0.005681)
8	75.50 (0.004363)	0.75 (0.007655)	0.79 (0.009603)	0.77 (0.005606)
9	75.51 (0.004307)	0.75 (0.007633)	0.79 (0.009603)	0.77 (0.005690)
10	75.51 (0.004300)	0.75 (0.007710)	0.79 (0.009603)	0.77 (0.005640)
Purata	**75.51 (0.004318)**	**0.75 (0.007694)**	**0.79 (0.009667)**	**0.77 (0.005650)**

SD = Sisihan Piawai

Jadual 4.24 Prestasi Penggabungan Algoritma KNN+DT+NB

Algoritma KNN+DT+NB				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.55 (0.002311)	0.76 (0.006402)	0.79 (0.009408)	0.77 (0.004079)
2	75.55 (0.002322)	0.76 (0.006496)	0.79 (0.009285)	0.77 (0.004130)
3	75.54 (0.002349)	0.76 (0.006536)	0.79 (0.009386)	0.77 (0.004013)
4	75.55 (0.002253)	0.76 (0.006441)	0.79 (0.009260)	0.77 (0.004086)
5	75.54 (0.002347)	0.75 (0.006369)	0.79 (0.009544)	0.77 (0.004073)
6	75.54 (0.002393)	0.76 (0.006321)	0.79 (0.009405)	0.77 (0.004106)
7	75.54 (0.002368)	0.76 (0.006453)	0.79 (0.009252)	0.77 (0.004053)
8	75.55 (0.002312)	0.76 (0.006572)	0.79 (0.009272)	0.77 (0.003974)
9	75.55 (0.002376)	0.76 (0.006472)	0.79 (0.009266)	0.77 (0.004063)
10	75.54 (0.002292)	0.76 (0.006474)	0.79 (0.009390)	0.77 (0.004124)
Purata	**75.55 (0.002332)**	**0.76 (0.006454)**	**0.79 (0.009347)**	**0.77 (0.004070)**

SD = Sisihan Piawai

Jadual 4.25 Prestasi Penggabungan Algoritma RF+ KNN

Algoritma RF+KNN				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	76.58 (0.005071)	0.73 (0.009443)	0.87 (0.007562)	0.80 (0.005841)
2	76.63 (0.004868)	0.73 (0.009775)	0.87 (0.006839)	0.80 (0.005716)
3	76.63 (0.005319)	0.73 (0.009805)	0.87 (0.007920)	0.80 (0.005393)
4	76.62 (0.005226)	0.73 (0.009379)	0.87 (0.007153)	0.80 (0.005941)
5	76.62 (0.005056)	0.73 (0.009879)	0.87 (0.007004)	0.80 (0.005662)
6	76.58 (0.005370)	0.73 (0.009635)	0.87 (0.006817)	0.80 (0.005692)
7	76.65 (0.004682)	0.73 (0.010040)	0.87 (0.007693)	0.80 (0.005835)
8	76.63 (0.005265)	0.73 (0.009947)	0.87 (0.006920)	0.80 (0.005644)
9	76.62 (0.005147)	0.73 (0.009688)	0.87 (0.007445)	0.80 (0.005739)
10	76.61 (0.005069)	0.73 (0.009779)	0.87 (0.007504)	0.80 (0.005673)
Purata	**76.62 (0.005107)**	**0.73 (0.009737)**	**0.87 (0.007286)**	**0.80 (0.005714)**

SD = Sisihan Piawai

Jadual 4.26 Prestasi Penggabungan Algoritma RF+ DT

Algoritma RF+DT				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	75.52 (0.007544)	0.73 (0.010489)	0.85 (0.008607)	0.78 (0.005940)
2	75.52 (0.006753)	0.73 (0.010786)	0.84 (0.007925)	0.78 (0.005713)
3	75.50 (0.007896)	0.73 (0.010284)	0.85 (0.007070)	0.78 (0.005357)
4	75.53 (0.007524)	0.73 (0.010308)	0.85 (0.007287)	0.78 (0.005813)
5	75.52 (0.007300)	0.73 (0.010397)	0.85 (0.008282)	0.78 (0.006179)
6	75.47 (0.006956)	0.73 (0.010733)	0.85 (0.007284)	0.78 (0.005933)
7	75.51 (0.007337)	0.73 (0.010830)	0.84 (0.007596)	0.78 (0.005298)
8	75.47 (0.007306)	0.73 (0.010905)	0.85 (0.007441)	0.78 (0.005832)
9	75.52 (0.007510)	0.73 (0.010789)	0.84 (0.007553)	0.78 (0.005214)
10	75.51 (0.007266)	0.73 (0.010761)	0.85 (0.007611)	0.78 (0.005794)
Purata	**75.51 (0.007339)**	**0.73 (0.010628)**	**0.85 (0.007666)**	**0.78 (0.005707)**

SD = Sisihan Piawai

Jadual 4.27 Prestasi Penggabungan Algoritma KNN+ DT

Algoritma KNN+DT				
Bil. Ujikaji	**Ketepatan (SD)**	**Kejituan (SD)**	**Dapatan Semula (SD)**	**F-measure (SD)**
1	74.96 (0.006275)	0.71 (0.010284)	0.89 (0.007905)	0.79 (0.005135)
2	74.95 (0.006351)	0.71 (0.010239)	0.89 (0.007541)	0.79 (0.005074)
3	74.96 (0.006341)	0.71 (0.010304)	0.88 (0.007565)	0.79 (0.005231)
4	74.95 (0.006246)	0.71 (0.010165)	0.89 (0.007819)	0.79 (0.005169)
5	74.98 (0.006099)	0.71 (0.010229)	0.89 (0.007541)	0.79 (0.005033)
6	74.96 (0.006266)	0.71 (0.010309)	0.89 (0.007819)	0.79 (0.004995)
7	74.97 (0.006130)	0.71 (0.010192)	0.89 (0.007508)	0.79 (0.005066)
8	74.95 (0.006277)	0.71 (0.010158)	0.88 (0.007593)	0.79 (0.005101)
9	74.96 (0.006332)	0.71 (0.010244)	0.89 (0.007905)	0.79 (0.005122)
10	74.95 (0.006344)	0.71 (0.010123)	0.89 (0.007541)	0.79 (0.005115)
Purata	**74.96 (0.006266)**	**0.71 (0.010225)**	**0.89 (0.007674)**	**0.79 (0.005104)**

SD = Sisihan Piawai

Hasil daripada penggabungan algoritma yang telah dijalankan menunjukkan bahawa penggabungan model dapat meningkatkan prestasi model yang dibangunkan. Sebagai contoh, model RF yang digabungkan dengan model KNN dan DT telah menunjukkan peningkatan prestasi berbanding model RF tunggal. Walaupun model RF

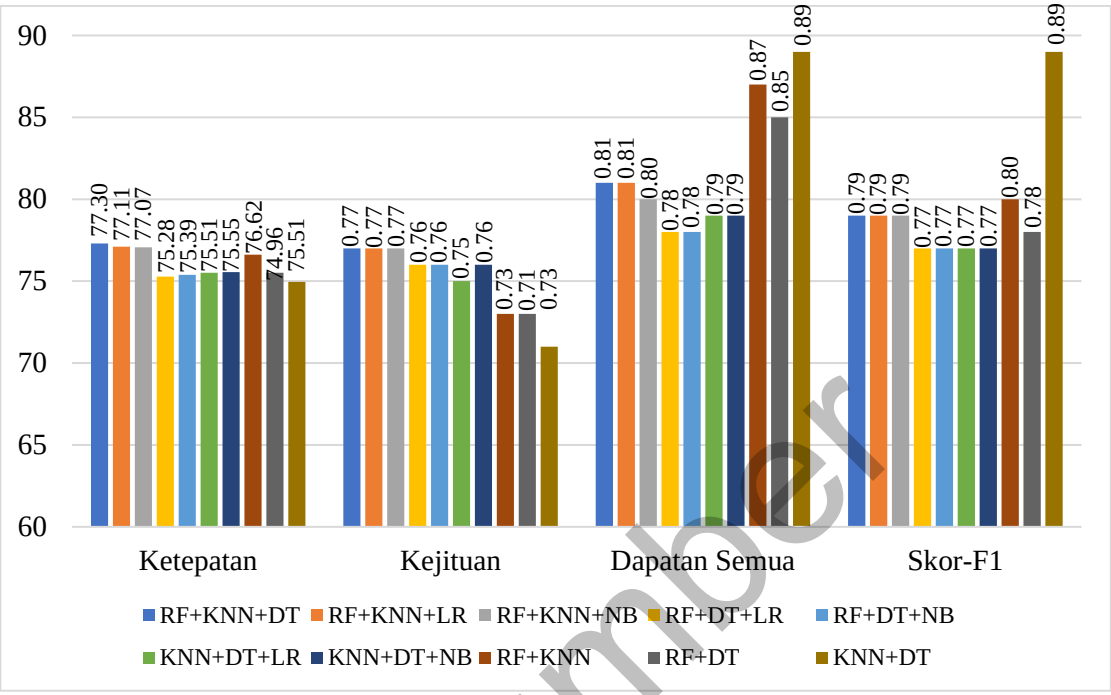
menunjukkan peningkatan prestasi apabila digabungkan dengan model lain, terdapat model yang menunjukkan penurunan prestasi apabila digabungkan seperti model KNN yang digabungkan dengan model DT dan LR yang telah menunjukkan penurunan prestasi ketepatan dari 75.80 peratus kepada 75.51 peratus.

Justeru, pemilihan model yang tepat bagi tujuan pembelajaran penggabungan memainkan peranan penting dalam memastikan prestasi model yang digabungkan dapat dipertingkatkan berbanding model tunggal. Analisis prestasi penggabungan algoritma dapat ditunjukkan dalam Jadual 4.28.

Jadual 4.28 Analisis Prestasi Model Penggabungan

Al	AC	SAC	P	SP	R	SR	F	SF	JS	K
RF+DT+KNN	77.30	10	0.77	9	0.81	6.5	0.79	7.5	33	**1**
RF+KNN +LR	77.11	9	0.77	9	0.81	6.5	0.79	7.5	32	**2**
RF+KNN+NB	77.07	8	0.77	9	0.80	5	0.79	7.5	29.5	**3**
RF+DT +LR	75.28	2	0.76	6	0.78	1.5	0.77	2.5	12	**10**
RF+DT+NB	75.39	3	0.76	6	0.78	1.5	0.77	2.5	13	**9**
KNN+DT+LR	75.51	4.5	0.75	4	0.79	3.5	0.77	2.5	14.5	**8**
KNN+DT+NB	75.55	6	0.76	6	0.79	3.5	0.77	2.5	18	**7**
RF+KNN	76.62	7	0.73	2.5	0.87	9	0.80	10	28.5	**4**
DT+KNN	74.96	1	0.71	1	0.89	10	0.79	7.5	19.5	**6**
RF+DT	75.51	4.5	0.73	2.5	0.85	8	0.78	5	20	**5**

Petunjuk: Al = Algoritma, AC= Ketepatan, P= Kejituan, R= Dapatan Semula, F= Skor F-measure, JS= Jumlah Skor, K= Kedudukan Prestasi, SAC= Skor Ketepatan, SP= Skor Kejituan, SR= Skor Dapatan Semula, SF= Skor F-measure



Rajah 4.8 Perbandingan Prestasi Penggabungan Model Pembelajaran Mesin

Berdasarkan Jadual 4.28 dan Rajah 4.8 telah menunjukkan model penggabungan RF+KNN+DT telah menghasil prestasi terbaik dengan jumlah skor terbesar iaitu 33 mata diikuti oleh model RF+KNN+LR dengan 32 mata dan RF+KNN+NB dengan 29.5 mata. Manakala model RF+DT+LR menunjukkan prestasi yang tercorot dengan jumlah skor sebanyak 12 mata. Walaubagaimanapun, secara amnya model-model yang dibangunkan dalam kajian ini menunjukkan hasil ketepatan yang bersifat sederhana iaitu dibawah 80 peratus. Namun terdapat model yang menghasilkan keputusan dapatan semula yang tinggi iaitu model KNN+DT dengan skor dapatan semula 0.89.

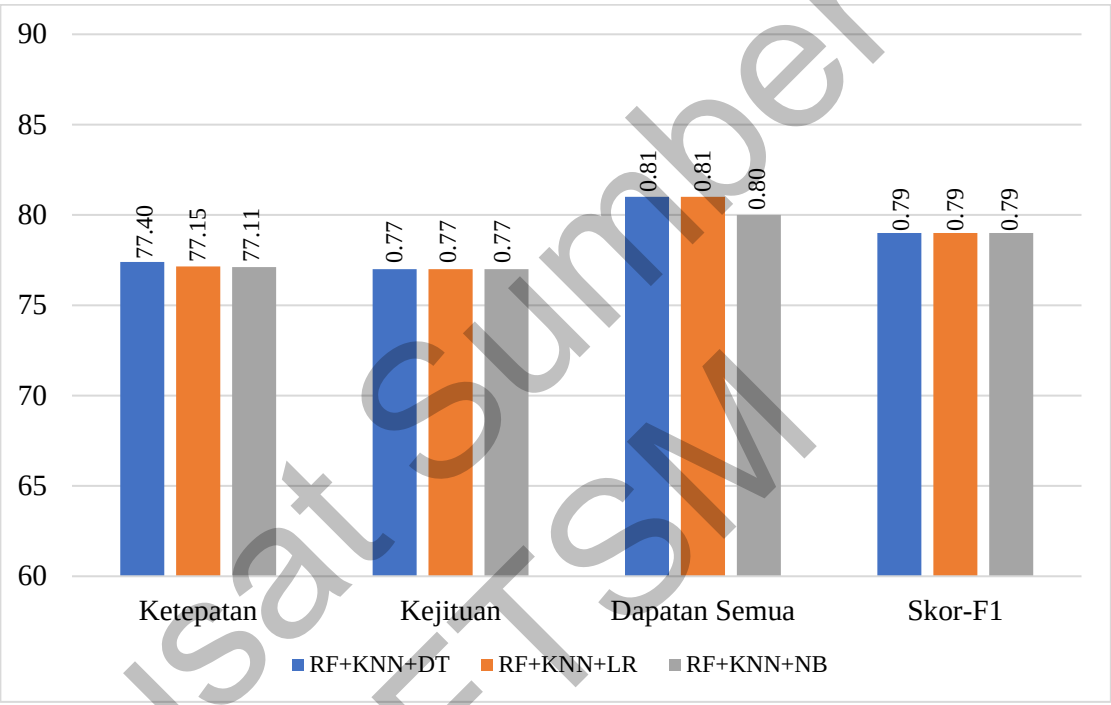
4.4 PENGUJIAN MODEL KE ATAS SET DATA UJIAN

Proses pengujian model telah dilaksanakan kepada tiga (3) model terbaik berdasarkan jumlah skor prestasi yang telah diperolehi. Model-model tersebut ialah model RF+DT+KNN, RF+KNN+LR dan RF+KNN+NB. Model-model tersebut diuji dengan menggunakan set data ujian yang telah disediakan pada awal proses pemodelan. Keputusan ujian ke atas set data ujian adalah seperti yang ditunjukkan dalam Jadual 4.29 dan Rajah 4.9.

Jadual 4.29 Analisis Prestasi Set Data Ujian

Al	AC	SAC	P	SP	R	SR	F	SF	JS	K
RF+DT+KNN	77.40	3	0.77	2	0.81	2.5	0.79	2	9.5	**1**
RF+KNN +LR	77.15	2	0.77	2	0.81	2.5	0.79	2	8.5	**2**
RF+KNN+NB	77.11	1	0.77	2	0.80	1	0.79	2	6	**3**

Petunjuk: Al = Algoritma, AC= Ketepatan, P= Kejituan, R= Dapatan Semula, F= Skor F-measure, JS= Jumlah Skor, K= Kedudukan Prestasi, SAC= Skor Ketepatan, SP= Skor Kejituan, SR= Skor Dapatan Semula, SF= Skor F-measure



Rajah 4.9 Analisis Prestasi Set Data Ujian

Berdasarkan Jadual 4.29 dan Rajah 4.9, hasil uji kaji terhadap set data ujian menunjukkan kedudukan prestasi model adalah selari dengan set data latihan dimana model RF+DT+KNN memperolehi jumlah skor tertinggi berbanding 2 model yang lain. Perbezaan ketepatan set data latihan dan ujian adalah 0.1 %. Ini menunjukkan model yang dibangunkan adalah stabil untuk menguji set data bagi kajian ini.

4.5 PENILAIAN MODEL KAJIAN

Proses penilaian model dalam kajian ini berdasarkan kepada ketepatan, kejituan, dapatan semula dan *F-measure* dimana berpandukan kepada objektif utama kajian iaitu mengelaskan graduan TVET samada bekerja di dalam bidang atau di luar bidang pengajian. Hasil kajian mendapati nilai tertinggi ketepatan model dalam kajian ini

adalah 77.40 peratus dengan menggabungkan beberapa model pembelajaran mesin. Nilai ketepatan yang sederhana ini diperolehi hasil daripada nilai kejituan yang sederhana jika dibandingkan dengan nilai dapatan semula bagi model-model yang telah dibangunkan. Nilai kejituan yang sederhana ini telah mengakibatkan ketepatan model berkurangan dimana model tidak dapat mengenal pasti graduan yang bekerja dalam bidang dengan baik tetapi perkara sebaliknya berlaku kepada graduan yang bekerja di luar bidang.

Ini mungkin disebabkan oleh kaedah pengumpulan dan pelabelan data bagi graduan yang bekerja di dalam bidang adalah kurang tepat sebaliknya pelabelan data bagi graduan yang bekerja luar bidang adalah lebih tepat. Ini dapat dilihat melalui dapatan kajian yang menunjukkan terdapat model yang memperolehi skor dapatan semula yang tinggi iaitu model RF+DT, RF+KNN dan KNN+DT dengan skor melebihi 0.85.

4.6 PENILAIAN MATRIKS KEKELIRUAN

Dalam bahagian ini, penilaian matriks kekeliruan dilaksanakan kepada tiga (3) model yang mendapat skor yang tertinggi iaitu model RF+KNN+DT, RF+KNN+LR dan RF+KNN+NB serta model-model tunggal yang sebelum dan selepas dilaksanakan penalaan parameter seperti di Jadual 4.30 dan 4.31.

Jadual 4.30 Matriks Kekeliruan RF, KNN, DT, LR dan NB

Matriks Kekeliruan

Model	**Sebelum Penalaan Parameter**			**Selepas Penalaan Parameter**		
		RAMALAN			**RAMALAN**	
		1	2		**1**	**2**
RF	**SEBENAR** 1	5269	1459	**SEBENAR** **1**	**5418**	**1310**
	SEBENAR 2	1655	4429	**SEBENAR** **2**	**1540**	**4544**

KNN

		RAMALAN	
		1	2
SEBENAR	1	5238	1490
	2	1658	4426

		RAMALAN	
		1	2
SEBENAR	1	5381	1347
	2	1677	4407

DT

		RAMALAN	
		1	2
SEBENAR	1	5274	1454
	2	2002	4082

		RAMALAN	
		1	2
SEBENAR	1	5338	1390
	2	1854	4230

LR

		RAMALAN	
		1	2
SEBENAR	1	4803	1925
	2	1927	4157

		RAMALAN	
		1	2
SEBENAR	1	4851	1877
	2	1951	4133

NB

		RAMALAN	
		1	2
SEBENAR	1	4390	2338
	2	1622	4462

		RAMALAN	
		1	2
SEBENAR	1	4779	1949
	2	1856	4228

Petunjuk: 1= Kelas Positif (Dalam Bidang), 2= Kelas Negatif (Luar Bidang)

Berdasarkan matriks kekeliruan yang ditunjukkan dalam Jadual 4.30, model yang telah melalui proses penalaan parameter telah menghasilkan keputusan ramalah yang lebih baik berbanding model yang tidak melalui proses penalaan parameter. Hasil dapatan juga mendapati bahawa model RF menghasilkan keputusan matriks kekeliruan yang terbaik berbanding model-model yang lain diikuti oleh model KNN dan DT.

Jadual 4.31 Matriks Kekeliruan RF+KNN+DT, RF+ KNN+LR dan RF+ KNN+NB

Model	**Matriks Kekeliruan**

RF+KNN+DT

		RAMALAN	
		1	**2**
SEBENAR	**1**	**5525**	**1203**
	2	**1614**	**4470**

RF+ KNN+LR

		RAMALAN	
		1	2
SEBENAR	1	5475	1253
	2	1579	4505

RF+ KNN+NB

		RAMALAN	
		1	2
SEBENAR	1	5438	1290
	2	1533	4551

Petunjuk: 1= Kelas Positif (Dalam Bidang), 2= Kelas Negatif (Luar Bidang)

Berdasarkan Jadual 4.31, model penggabungan RF+KNN+DT menghasilkan keputusan matriks kekeliruan yang terbaik diantara model-model penggabungan dan model-model tunggal yang telah dibangunkan. Tambahan pula, dapatan kajian mendapati bahawa semua model penggabungan menunjukkan peningkatan keputusan matriks kekeliruan dengan kaedah pemilihan penggabungan model yang sesuai.