Model Ramalan Dalam Kebolehpasaran Graduan UKM

Yong Kah Thean[1], Azuraliza Abu Bakar[2]

Centre for Artificial Intelligence Technology, Faculty of Information Science and Technology University Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, MALAYSIA

***Abstract*— Pada zaman ini, semakin ramai pelajar lulus pengajian dari universiti menyebabkan persaingan untuk mendapat pekerjaan antara mereka semakin sengit. Kebolehpasaran graduan adalah kebimbangan utama bagi institusi yang menawarkan pendidikan tinggi dan kaedah untuk ramalan awal kebolehpasaran pelajar selalu diinginkan untuk mengambil tindakan tepat pada masanya. Jadi, kajian ini dijalankan untuk menghasilkan satu model tentang kebolehpasaran graduan untuk meramal kebolehdapatan kerja graduan UKM. Perlombongan data adalah platform untuk mengekstrak pengetahuan tersembunyi dalam koleksi data. Oleh itu, perlombongan data merupakan teknik yang penting dalam kajian ini. Kajian ini menggunakan pendekatan perlombongan data yang terdiri daripada empat fasa. Fasa-fasa tersebut adalah pemahaman perniagaan, pengumpulan data, pra-pemprosesan data dan penyediaan data, dan pembangunan model. Kajian ini menggunakan 19792 data Graduan UKM pada tahun 2016 hingga 2018 yang dikumpul dari Sistem Kajian Pengesanan Graduan (SKPG) yang dikelolakan daripada Unit Kajian Pengesanan Graduan, Kementerian Pengajian Tinggi Malaysia. Objektif kajian ini adalah untuk mengurangkan jurang perbezaan pekerjaan yang ditawar dan bidang di universiti. Selain itu, kajian ini juga menggunakan teknik klasifikasi yang mengandungi teknik pengelasan iaitu kaedah Pohon Keputusan. Selain daripada model algoritma Pohon Keputusan (J48), kajian telah membangunkan model dengan menggunakan algoritma seperti Regresi Logistik, Naive Bayes dan Mesin Vektor Sokongan (SVM) untuk mengenal pastikan hubungan antara faktor-faktor yang akan mempengaruhi kegagalan mendapat kerja dan kebolehdapatan kerja. Keputusan analisis perbandingan menyimpulkan bahawa kaedah mesin vektor sokongan iaitu algoritma SMO dan regresi logistik adalah paling sesuai digunakan untuk membina model kebolehpasaran dengan ketepatan pengelasan.**

***Keywords*— Pohon Keputusan; Naïve Bayes; Mesin Vektor Sokongan; Regresi Logistik; kebolehpasaran; perlombongan data**

I. PENGENALAN

Kebolehpasaran graduan telah ditakrif secara umum. Kebolehpasaran graduan bermakna bahawa seorang siswazah dapat memperoleh pekerjaan dalam pelbagai bentuk. Dalam pendidikan tinggi, para graduan dapat mencari pekerjaan yang relevan dengan kelayakan mereka adalah sangat penting. Zaman kini, bilangan graduan semakin meningkat berbanding permintaan dari pasaran. Kadar pengangguran di kalangan graduan telah meningkat dari tahun ke tahun. Seramai 200,000 graduan daripada Universiti yang dikeluarkan setiap tahun dan bilangan graduan ini saya dapat cari dari laporan internet. Daripada penyelidikan yang saya dapat cari adalah Michelle, L (2016), 2 daripada 5 graduan tetap menganggur enam bulan selepas tamat pengajian dari Universiti dan majoriti sebagai pemegang ijazah. Ini sangat susah untuk mengetahui sejauh mana graduan sudah menyediakan diri untuk bekerja, walaupun kawasan ini tidak dikaji. Menurut Jabatan Statistik Malaysia, Malaysia: Youth Employment (2019), kadar pengangguran di Malaysia masih mengekalkan di 3.30% pada Jun dan Julai 2019. Kadar pengangguran ini berubah setiap bulan, daripada Januari hingga Julai 2019, dengan kadar purata 3.40%. Data ini mencapai tahap tertinggi iaitu 3.60% pada bulan November 2016 dan rekod terendah pada 3.20% pada Januari 2019.

Teknik perlombongan data adalah proses secara automatik meneroka pengetahuan berguna dalam repositori data yang besar. Tugas perlombongan data dibahagikan kepada dua kategori utama yang merupakan tugas ramalan (prediktif) dan tugas deskriptif. Teknik ini dapat mendedahkan bahawa meramalkan nilai sesuatu atribut tertentu bergantung pada nilai atribut yang lain adalah matlamat tugas ramalan. Klasifikasi adalah salah satu teknik perlombongan data yang telah menjadi menarik topik kepada penyelidik kerana ketepatan dan kecekapannya untuk mengklasifikasikan data untuk penemuan pengetahuan. Penyelidikan mengenai ramalan

prestasi pelajar Universiti ini dijalankan menggunakan empat model klasifikasi data seperti Pohon Keputusan, Regresi Logistik, Naive Bayes dan Mesin Vektor Sokongan (SVM). Data diperolehi melalui tinjauan di kalangan pelajar universiti yang dijalankan dalam tempoh 2016 hingga 2018. Teknik perlombongan data digunakan untuk mengenal pasti dan meramal data graduan kebolehpasaran supaya dapat mencari faktor yang mempengaruhi kebolehpasaran graduan. Kajian-kajian dijalankan adalah untuk memudahkan dan membantu graduan mempunyai peluang yang banyak mencari pekerjaan atau pekerjaan yang sesuai.

II. KAJIAN KESUSASTERAAN

Melalui kajian lepas selidik Mohd Tajul Rizal dan Yuhanis Yusof (2017) menerangkan kajian ini mengemukakan penerapan teknik perlombongan data dalam meramalkan jenis pekerjaan graduan KPM. Dalam perlombongan data, terdapat tiga tugas utama; pengelasan, pengelompokan dan pergaulan. Tujuan kajian ini adalah untuk meramalkan sama ada lulusan tertentu akan "diambil bekerja", "menganggur" atau "melanjutkan pelajaran" 6 bulan setelah tamat pengajiannya. Eksperimen yang dilakukan merangkumi penggunaan lima teknik perlombongan data iaitu Naive Bayes, Logistic Regression, multilayer perceptron, K-terdekat tetangga dan keputusan. Selanjutnya, penyediaan eksperimen berdasarkan tiga jenis perkadaran data (latihan-ujian) 70-30, 80-20 dan 90-10. Berdasarkan hasil yang diperoleh, dapat diketahui bahawa regresi logistik adalah pengelasan terbaik untuk set data di tangan.

Kajian lepas daripada Myzatul Akmam Sapaat (2011) membina graduan kebolehdapatan kerja model dengan menggunakan tugas klasifikasi dalam perlombongan data. Eksperimen klasifikasi dilakukan menggunakan pelbagai algoritma Bayes untuk menentukan sama ada bekerja, masih menganggur atau dalam keadaan yang tidak dapat ditentukan. Prestasi algoritma Bayes juga dibandingkan dengan beberapa algoritma berasaskan pokok. Keputusan menunjukkan bahawa J48, satu variasi algoritma pokok keputusan dilakukan dengan ketepatan tertinggi, iaitu 92.3% berbanding dengan purata 91.3% daripada algoritma Bayes yang lain. Ini membawa kepada kesimpulan bahawa pengeluar berasakan pokok lebih sesuai untuk data pengesan akibat strategi mendapatkan maklumat.

Satu lagi penyelidikan daripada Anatoli Nachev (2018) membentangkan kajian ini adalah untuk menganalisis data dari Quarterly Tinjauan Rumah Tangga Nasional (QNHS), berskala besar di seluruh negara tinjauan, dengan kaedah SVM. Ini mengatasi beberapa jurang dalam penyelidikan terdahulu, memberi tumpuan kepada mengukur faktor dan memberi pandangan mengenai peranan mereka dalam pekerjaan. Kami meneroka secara eksperimen bagaimana pilihan kernel dan parameter SVM nilai mempengaruhi model dan dibina dengan prestasi terbaik. Hasil kajian menunjukkan bahawa kernel RBF Gaussian dengan sigma = 0,051 dan C = 3.5 adalah yang terbaik untuk set data tersebut. Prestasi diukur oleh ketepatan ramalan dan AUC dari analisis ROC. Menganalisis kepentingan berubah untuk model, kami mengira faktor yang mempengaruhi status pekerjaan responden. Selepas menentukan faktor-faktor tersebut, kami memberi tumpuan kepada yang paling penting: kelas umur, pendidikan semasa dan tamat, sifat penghunian, kewarganegaraan, dan status perkahwinan. Maklumat lebih lanjut untuk setiap faktor diberikan oleh analisis VEC, yang menunjukkan bagaimana nilai faktor tersebut menyumbang kepada status pekerjaan. Sebagai kesimpulan, kami dapati SVM sebagai alat perlombongan data yang kuat, yang memungkinkan untuk menganalisis data mentah untuk memperoleh yang berharga pandangan dengan makna praktikal dalam domain pekerjaan.

Kajian daripada Kian Lam Tan (2019) menerangkan Teknik perlombongan data banyak digunakan di kejuruteraan, perubatan, industri, pertanian dan juga digunakan di pendidikan untuk meramalkan keadaan masa depan. Dalam makalah ini, yang digunakan teknik perlombongan data yang diterapkan dalam pemilihan ciri dan tentukan model terbaik yang boleh digunakan untuk meramalkan status pekerjaan Institusi Awam lulusan baru sama ada bekerja atau tidak bekerja, enam bulan selepas tamat pengajian. Dalam Metodologi CRISP-DM, enam fasa diadopsi. algoritma dalam pembelajaran yang diselia dan tidak diselia; K-Terdekat Jiran, Naive Bayes, Pohon Keputusan, Rangkaian Neural, Mesin Vektor Regresi Logistik dan Sokongan adalah dibandingkan dengan menggunakan set data latihan dari Tracer Study hingga menentukan ketepatan tertinggi pada gilirannya digunakan sebagai ramalan model. Rapid Miner sebagai alat perlombongan data digunakan untuk data algoritma analisis.

Dalam kajian ini akan menggunakan teknik klasifikasi yang mengandungi teknik pengelasan iaitu kaedah Pohon Keputusan. Kajian ini adalah untuk mencari faktor yang mempengaruhi kebolehpasaran kerja graduan UKM. Selain daripada model algoritma J48, kajian telah membangunkan model dengan menggunakan algoritma seperti Regresi Logistik, Naive Bayes dan SVM. Teknik klasifikasi yang dipilih untuk membangunkan model bagi kajian ini adalah yang akan membawa ketepatan yang tinggi dan selalu dipilih oleh orang yang membuat ramalan untuk

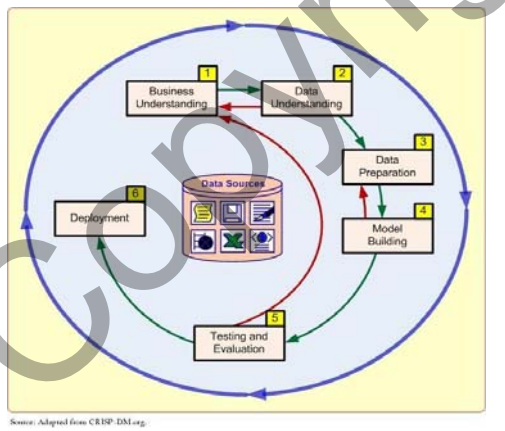
membangunkan model. Perbezaan antara kajian ini dengan kajian lepas adalah kajian ini hanya menentukan hasil bekerja atau tidak bekerja sahaja. Jadual 1.1 menunjukkan jadual beberapa kajian kebolehpasaran graduan dan teknik ramalan model.

Jadual 1.1: Perbandingan aplikasi ramalan yang berkaitan dengan kajian

Bilangan	Rujukan	Teknik Klasifikasi
1.	Mohd Tajul Rizal dan Yuhanis Yusof (2017)	Regresi Logistik
2.	Myzatul Akmam Sapaat (2011)	Pohon Keputusan (J48)
3.	Anatoli Nachev (2018)	Mesin Vektor Sokongan (SVM)
4.	Kian Lam Tan (2019)	Naive Bayes

III. KAEDAH KAJIAN

Metodologi kajian yang digunakan adalah dijalankan mengikuti kaedah CRISP-DM. Terdapat enam fasa dalam CRISP-DM untuk menjalankan kajian ini adalah pemahaman perniagaan, pemahaman data, penyediaan data, pembangunan model, penilaian model dan pengaplikasian. Pendekatan kajian ini berdasarkan kepada penyelidikan eksperimental iaitu berasaskan kepada enam fasa penyelidikan seperti Rajah 1.1. Setiap fasa berstruktur dan berdefinisi dengan jelas supaya mudah untuk diaplikasikan.



Rajah 1.1: Model Data Mining CRISP-DM

A. Pemahaman Perniagaan

Pemahaman perniagaan ini merupakan langkah pertama dalam mencari peraturan kesatuan dan melaksanakan teknik perlombongan data dalam menggunakan metodologi CRISP-DM. Sejak kajian ini adalah untuk mengenal pastikan hubungan antara faktor-faktor yang akan mempengaruhi kegagalan graduan mendapat kerja. Pakar UKM-karier diperlukan terlibat dalam kajian ini untuk menguji model dan keputusan yang didapati pada masa akan datang supaya menyebabkan peningkatan keyakinan terhadap kajian ini. Faktor-faktor yang menyebabkan graduan gagal memperoleh pekerjaan adalah sifat peribadi graduan yang mana kemahiran komunikasi, tahap, kerja berpasukan, kemahiran profesional dan membuat keputusan. Maklumat tersebut didapati dari jurnal. Graduan tidak menguasai kemahiran tersebut. Oleh itu, kajian ini akan dijalankan dengan menjelaskan faktor-faktor kebolehpasaran graduan dengan meminta graduan UKM mengisi borang.

Tugas utama memahami matlamat perniagaan untuk analisis data adalah di mana matlamat dan soalan perniagaan perlu ditakrifkan. Matlamat perniagaan kajian ini untuk membangunkan model ramalan untuk klasifikasi kebolehpasaran graduan di UKM. Kami perlu menyediakan unjuran karier di Malaysia untuk tempoh 5 tahun dan mengurangkan jurang perbezaan pekerjaan yang ditawar dan bidang di universiti untuk mencapai matlamat perniagaan kajian ini. Meningkatkan kebolehpasaran kerja graduan dan menentukan laluan kerjaya yang sesuai mengikut pencapaian akademik, kokurikulum, psikometrik, disiplin dan sahsiah mengikut pasaran semasa juga merupakan matlamat yang penting dalam perniagaan kajian ini. Soalan perniagaan kami adalah seperti berikut:

1. Apakah top 10 karier di Malaysia untuk 5 tahun akan datang?
2. Apakah hubungan demografi, akademik, sosial pelajar, program pengajian dengan pekerjaan?
3. Meramal top 5 kerjaya yang sesuai untuk setiap pelajar lepasan program pengajian (ijazah, diploma, sijil). Apakah kerjaya yang bersesuaian dengan individu.
    I. Lepasan SPM dan lepasan STPM setaraf.
    II. Lepasan sijil, lepasan diploma.
4. Mengenal pasti pelajar yang berisiko tidak mendapat pekerjaan selepas graduan.
5. Mengenal pasti pelajar yang berpeluang mendapat pekerjaan selepas graduan.

B. Pemahaman Data

Pemahaman data merupakan proses yang mengumpul data awal dari semua sumber yang kita ada. Kemudian, sifat-sifat data yang diperolehi perlu mengetahui data untuk mengenal pasti masalah kualiti data. Untuk mengetahui tahap pandangan awal dalam data atau mengesan sub set

menarik untuk membentuk hipotesis untuk maklumat yang tersembunyi.

Sifat-sifat data dikumpulkan daripada graduan UKM yang mengisi borang. Untuk menentukan status kebolehdapatan kerja terhadap graduan UKM, klasifikasi sifat-sifat data graduan akan dijalankan. Terdapat sebanyak 337 sifat-sifat dalam data set yang dikumpul. Atribut ini merangkumi nama, jantina, CGPA, umur dan sebagainya. Daripada data yang dikumpulkan, kita akan mengkategorikan profesion graduan ikuti fakulti untuk membanding mana satu sektor pekerjaan lebih memerlukan pekerja daripada sektor lain. Jadual 1.2 di bawah menunjukkan sifat-sifat kebolehpasaran graduan UKM dengan catat nilai dan penerangan. Melalui penyelidikan di Internet, penggunaan data dapat mengklasifikasikan sifat-sifat dari data pelajar dengan memastikan status pelajar sama ada dipilih bekerja, melanjutkan pelajaran, kemahiran menaik taraf, menunggu penempatan kerja atau menganggur.

Jadual 1.2 Sifat-sifat kebolehpasaran graduan UKM

No	Sifat-sifat	Nilai	Penerangan
1	Jantina	Lelaki, Perempuan	Jantina Lulusan
2	Umur	20-49 tahun	Umur
3	CGPA	{2.00-2.49, 2.50-2.99, 3.00-3.66, 3.67-4.00, failed}	CGPA untuk kelayakan semasa

Kajian ini menggunakan data Graduan UKM pada tahun 2016 hingga 2018 yang diperolehi dari Sistem Kajian Pengesanan Graduan (SKPG) yang dikelolakan daripada Unit Kajian Pengesanan Graduan, Kementerian Pengajian Tinggi Malaysia. Data ini mengandungi 3 tahun data graduan yang bermula daripada tahun 2016 hingga tahun 2018. Jumlah data pada tahun 2016 adalah 5889 respons, 7353 respons pada tahun 2017 dan 6550 respons pada tahun 2018. Terdapat sebanyak 19792 data yang merangkumi 358 atribut telah dikumpul. Keterangan data telah ditunjukkan dalam lampiran. Rajah 3.1 menunjukkan senarai keseluruhan atribut. (Soo,2017).

C. Penyediaan Data

Penyediaan data dan pra-pemprosesan data melibatkan empat proses utama iaitu, pembersihan data, pengurangan data, transformasi data dan pelabelan data. Tujuan fasa ini dilakukan untuk memastikan set data adalah bersih dan sedia digunakan dalam fasa perlombongan data. Fasa ini merupakan fasa yang penting sebelum proses perlombongan data dijalankan kerana fasa ini sangat berguna bagi memperolehi pengetahuan. Program Waikato Environment for Knowledge Analysis (Weka) telah digunakan dalam kajian ini kerana mudah diterapkan pada set data. Jenis fail yang digunakan dalam kajian ini untuk menganalisis data dalam Weka ialah jenis fail *Attribute-Relation File Format* (arff) dan data juga diimport dalam format seperti Comma-Separated Values (CSV).

Proses pembersihan data adalah langkah yang sangat penting dan terawal dalam proses pra-pemprosesan data dijalankan untuk mengurangkan atribut yang sama, mengendalikan nilai-nilai yang hilang dan data yang tidak konsisten dan mengenal pasti outliers. Bagi menangani kehilangan nilai, data tahun 2016 dan 2018 mempunyai kehilangan nilai manakala tahun 2017 tidak mempunyai kehilangan nilai. Jadi, kaedah penapis 'ReplaceMissingValues' akan digunakan untuk menggantikan nilai supaya tidak mempunyai kehilangan nilai. Kehilangan nilai akan digantikan dengan nilai 'mean'. Kaedah menyingkirkan outlier adalah untuk menghapuskan data yang tidak sama dengan data lain dan akan menjadi data tidak konsisten. Data untuk tahun 2016 tidak mempunyai data tidak konsisten manakala data untuk tahun 2017 dan tahun 2018 tidak mempunyai kehilangan data. Oleh itu, penapis '*InterquartileRange*' digunakan untuk mencari outlier dalam set data ini. Selepas mencari outlier, kaedah mengeluarkan outlier telah dijalankan dengan menggunakan penapis '*RemoveWithValue*'.

Langkah kedua dalam pra-pemprosesan data adalah pengurangan data. Dalam kajian ini, proses tersebut akan dijalankan untuk mengurangkan data yang lebih atau tidak membawa kesan yang besar terhadap kajian ini.

Transformasi data adalah proses yang penting dalam pra-pemprosesan data. Proses transformasi data digunakan untuk menyatukan data mengikut kelas yang diperlukan. Selain itu, proses tersebut juga boleh menukarkan format data ke format lain. Kaedah penyatu data adalah menukar daripada struktur kumpulan besar kepada struktur kumpulan kecil. Proses penukaran jenis atribut membantu bertukar jenis atribut daripada '*numeric*' kepada '*nominal*'. '*Nominal*' ialah satu jenis atribut digunakan untuk melabelkan data tanpa memberi sebarang nilai kuantitatif. Penapis '*NumericToNominal*' digunakan untuk menukar jenis atribut dalam kajian ini. Atribut jantina daripada data set 2017 telah menukar daripada '*numeric*' kepada '*nominal*' selepas menggunakan penapis '*NumericToNominal*'.

Proses yang terakhir dalam pra-pemprosesan data adalah pelabelan data. Proses ini adalah menukarkan label atribut seperti e_pendapatan kepada pendapatan keluarga untuk senang difahami. Selain daripada itu, label kumpulan juga boleh ditukarkan dengan menggunakan excel. Jadual akan menunjukkan sebelum dan selepas menukar label kumpulan masing-masing supaya lebih mudah difahami.

D. Pemodelan

Pemilihan teknik pemodelan adalah langkah pertama yang diambil diikuti oleh penjanaan senario ujian untuk mengesahkan kualiti model. Fasa pemodelan boleh dilakukan dengan menguji dan mencuba model teknik yang berbeza dengan ciri-ciri yang berbeza. Ujian ini boleh diulang dengan parameter perubahan sedikit, memantau keputusan dan membina beberapa kesimpulan awal pada model. Sebagai contoh, KNN, Naive Bayes, Pohon Keputusan, Rangkaian Neural, Regresi Logistik, SVM dan sebagainya digunakan sebagai pembelajaran mesin untuk proses pemodelan dan eksperimen.

Dalam kajian ini, algoritma Naive Bayes, Mesin Vektor Sokongan, pohon keputusan dan regresi logistik akan digunakan untuk membina model bagi meramalkan kebolehpasaran graduan. Shaneth C. Ambat. (2016) mengatakan bahawa model pohon keputusan tidak hanya memberikan ukuran seberapa relevannya peramal (ukuran pekali), tetapi juga arah perkaitannya (positif atau negatif). Umunya, ini amat penting untuk menguji model latihan data tetapi pada data baru yang belum digunakan untuk latihan. Dengan beberapa data set dan untuk mengelakkan kelebihan tepat, berpecah kaedah pengesahan digunakan dengan pembahagian peratusan 80/20, 70/30, 60/40 dan 50/50. Sebagai contoh, nisbah untuk memisahkan data adalah 70% daripada data lengkap untuk latihan yang membina model dan baki 30% untuk menguji model yang digunakan untuk menilai prestasi. Peratusan ketepatan dikira menggunakan pengesahan berpecah. Pengesahan berpecah mempunyai latihan dan sub ujian proses. Proses sub latihan digunakan untuk pembelajaran dan membina model latihan maka model latihan diterapkan di proses sub ujian. Prestasi model diukur semasa fasa ujian.

Tahap ini menjelaskan pembelajaran mesin yang digunakan dalam kajian ini adalah teknik klasifikasi. Dalam kajian ini, teknik klasifikasi digunakan kerana keupayaan pendekatan klasifikasi bukan sahaja mengendalikan jumlah data yang besar seperti data set yang diperoleh oleh dari tahun 2016, 2017 dan 2018 untuk menemui pola tersembunyi dan hubungan membantu dalam membuat keputusan, tetapi juga mengurangkan kelenturan struktur generasi data. Ujian ini boleh diulang dengan berbeza perubahan parameter untuk mendapat keputusan yang paling tepat. Proses pra-pemprosesan data juga amat penting kerana proses ini dapat membantu membersihkan data supaya mendapat data kualiti yang baik dan mendapat keputusan yang tepat. Fasa-fasa pra-pemprosesan data yang telah dijalankan adalah membersihkan data, mengurangkan data, transformasi data dan pelabelan data.

Selepas proses pra-pemprosesan data, pembelajaran teknik klasifikasi mengetahui fungsi perlombongan data menyerahkan item dalam koleksi untuk menyasarkan kategori atau kelas. Matlamat klasifikasi adalah dengan tepat meramalkan kelas sasaran untuk setiap kes dalam data. Sebagai contoh, model klasifikasi digunakan dalam kajian ini untuk mengenal pasti bilangan yang graduan dapat bekerja atau tidak bekerja. Perbandingan antara 4 algoritma iaitu, Pohon Keputusan, Regresi Logistik, Naive Bayes dan Mesin Vektor Sokongan adalah untuk menepatkan kumpulan data latihan mengenai beberapa faktor yang akan mempengaruhi kebolehpasaran graduan. Seterusnya, perisian Program Waikato Environment for Knowledge Analysis (Weka) telah dipilih dan digunakan dalam kajian ini kerana mudah diterapkan pada set data.

Dalam kajian ini, terdapat dua jenis asas data set; data set latihan yang digunakan oleh pengelasan untuk membina model dengan belajar dari data yang diberikan dan set ujian iaitu dikenali sebagai data set pengesahan yang bertujuan untuk menganggarkan prestasi model ramalan. Pengesahan silang K-fold telah dipilih yang diambil kira untuk membahagikan kumpulan data latihan ke dalam subset k dengan saiz yang sama. Dalam setiap lelaran, satu bahagian dikhaskan untuk data set pengesahan dan sisanya k-1 dipertahankan sebagai data latihan. Sebagai tambahan, kriteria perbandingan antara algoritma boleh dinilai berdasarkan kriteria berikut:

I. Ketepatan klasifikasi: Keupayaan model untuk meramalkan betul label kelas yang dinyatakan sebagai peratusan.

II. Kelajuan: Kelajuan merujuk kepada masa yang diambil untuk menetapkan model.

III. Kekuatan: Keupayaan untuk meramalkan model dengan betul walaupun ada tersebut mempunyai pemerhatian yang bising dan nilai yang hilang.

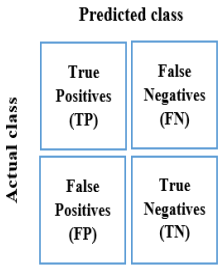
IV. Skalabiliti: Keupayaan model menjadi tepat dan produktif semasa mengendalikan peningkatan jumlah data.

V. Kebolehfahaman: Tahap pemahaman yang

disediakan oleh model.

VI. Struktur Peraturan: Kefahaman struktur peraturan algoritma.

Pengiraan perbandingan antara empat algoritma klasifikasi mengikut enam kriteria statistik. Nilai-nilai kriteria statistik yang dibandingkan dengan algoritma klasifikasi yang digunakan dikira dengan menggunakan matriks kekeliruan iaitu, kadar ketepatan, ketepatan, pengingat, F-ukuran, kawasan ROC dan kuadrat rata akar. Kadar ketepatan adalah peratusan ramalan yang betul manakala ketepatan adalah pecahan yang betuk meramalkan pemerhatian positif di antara jumlah tinjauan positif yang diramalkan. Selain itu, pengingat merujuk pecahan yang betul meramalkan pemerhatian positif di kalangan semua pemerhatian di dalam kelas. Tambahan pula, F-ukuran merupakan kriteria ketepatan dan pengingat boleh ditafsirkan bersama dan bukan secara individu. Rajah 1.2 menunjukkan formula matriks kekeliruan.



Measure	formula
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Precision	TP/ (TP+FP)
Recall	TP/ (TP+FN)
F-Measure	2*Precision*Recall/ (Precision+Recall)

Rajah 1.2 Formula Matriks Kekeliruan

Model akan dilahirkan melalui Weka dan melahirkan hubungan di antara set data untuk mengetahui bagaimana ciri-ciri atribut mempengaruhi status bekerja atau tidak bekerja. Dalam Weka, peratusan berpecah akan dilaraskan mengikut peratusan yang ditetapkan dengan nisbah seperti [90:10], [80:20], [70:30], [60:40], [50:50], [40:60], [30:70], [20:80] dan [10:90]. Sebagai contoh, nisbah-nisbah tersebut adalah 80% data latihan dan 20% data ujian. Nisbah-nisbah untuk memisahkan data adalah 80% daripada data lengkap untuk latihan yang membina model dan baki 20% untuk menguji model yang digunakan untuk menilai prestasi. Jadi, keputusan yang didapati mesti melebihi 80% supaya keputusan tersebut lebih tepat dan memilih model yang mempunyai ketepatan yang tinggi sebagai model terbaik dalam kajian ini.

E. Penilaian

Penilaian adalah untuk menguji dengan teliti model dan mengkaji semula langkah-langkah yang dilaksanakan untuk membina model untuk memastikan ia mencapai matlamat kajian ini. Pada akhir fasa ini, keputusan mengenai penggunaan keputusan perlombongan data harus dicapai. Fasa ini adalah untuk menilai hasil pemodelan dan mengkaji semula langkah-langkah yang dilaksanakan untuk membina model untuk memastikan ia mencapai matlamat kajian ini. Perbandingan antara model klasifikasi yang berbeza perlu dilakukan untuk mendapatkan ketepatan klasifikasi. Alat perlombongan data menyediakan pilihan untuk pembahagian kepada tiga bahagian yang berbeza adalah ujian, pengesahan dan latihan.

Dalam kajian ini, penilaian keputusan perlombongan data akan dijalankan dengan mengambil keputusan perlombongan data melibatkan model klasifikasi yang semestinya berkaitan dengan objektif kajian tersebut. Hasil penilaian ringkasan dari segi kriteria kejayaan ketepatan seluruh diperolehi oleh kaedah pengesahan model. Model yang diluluskan juga penting dalam fasa ini, setelah menilai model berkenaan dengan kriteria kejayaan model klasifikasi yang dihasilkan yang memenuhi kriteria terpilih menjadi model yang diluluskan. Salah satu kaedah yang terkenal adalah kaedah menvisualisasikan prestasi ramalan metrik yang boleh didapati dengan membina matriks kekeliruan.

F. Pengaplikasian

Pengaplikasian adalah perlu menentukan bagaimana hasilnya akan digunakan. Pengetahuan yang diperolehi perlu disusun dan disajikan dengan cara yang dapat digunakan oleh pelanggan. Bergantung kepada keperluan, fasa ini boleh semudah menghasilkan laporan seperti melaksanakan proses perlombongan data yang boleh diulangi di seluruh persusahan.

Dalam kajian ini, keputusan didapati bahawa perbandingan empat model akan memberi ketepatan yang berbeza. Model klasifikasi memberi data yang berguna dan berkesan dalam menjalani kajian ini. Sekiranya, model klasifikasi yang memberi ketepatan yang tinggi adalah penting untuk menganalisis atribut yang memberi kesan kepada kebolehpasaran graduan. Selepas kajian telah disiapkan, hasil kajian akan disampaikan kepada penyelia untuk mengubahsuai. Penyelidikan ini akan diterbitkan sekiranya dilakukan dengan sempurna dan tidak bersalah supaya dapat memberi pengajaran dan sebagai rujukan untuk graduan pada masa akan datang.

IV. KEPUTUSAN DAN PERBINCANGAN

Dalam bahagian ini, hasil pemodelan untuk empat algoritma adalah dibandingkan bagi tiga tahun iaitu 2016, 2017 dan 2018. Keempat-empat algoritma adalah Pohon Keputusan, Regresi Logistik, Naive Bayes dan Mesin Vektor Sokongan (SMO). Pengesahan silang K-fold telah dipilih yang diambil kira untuk membahagikan kumpulan data latihan ke dalam subset k dengan saiz yang sama. Mekanisme ujian yang digunakan ialah silang 10 kali ganda pengesahan. Dalam WEKA, dengan pengesahan silang sampel data bahagikan set data menjadi 10 keping ("lipatan"), kemudian tahan setiap bahagian secara bergilir untuk menguji dan melatih 9 baki bersama. Ini memberikan 10 hasil penilaian yang merata-rata. Dalam pengesahan silang berstrata, ketika melakukan pembahagian awal memastikan bahawa setiap lipatan mengandungi kira-kira bahagian nilai kelas yang betul. Setelah melakukan pengesahan silang 10 kali ganda dan menghitung hasil penilaian, WEKA menggunakan algoritma pembelajaran pada akhir (ke-11) pada keseluruhan kumpulan data untuk mendapatkan model yang dicetaknya. Prestasi algoritma dibandingkan berdasarkan pada ketepatan ROC (*Receiver Operating Characteristic*), RMSE (*Root mean Squared Error*) dan masa yang diperlukan untuk membina model.

a) Keputusan pengujian model

Data set tahun 2016 akan menjalankan teknik klasifikasi dengan menggunakan algoritma J48, regresi logistik, Naive Bayes dan SMO. Daripada hasil yang menjalani teknik klasifikasi, satu algoritma akan dipilih untuk membina model bagi setiap tahun dalam kajian ini.

Jadual 1.3 Data set tahun 2016 keputusan pengesahan silang 10 kali lipatan bagi J48

Pohon Keputusan (J48)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.29	60.7843	0.4718	0.644
80	20	0.16	60.7443	0.4731	0.638
70	30	0.17	60.7043	0.4722	0.636
60	40	0.15	60.9844	0.472	0.634
50	50	0.17	61.2645	0.4707	0.638
40	60	0.15	61.2245	0.4715	0.632
30	70	0.19	61.1044	0.4731	0.632
20	80	0.15	61.5846	0.4704	0.626
10	90	0.16	61.1845	0.4722	0.625
Pengesahan silang 10 kali ganda		0.17	61.0644	0.4718	0.634

Berdasarkan jadual 1.3, ketepatan purata bagi J48 adalah 61.0644. Masa purata untuk membina model adalah 0.17 saat. Nilai RMSE dan ROC adalah 0.4718 dan 0.634. Ketepatan yang terbaik adalah pada 20 latihan dan 80 ujian kali lipatan. Keputusan bagi nilai RMSE dan ROC tidak mengalami terlalu peningkatan dan penurunan.

Jadual 1.4 Data set tahun 2016 keputusan pengesahan silang 10 kali lipatan bagi regresi logistik

Regresi Logistik					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	21.58	88.1553	0.3197	0.912
80	20	18.15	88.1152	0.3206	0.907
70	30	22.79	88.0752	0.3228	0.903
60	40	20.74	88.4754	0.3203	0.907
50	50	26.82	88.4354	0.3179	0.908
40	60	31.31	88.1152	0.3217	0.905
30	70	35.89	88.7155	0.3171	0.908
20	80	54.51	88.4754	0.3185	0.907
10	90	72.36	88.1152	0.3226	0.903
Pengesahan silang 10 kali ganda		35.32	88.2975	0.3201	0.907

Berdasarkan jadual 1.4, ketepatan purata bagi regresi logistik adalah 88.2975. Nilai RMSE dan ROC adalah 0.3201 dan 0.907. Ketepatan yang terbaik adalah pada 30 latihan dan 70 ujian kali lipatan. Keputusan bagi nilai RMSE dan ROC tidak mengalami terlalu peningkatan dan penurunan.

Jadual 1.5 Data set tahun 2016 keputusan pengesahan silang 10 kali lipatan bagi Naive Bayes

Naive Bayes					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.02	70.068	0.4808	0.799
80	20	0	69.5078	0.4813	0.798
70	30	0	69.4678	0.4817	0.797
60	40	0	69.5878	0.4815	0.798
50	50	0	69.7079	0.4811	0.798
40	60	0	69.5078	0.4817	0.797
30	70	0	69.4278	0.482	0.797
20	80	0	69.4278	0.4817	0.797
10	90	0	69.3477	0.482	0.797
Pengesahan silang 10 kali ganda		0.002	69.5612	0.4815	0.798

Berdasarkan jadual 1.5, purata ketepatan pengesahan silang 10 kali lipatan adalah 69.5612%. Jadual ini menunjukkan mempunyai ketepatan yang terendah pada 10/90 kali lipatan dan tertinggi pada 90/10 kali lipatan. Ukuran untuk RMSE dan ROC masih stabil dan tidak mempunyai terlalu banyak perubahan dalam sepuluh lipatan. Purata untuk RMSE adalah 0.4815 dan ROC adalah 0.798.

Jadual 1.6 Data set tahun 2016 keputusan pengesahan silang 10 kali lipatan bagi SMO

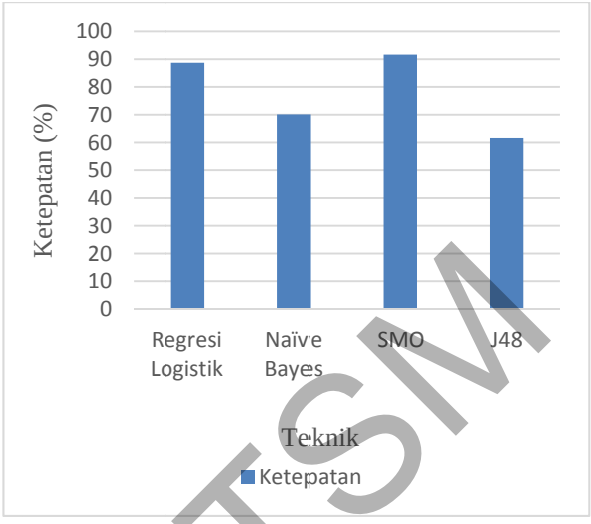
Mesin Vektor Sokongan (SMO)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	14.83	91.3966	0.2933	0.916
80	20	10.62	91.4366	0.2926	0.917
70	30	14.79	91.5166	0.2913	0.918
60	40	8.93	91.5966	0.2899	0.918
50	50	8.74	91.4366	0.2926	0.917
40	60	8.76	91.5566	0.2906	0.918
30	70	8.75	91.6367	0.2892	0.919
20	80	8.67	91.4766	0.2919	0.917
10	90	8.70	91.4766	0.2919	0.917
Pengesahan silang 10 kali ganda		10.31	91.5032	0.2914	0.917

Berdasarkan jadual 1.6, ketepatan pengesahan silang 10 kali lipatan pada 80/20 kali lipatan dan 50/50 kali lipatan adalah sama, iaitu 91.4366%. Bagi ketepatan yang tertinggi adalah 91.6367% pada 30/70 kali lipatan dan purata ketepatan pengesahan silang 10 kali lipatan adalah 91.5032%. Jadual ini menunjukkan trend yang semakin meningkat dari 90/10 kali lipatan hingga 60/40 kali lipatan seterusnya menurun pada 50/50 kali lipatan dan meningkat sekali dari 40/60 kali lipatan hingga 30/70 kali lipatan. Ketepatan terendah adalah 91.3966% pada 90/10 kali lipatan. RMSE dan ROC nilai untuk pengesahan silang 10 kali lipatan telah menunjukkan dalam jadual. RMSE yang tertinggi adalah 0.2933 pada 90/10 kali lipatan dan mengalami penurunan trend. Manakala nilai ROC yang tertinggi adalah 0.919 pada 30/70 kali ganda dan menunjukkan trend yang stabil.

Jadual 1.7 Data set tahun 2016 keputusan untuk pengesahan silang 10 kali lipatan

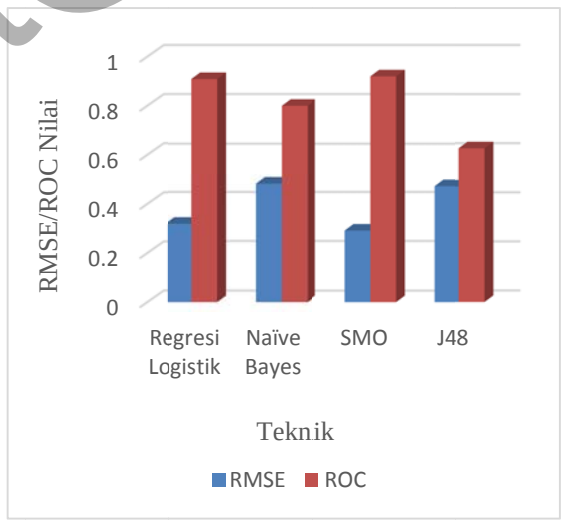
Teknik	Algoritma	Masa (s)	Ketepatan (%)	RMSE	ROC
Regresi Logistik	Regresi Logistik	35.89	88.7155	0.3171	0.908
Naive Bayes	Naive Bayes	0.02	70.068	0.4808	0.799
Mesin Vektor Sokongan	SMO	8.75	91.6367	0.2892	0.919
Pohon Keputusan	J48	0.15	61.5846	0.4704	0.626

Jadual 1.7 menunjukkan keempat-empat algoritma yang telah menjalani teknik klasifikasi untuk mendapat ketepatan supaya memilih algoritma yang sesuai untuk membina model. Dari jadual 1.7, algoritma SMO mendapat ketepatan yang tertinggi berbanding dengan algoritma J48, Regresi Logistik dan Naive Bayes, iaitu 91.6367%. Perbezaan ketepatan antara algoritma SMO dengan Naive Bayes adalah terlalu besar tetapi terlalu kecil berbanding dengan Regresi Logistik, iaitu hampir 3%. Algoritma J48 mempunyai ketepatan yang paling rendah adalah 61.5864%.



Rajah 1.3 Data set tahun 2016 perbandingan ketepatan antara J48, Regresi Logistik, Naive Bayes dan SMO

Dari rajah 1.3, algoritma SMO menunjukkan prestasi yang baik dalam purata ketepatan manakala algoritma J48, Regresi Logistik dan Naive Bayes tidak menunjukkan prestasi yang baik dalam kajian ini.



Rajah 1.4 Data set tahun 2016 perbandingan nilai RMSE dan ROC antara J48, Regresi Logistik, Naive Bayes dan SMO

Tambahan pula, nilai RMSE bagi SMO adalah terendah berbanding dengan algoritma J48, Regresi Logistik dan Naive Bayes. Bagi ROC nilai, SMO menunjukkan nilai tertinggi, iaitu 0.919 yang berdekatan dengan 1 (nilai ROC menghampiri 1 adalah terbaik). SMO mempunyai ketepatan dan nilai ROC yang tertinggi berbanding dengan J48, Regresi Logistik dan Naive Bayes.

Ketepatan SMO mempunyai ketepatan yang terbaik daripada algoritma J48, Regresi Logistik dan Naive Bayes. Dalam kajian oleh Binh et al. (2018), pengkelasan SVM mempunyai ketepatan yang terbaik daripada algoritma Bayesian seperti NB, NBT, BN dan DTNB. Walaupun masa yang diambil untuk membina model bagi algoritma Naive Bayes adalah pendek berbanding dengan algoritma J48, SMO dan Regresi Logistik namun SMO mempunyai ketepatan yang tertinggi. Dalam kajian ini, berbeza diawasi teknik pengkelasan seperti (Jin era l., 2003), Naive Bayes dan Mesin Vektor Sokongan (SMO) digunakan. Oleh itu, SMO mempunyai 91.6367% pada 30/70 kali lipatan merupakan ketepatan yang terbaik berbanding dengan algoritma J48, Regresi Logistik dan Naive Bayes. Jadi, SMO dipilih untuk membina model ramalan dalam kebolehpasaran graduan UKM dalam data set tahun 2016.

Seterusnya, data set tahun 2017 akan menjalankan teknik klasifikasi untuk mencari ketepatan yang tinggi supaya dapat membina model dalam kajian ini.

Jadual 1.8 Data set tahun 2017 keputusan pengesahan silang 10 kali lipatan untuk J48

Pohon Keputusan (J48)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.02	62.1003	0.5035	0.610
80	20	0.01	62.4637	0.5007	0.618
70	30	0.02	62.0276	0.4994	0.621
60	40	0.02	62.3183	0.4994	0.621
50	50	0.01	62.2820	0.4992	0.616
40	60	0.02	62.1366	0.4992	0.621
30	70	0.01	61.8823	0.4991	0.619
20	80	0.01	62.1366	0.4984	0.621
10	90	0.01	62.3910	0.4962	0.623
Pengesahan silang 10 kali ganda		0.01	62.1931	0.4995	0.618

Berdasarkan jadual 1.8, ketepatan purata bagi J48 adalah 62.1931%. Ketepatan menunjukkan arah aliran yang meningkat seterusnya menurun dan meningkat lagi. Ketepatan yang terbaik adalah pada 80 latihan dan 20 kali ujian. Sementara itu, jadual 1.8 menunjukkan dua ukuran adalah nilai RMSE dan ROC untuk 10 leleran pengesahan silang 10 kali lipatan J48. Hasilnya menunjukkan perubahan yang tidak signifikan dalam kedua-dua ukuran tersebut.

Jadual 1.9 Data set tahun 2017 keputusan pengesahan silang 10 kali lipatan untuk Regresi Logistik

Regresi Logistik					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	1.03	62.5727	0.4772	0.649
80	20	1.05	62.5727	0.4762	0.652
70	30	0.71	62.7907	0.4766	0.651
60	40	0.55	62.5363	0.4760	0.653
50	50	0.56	62.0640	0.4770	0.650
40	60	0.54	62.4637	0.4764	0.651
30	70	0.65	62.0640	0.4764	0.651
20	80	0.61	62.2093	0.4759	0.653
10	90	0.58	62.2820	0.4761	0.652
Pengesahan silang 10 kali ganda		0.65	62.3950	0.4764	0.651

Berdasarkan jadual 1.9, ketepatan purata bagi regresi logistik adalah 62.3950%. Ketepatan menunjukkan arah aliran yang meningkat seterusnya menurun dan meningkat lagi. Ketepatan yang terbaik adalah pada 70 latihan dan 30 kali ujian. Selain itu, jadual 1.9 menunjukkan dua ukuran adalah nilai RMSE dan ROC untuk 10 leleran pengesahan silang 10 kali lipatan regresi logistik. Hasilnya menunjukkan perubahan yang tidak signifikan dalam kedua-dua ukuran tersebut.

Jadual 1.10 Data set tahun 2017 keputusan pengesahan silang 10 kali lipatan untuk Naive Bayes

Naive Bayes					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.01	60.8285	0.5124	0.650
80	20	0	61.2282	0.5115	0.651
70	30	0	60.9375	0.5115	0.651
60	40	0	61.1555	0.5115	0.651
50	50	0	60.8648	0.5128	0.649
40	60	0	61.0102	0.5117	0.650
30	70	0	61.2645	0.5119	0.650
20	80	0	61.1192	0.5112	0.651
10	90	0	61.0102	0.5114	0.651
Pengesahan silang 10 kali ganda		0.001	61.0465	0.5117	0.650

Berdasarkan jadual 1.10, pengesahan silang 10 kali lipatan memberi purata ketepatan 61.0465%. Jadual ini menunjukkan Naive Bayes mempunyai paling rendah pada 90/10 lipatan dan tertinggi pada 30/70 lipatan. Dalam jadual 1.10 menunjukkan nilai RMSE dan ROC untuk pengesahan silang 10 kali lipatan. Ukuran untuk RMSE dan ROC mempunyai kecil perubahan dalam sepuluh lipatan. Purata untuk RMSE adalah 0.5117 dan ROC adalah 0.65.

Jadual 1.11 Data set tahun 2017 keputusan pengesahan silang 10 kali lipatan untuk Mesin Vektor Sokongan (SMO)

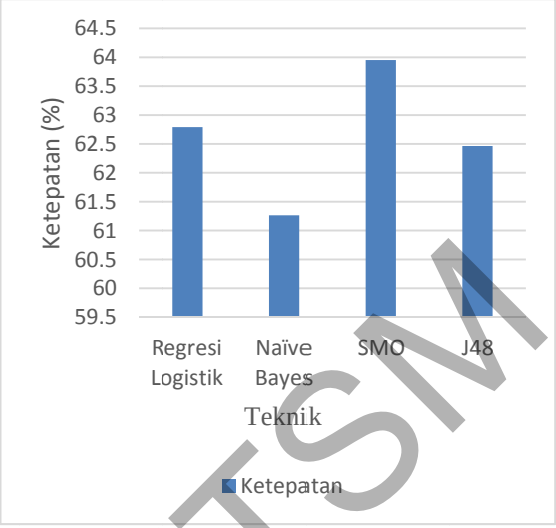
Mesin Vektor Sokongan (SMO)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	4.37	63.0814	0.6076	0.594
80	20	3.78	63.4448	0.6046	0.597
70	30	3.91	63.2267	0.6064	0.596
60	40	3.97	63.6628	0.6028	0.599
50	50	3.99	63.5901	0.6034	0.600
40	60	4.13	63.5901	0.6034	0.600
30	70	3.92	63.5174	0.604	0.598
20	80	4.08	63.9535	0.6004	0.603
10	90	4.14	63.8808	0.601	0.602
Pengesahan silang 10 kali ganda		4.03	63.5497	0.6037	0.598

Berdasarkan jadual 1.11, pembahagian latihan 20% dan 80% ujian untuk SMO memberikan yang tertinggi dikelaskan dengan betul iaitu, 63.9535%. Secara purata, ketepatannya adalah 63.5497%. Jadual telah menunjukkan peningkatan trend ketepatan SMO. Sebagai ditunjukkan dalam jadual, ketepatan tertinggi ialah 20/80 kali ganda dan terendah ialah 90/10 kali lipatan. RMSE mempunyai nilai yang tertinggi pada 90/10 kali lipatan dan menunjukkan penurunan trend. Tambahan pula, nilai ROC mempunyai terendah pada 90/10 kali lipatan dan mengalami kenaikan sedikit.

Jadual 1.12 Data set tahun 2017 keputusan untuk pengesahan silang 10 kali lipatan

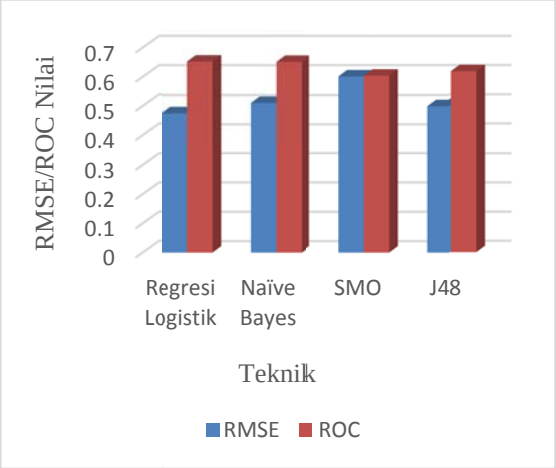
Teknik	Algoritma	Masa (s)	Ketepatan (%)	RMSE	ROC
Regresi Logistik	Regresi Logistik	0.71	62.7907	0.4766	0.651
Naive Bayes	Naive Bayes	0	61.2645	0.5119	0.650
Mesin Vektor Sokongan	SMO	4.08	63.9535	0.6004	0.603
Pohon Keputusan	J48	0.01	62.4637	0.5007	0.618

Jadual 1.12 menunjukkan ketepatan tertinggi diperoleh dengan mengaplikasikan algoritma SMO berbanding dengan algoritma lain, iaitu 63.9535%. Untuk tertinggi kedua ialah 62.7907% diperoleh dari regresi logistik. Algoritma Naive Bayes mengandungi ketepatan yang paling rendah berbanding dengan algoritma J48, Regresi Logistik dan SMO, iaitu 61.2645%. Selain itu, ketepatan SMO lebih tinggi berbanding dengan algoritma Regresi Logistik tetapi algoritma Regresi Logistik mengambil masa yang pendek untuk membina model dalam masa 0.71 saat berbanding dengan algoritma SMO. Sementara itu, SMO mengambil masa 4.08 saat. Algoritma Naive Bayes mengambil masa yang terpendek membina model adalah 0 saat.



Rajah 1.5 Data set tahun 2017 perbandingan ketepatan antara J48, Regresi Logistik, Naive Bayes dan SMO

Prestasi keempat-empat teknik telah ditunjukkan dalam rajah 1.5. Regresi Logistik dan SMO telah menunjukkan prestasi yang baik di segi peratusan ketepatan. Sementara itu, Naive Bayes dan J48 tidak menunjukkan prestasi yang baik dalam kajian ini.



Rajah 1.6 Data set tahun 2017 perbandingan antara nilai RMSE dan ROC untuk J48, Regresi Logistik, Naive Bayes dan SMO

Di samping itu, nilai RMSE untuk Regresi Logistik adalah yang paling rendah, iaitu 0.4766 berbanding J48, Naive Bayes dan SMO. Matriks prestasi ROC untuk Regresi Logistik menunjukkan nilai yang tertinggi (nilai ROC menghampiri 1 adalah lebih baik), 0.651 berbanding dengan J48, 0.618, Naive Bayes, 0.650 dan SMO, 0.603. Selain itu, ketepatan Naive Bayes adalah terendah berbanding dengan J48, Regresi Logistik dan SMO,

jadi Naive Bayes tidak sesuai untuk membina model untuk kajian ini. Seterusnya, algoritma Regresi Logistik mempunyai ketepatan yang rendah berbanding dengan SMO tetapi masa untuk membina model bagi algoritma Regresi Logistik adalah pendek, iaitu 0.71 saat berbanding dengan SMO, 4.08 saat dan juga nilai ROC lebih baik dari SMO pada 70/30 kali lipatan. Algoritma Naive Bayes mempunyai yang paling pendek berbanding dengan regresi logistik dan SMO tetapi ketepatan tidak melebihi algoritma regresi logistik dan SMO. Oleh itu, algoritma Regresi Logistik adalah sesuai untuk membina model bagi data set tahun 2017.

Data set tahun 2018 telah menjalankan teknik klasifikasi dengan menggunakan WEKA untuk mendapat ketepatan yang terbaik supaya dapat mencari algoritma yang sesuai untuk membina model dalam kajian ini.

Jadual 1.13 Data set tahun 2018 keputusan pengesahan silang 10 kali lipatan untuk J48

Pohon Keputusan (J48)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.5	58.981	0.4795	0.584
80	20	0.26	59.1249	0.4783	0.584
70	30	0.42	59.2976	0.4776	0.584
60	40	0.42	59.2401	0.4776	0.580
50	50	0.4	59.2976	0.4773	0.582
40	60	0.39	59.2976	0.4773	0.576
30	70	0.42	59.2976	0.4773	0.573
20	80	0.3	59.3552	0.4772	0.573
10	90	0.42	59.2976	0.4773	0.570
Pengesahan silang 10 kali ganda		0.39	59.2432	0.4777	0.578

Berdasarkan jadual 1.13, purata ketepatan pengesahan silang 10 kali lipatan adalah 59.2432%. Ketepatan yang paling tinggi adalah 59.3552% pada 20/80 lipatan. Dari jadual menunjukkan trend yang semakin meningkat dalam pengesahan silang 10 kali lipatan. Ketepatan yang paling rendah adalah 58.981% pada 90 latihan dan 10 ujian lipatan. Sementara itu, nilai RMSE semakin menurun dari 90/10 lipatan hingga 10/90 lipatan. Nilai ROC menunjukkan trend yang stabil dan tidak mempunyai perubahan dalam jadual.

Jadual 1.14 Data set tahun 2018 keputusan pengesahan silang 10 kali lipatan untuk Regresi Logistik

Regresi Logistik					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	30.38	85.9816	0.3372	0.900
80	20	35.90	86.2406	0.3350	0.902
70	30	41.81	85.9816	0.3368	0.901
60	40	52.46	86.3270	0.3349	0.902
50	50	66.44	86.1543	0.3359	0.902
40	60	78.82	86.0391	0.3380	0.900
30	70	90.49	86.1831	0.3359	0.902
20	80	104.42	86.1255	0.3361	0.902
10	90	113.59	86.0967	0.3360	0.902
Pengesahan silang 10 kali ganda		68.26	86.1255	0.3362	0.901

Berdasarkan jadual 1.14, purata ketepatan pengesahan silang 10 kali lipatan adalah 86.1255%. Ketepatan yang paling tinggi adalah 86.3270% pada 60/40 lipatan. Dari jadual di atas menunjukkan trend yang berbukit dalam pengesahan silang 10 kali lipatan. Ketepatan yang paling rendah adalah 85.9816% pada 90 latihan dan 10 ujian lipatan dan juga 70 latihan dan 30 ujian lipatan. Sementara itu, nilai RMSE semakin menurun dari 90/10 lipatan hingga 10/90 lipatan. Nilai ROC menunjukkan trend yang stabil dan tidak mempunyai perubahan dalam jadual.

Jadual 1.15 Data set tahun 2018 keputusan pengesahan silang 10 kali lipatan untuk Naive Bayes

Naive Bayes					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	0.02	72.5676	0.4591	0.783
80	20	0	72.8555	0.4575	0.785
70	30	0	72.7691	0.4569	0.786
60	40	0	72.7116	0.4568	0.786
50	50	0	72.9419	0.4575	0.785
40	60	0	72.7116	0.4568	0.786
30	70	0	72.7979	0.4568	0.786
20	80	0	72.8555	0.457	0.785
10	90	0	72.9419	0.4565	0.786
Pengesahan silang 10 kali ganda		0.002	72.7947	0.4572	0.785

Berdasarkan jadual 1.15 purata ketepatan pengesahan silang 10 kali lipatan adalah 72.7947%. Masa purata yang diambil untuk membina model terlalu pendek adalah 0.002 saat. Jadual menunjukkan ketepatan yang sama pada 50/50 lipatan dan 10/90 lipatan adalah 72.9419%. Pada 90/10 lipatan, ketepatan adalah terendah, iaitu 72.5676%. Nilai RMSE dan ROC menunjukkan trend yang stabil dan tidak mempunyai perubahan dalam pengesahan silang 10 kali lipatan bagi Naive

Bayes. Purata RMSE adalah 0.4572 dan purata ROC adalah 0.785.

Jadual 1.16 Data set tahun 2018 keputusan pengesahan silang 10 kali lipatan untuk SMO

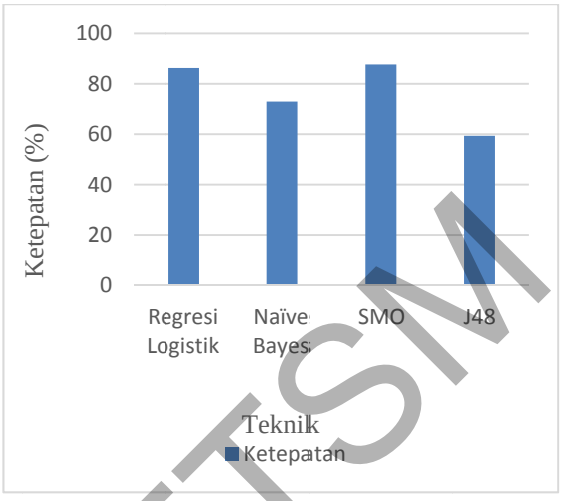
Mesin Vektor Sokongan (SMO)					
Latihan	Ujian	Masa (s)	Ketepatan (%)	RMSE	ROC
90	10	57.32	87.6799	0.351	0.886
80	20	40.7	87.4784	0.3539	0.884
70	30	41.29	87.2481	0.3571	0.882
60	40	54	87.5072	0.3535	0.885
50	50	36.27	87.3921	0.3551	0.884
40	60	39.03	87.2769	0.3567	0.882
30	70	47.9	87.6511	0.3514	0.886
20	80	63.23	87.3633	0.3555	0.883
10	90	72.55	87.3921	0.3551	0.883
Pengesahan silang 10 kali ganda		50.25	87.4432	0.3543	0.883

Berdasarkan jadual 1.16, purata ketepatan pengesahan silang 10 kali lipatan untuk SMO adalah 87.4432% dalam data set tahun 2018. Dari jadual 1.16 menunjukkan ketepatan pengesahan silang 10 kali lipatan untuk SMO. Ketepatan tertinggi adalah 87.6799% pada 90/10 lipatan dan terendah adalah 87.2481% pada 70/40 lipatan. RMSE adalah yang tertinggi pada lipatan 70/30 lipatan dan menunjukkan arah aliran menurun. Sementara itu, nilai ROC mempunyai terendah pada 70/30 lipatan dan 40/60 lipatan.

Jadual 1.17 Keputusan pengesahan silang 10 kali lipatan

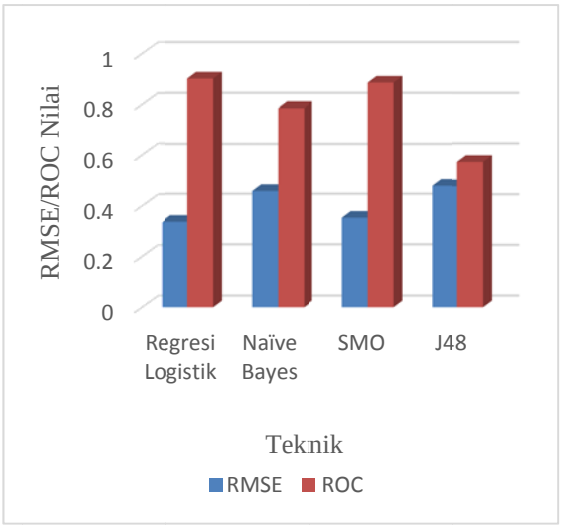
Teknik	Algoritma	Masa (s)	Ketepatan (%)	RMSE	ROC
Regresi Logistik	Regresi Logistik	52.46	86.3270	0.3349	0.902
Naive Bayes	Naive Bayes	0	72.9419	0.4565	0.786
Mesin Vektor Sokongan	SMO	57.32	87.6799	0.351	0.886
Pohon Keputusan	J48	0.3	59.3552	0.4772	0.573

Jadual 1.17 menunjukkan keputusan pengesahan silang 10 kali lipatan untuk algoritma J48, Regresi Logistik, Naive Bayes dan SMO. SMO mendapat ketepatan yang tertinggi berbanding dengan Regresi Logistik dan Naive Bayes, iaitu 87.6799%. Regresi Logistik dan Naive Bayes juga mendapat ketepatan 86.3270% dan 72.9419%. Ketepatan yang terendah bagi algoritma J48 adalah 59.3552%. Ketepatan SMO tertinggi tetapi masa yang diguna untuk membina model terlalu panjang, iaitu 57.32 saat. Oleh itu, Naive Bayes mempunyai ketepatan yang terendah tetapi masa yang diguna untuk membina model adalah pendek hanya 0 saat saja.



Rajah 1.7 Data set tahun 2018 perbandingan ketepatan antara J48, Regresi Logistik, Naive Bayes dan SMO

Prestasi antara keempat-empat teknik telah ditunjukkan dalam rajah 1.7. Regresi Logistik dan SMO telah menunjukkan prestasi yang baik dalam ketepatan pengesahan silang 10 kali lipatan. Naive Bayes dan J48 belum menunjukkan prestasi yang baik dalam kajian ini.



Rajah 1.8 Data set tahun 2018 perbandingan nilai RMSE dan ROC antara J48, Regresi Logistik, Naive Bayes dan SMO

Di samping itu, nilai RMSE untuk regresi logistik adalah terendah, iaitu 0.3349 berbanding dengan J48, SMO dan Naive Bayes. Nilai ROC untuk regresi logistik adalah tertinggi, iaitu 0.902 berbanding dengan Naive Bayes, 0.786. Nilai ROC bagi J48 adalah terendah, iaitu 0.573.

Dari jadual 1.17, ketepatan SMO adalah paling tinggi berbanding dengan algoritma J48, Regresi Logistik dan Naive Bayes, iaitu 87.6799%. Masa yang diambil untuk membina model bagi SMO adalah terlalu panjang berbanding dengan J48, SMO dan Naive Bayes, iaitu 57.32 saat. Jadi, keputusan ujian mendapati bahawa algoritma SMO merupakan model yang paling sesuai untuk membina model ramalan kebolehpasaran graduan UKM pada data set tahun 2018 dalam kajian ini.

b) Analisis Pengetahuan

Bagi kajian ini, output model ramalan telah ditukar menjadi pemahaman yang boleh diambil tindakan kerana penting untuk memahami faktor-faktor yang menyumbang paling banyak kepada ramalan. Pada data set tahun 2016, 2017 dan 2018, teknik yang mempunyai prestasi yang baik untuk membina model adalah regresi logistik dan mesin vektor sokongan (SMO). Namun, masa untuk membina model tidak mempengaruhi keputusan SMO menjadi algoritma yang paling sesuai membina model kerana ketepatan untuk SMO dan regresi logistik telah memenuhi standard bagi kajian ini adalah melebihi 75%. Dari data set tahun 2016 mempunyai sama matlamat iaitu ketepatan melebihi 75% tetapi bagi data set tahun 2017 ketepatan antara 4 algoritma tidak memenuhi standard ketepatan. Jadi, data set tahun 2017 akan mengambil kira masa untuk membina model merupakan faktor yang menyebabkan keputusan algoritma yang sesuai membina model bagi kajian ini. Jadi, kedua-dua teknik tersebut akan dianalisis selanjutnya. Walau bagaimanapun algoritma J48 mendapat ketepatan yang rendah, namun algoritma J48 mempunyai sub pohon keputusan yang baik dan boleh tunjuk faktor yang mempengaruhi kebolehpasaran graduan.

Dari algoritma regresi logistik, 70/30 kali lipatan mempunyai ketepatan tertinggi antara 90/10 hingga 10/90 kali lipatan dalam data set tahun 2017. Jadi 70/30 kali lipatan dipilih untuk membina model tersebut dalam kajian ini. Analisis di atas menentukan bahawa regresi logistik adalah algoritma klasifikasi terbaik bagi data set tahun 2017. Menurut hasil ini, kami mengaplikasikan algoritma regresi logistik sekali lagi untuk menentukan pemboleh ubah yang boleh menyebabkan risiko lalai. Menurut Antonio J. Pena (2019), analisis chi-square diterapkan melalui WEKA Select Panel atribut untuk menentukan pemboleh ubah yang menjelaskan model yang terbaik. Chi-square adalah analisis yang menunjukkan nilai pemboleh ubah yang dipilih bergantung pada pemboleh ubah kelas digunakan apabila pemboleh ubah adalah nominal. Analisis 1e mengesyorkan mengurangkan varian peringkat terendah dari model. Namun, kerana WEKA tidak dapat menggunakan chi-square analisis berdasarkan algoritma, lebih baik dikecualikan pemboleh ubah dari model. Keputusan peringkat menerapkan analisis chi-square terhadap pemboleh ubah ditunjukkan dalam jadual 1.18.

Jadual 1.18 Keputusan Chi-square

Peringkat rata-rata	Atribut
1 +- 0	6 Program Pengajian
2 +- 0	7 Pengkhususan
3 +- 0	4 Fakulti
4.1 +- 0.3	5 Bidang Pengajian Utama
4.9 +- 0.3	9 Pendapatan Keluarga
6 +- 0	1 Jantina
7 +- 0	8 Pencapaian Akademik
8 +- 0	3 Kelayakan Bahasa (MUET)
9 +- 0	2 Keturunan

Mengkaji analisis ini, melihat bahawa pemboleh ubah atribut program pengajian adalah berangka kedudukan terendah. Oleh itu, pemboleh ubah ini mesti sama ada akan ditinggalkan atau ditukar menjadi nilai nominal jika hendak termasuk dalam analisis. Oleh itu, mengubah atribut program pengajian menjadi pemboleh ubah kategori dan analisis regresi logistik dan analisis chi-square untuk pemilihan pemboleh ubah diulang.

Jadual 1.19 Keputusan regresi logistik atribut program pengajian nominal

Ketepatan (%)	62.7907
Nilai RMSE	0.4766
Nilai ROC	0.651

Ketepatan bagi atribut program pengajian nominal adalah 62.7907%.

Jadual 1.20 Keputusan regresi logistik tidak memasukkan atribut program pengajian nominal

Ketepatan (%)	62.8105
Nilai RMSE	0.4765
Nilai ROC	0.652

Oleh kerana dari hasil ini jelas bahawa klasifikasi data diperbaiki apabila pemboleh ubah program pengajian dikeluarkan yang pemboleh ubah akan dikeluarkan dan analisis diteruskan.

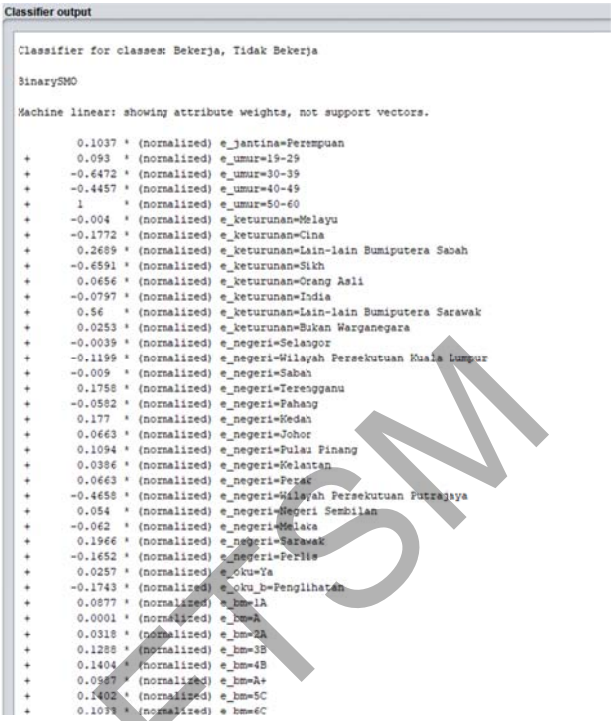
```
Odds Ratios...

                                                                    Clas
Variable                                                  Tidak Bekerj
================================================================================
Jantina=Lelaki                                                    0.733
Kelayakan Bahasa (MUET)=Band 4 hingga band 6                      1.046
Kelayakan Bahasa (MUET)=Band 1 hingga band 3                      0.909
Kelayakan Bahasa (MUET)=Tidak Berkaitan                           1.251
Fakulti=FAKULTI EKONOMI DAN PENGURUSAN                            0.829
Fakulti=FAKULTI KEJURUTERAAN DAN ALAM BINA                        1.099
Fakulti=FAKULTI PENDIDIKAN                                        1.302
Fakulti=FAKULTI PENGAJIAN ISLAM                                   0.894
Fakulti=FAKULTI SAINS DAN TEKNOLOGI                               1.309
Fakulti=FAKULTI SAINS KESIHATAN                                    0.53
Fakulti=FAKULTI SAINS SOSIAL DAN KEMANUSIAAN                      1.326
Fakulti=FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT                      0.583
Bidang Pengajian Utama=Sastera & Sains Sosial                     1.035
Bidang Pengajian Utama=Teknik                                     1.099
Bidang Pengajian Utama=Pendidikan                                 1.302
Bidang Pengajian Utama=Sains                                      1.050
Bidang Pengajian Utama=Teknologi Maklumat & Komunikasi            0.583
Pengkhususan=EKONOMI                                              1.221
Pengkhususan=KEUSAHAWANAN DAN INOVASI                             1.222
Pengkhususan=PENTADBIRAN PERNIAGAAN                               0.718
Pengkhususan=PERAKAUNAN                                           0.727
Pengkhususan=KEJURUTERAAN AWAM DAN SEKITARAN                      0.966
Pengkhususan=KEJURUTERAAN AWAM DAN STRUKTUR                       0.665
Pengkhususan=KEJURUTERAAN BIOKIMIA                                0.697
```

Rajah 1.9 Nisbah odds setiap atribut

Nisbah ini ditafsirkan dan dikaitkan dengan perbezaan kelas tanpa masuk ke sub kumpulan lalai. Menurut rajah 1.9, model meramalkan bahawa kemungkinan tidak bekerja adalah 0.7334 kali lebih tinggi bagi lelaki berbanding dengan wanita. Bagi graduan yang mendapat kelayakan bahasa (MUET) tidak berkaitan adalah 1.2513 kali tinggi berbanding dengan kelayakan bahasa (MUET) band 1 hingga band 6 akan tidak mendapat pekerjaan. Graduan yang dalam bidang pengajian pendidikan akan mempunyai 1.3022 kali tinggi tidak bekerja berbanding dengan bidang pengajian dalam sastera & sains sosial, teknik, sains dan teknologi maklumat & komunikasi.

Selain itu, mesin vektor sokongan (SVM) telah mendapat ketepatan yang paling tinggi dalam data set tahun 2016 dan 2018. Output yang didapati dari WEKA akan ditunjukkan dalam rajah 1.10. Dalam rajah 4.39, pengkelasan untuk kelas adalah bekerja atau tidak bekerja. Nombor-nombor seperti 0.1037, -0.6473 atau 1 adalah kepentingan ciri permutasi. Kepentingan ciri permutasi adalah teknik pemeriksaan model yang dapat digunakan untuk penaksir yang sesuai ketika data berbentuk tabel. Ini amat berguna untuk penganggar tidak linear atau legap. Kepentingan ciri permutasi didefinisikan sebagai penurunan dalam skor model apabila nilai satu fitur diacak secara rawak. Prosedur ini memutuskan hubungan antara ciri dan sasaran, sehingga penurunan skor model menunjukkan seberapa besar model bergantung pada fitur tersebut. Teknik ini mendapat manfaat daripada menjadi model agnostik dan dapat dikira berkali-kali dengan permutasi ciri yang berbeza.



```
Classifier output

Classifier for classes: Bekerja, Tidak Bekerja

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        0.1037 * (normalized) e_jantina=Perempuan
+       0.093  * (normalized) e_umur=19-29
+      -0.6472 * (normalized) e_umur=30-39
+      -0.4457 * (normalized) e_umur=40-49
+       1      * (normalized) e_umur=50-60
+      -0.004  * (normalized) e_keturunan=Melayu
+      -0.1772 * (normalized) e_keturunan=Cina
+       0.2689 * (normalized) e_keturunan=Lain-lain Bumiputera Sabah
+      -0.6591 * (normalized) e_keturunan=Sikh
+       0.0656 * (normalized) e_keturunan=Orang Asli
+      -0.0797 * (normalized) e_keturunan=India
+       0.56   * (normalized) e_keturunan=Lain-lain Bumiputera Sarawak
+       0.0253 * (normalized) e_keturunan=Bukan Warganegara
+      -0.0039 * (normalized) e_negeri=Selangor
+      -0.1199 * (normalized) e_negeri=Wilayah Persekutuan Kuala Lumpur
+      -0.009  * (normalized) e_negeri=Sabah
+       0.1758 * (normalized) e_negeri=Terengganu
+      -0.0582 * (normalized) e_negeri=Pahang
+       0.177  * (normalized) e_negeri=Kedah
+       0.0663 * (normalized) e_negeri=Johor
+       0.1094 * (normalized) e_negeri=Pulau Pinang
+       0.0386 * (normalized) e_negeri=Kelantan
+       0.0663 * (normalized) e_negeri=Perak
+      -0.4658 * (normalized) e_negeri=Wilayah Persekutuan Putrajaya
+       0.054  * (normalized) e_negeri=Negeri Sembilan
+      -0.062  * (normalized) e_negeri=Melaka
+       0.1966 * (normalized) e_negeri=Sarawak
+      -0.1652 * (normalized) e_negeri=Perlis
+       0.0257 * (normalized) e_oku=Ya
+      -0.1743 * (normalized) e_oku_b=Penglihatan
+       0.0877 * (normalized) e_bm=1A
+       0.0001 * (normalized) e_bm=A
+       0.0318 * (normalized) e_bm=2A
+       0.1288 * (normalized) e_bm=3B
+       0.1404 * (normalized) e_bm=4B
+       0.0987 * (normalized) e_bm=A+
+       0.1402 * (normalized) e_bm=5C
+       0.1031 * (normalized) e_bm=6C
```

Rajah 1.10 Output bagi SMO

Bagi algoritma J48, ketepatannya adalah 61.0644% paling rendah berbanding antara algoritma regresi logistik, naive bayes dan SMO dalam data set tahun 2016. Sebanyak 12 petua telah dihasilkan, atribut utama yang terlibat adalah umur, CGPA, e_17. Kajian ini perlu dilakukan bagi menjelaskan pengetahuan yang diperoleh daripada pohon keputusan yang lebih bermakna. Saiz pohon keputusan adalah 8 daun dan sejumlah 12 simpul. Jadual 1.21 menunjukkan contoh jadual petua-petua.

Dari petua-petua tersebut menjelaskan graduan yang berumur 50 tahun hingga 60 tahun tidak bekerja. Graduan yang berumur 30 tahun hingga 39 tahun dan 40 tahun hingga 49 tahun akan bekerja. Graduan yang mendapat yang berumur 19 tahun hingga 29 tahun mendapat CGPA 2.00 hingga 3.00 akan bekerja. Seterusnya, graduan yang mendapat yang berumur 19 tahun hingga 29 tahun mendapat CGPA 3.00 hingga 4.00 yang mendaftar untuk melanjutkan pengajian setelah anda menamatkan pengajian telah tidak bekerja. Jika tidak mendaftar akan mendapat pekerjaan sebagai graduan. graduan yang mendapat yang berumur 19 tahun hingga 29 tahun mendapat CGPA 0.00 hingga 2.00 yang pernah ikuti LI/internship/praktikal/attachment semasa di IPT akan tidak bekerja. Jika tidak pernah ikuti LI/internship/praktikal/attachment semasa di IPT akan mendapat pekerjaan.

Jadual 1.21 Contoh jadual petua-petua

No.	Syarat	Keputusan
1.	IF e_umur = 19-29 AND e_cgpa = 2.00-3.00	THEN Status Pekerjaan Graduan = Bekerja
2.	IF e_umur = 19-29 AND e_cgpa = 3.00-4.00 AND e_33 = Tidak	THEN Status Pekerjaan Graduan = Bekerja
3.	IF e_umur = 19-29 AND e_cgpa = 3.00-4.00 AND e_33 = Ya	THEN Status Pekerjaan Graduan = Tidak Bekerja
4.	IF e_umur = 19-29 AND e_cgpa = 0.00-2.00 AND e_17 = Ya	THEN Status Pekerjaan Graduan = Tidak Bekerja
5.	IF e_umur = 19-29 AND e_cgpa = 0.00-2.00 AND e_17 = Tidak	THEN Status Pekerjaan Graduan = Bekerja
6.	IF e_umur = 30-39	THEN Status Pekerjaan Graduan = Bekerja
7.	IF e_umur = 40-49	THEN Status Pekerjaan Graduan = Bekerja
8.	IF e_umur = 50-60	THEN Status Pekerjaan Graduan = Tidak Bekerja

V. KESIMPULAN

Dalam kajian ini, teknik perlombongan data digunakan untuk mengelaskan faktor yang mempengaruhi kebolehpasaran graduan UKM. Teknik klasifikasi yang digunakan seperti J48, naive bayes, regresi logistik dan SVM. Hasil telah menunjukkan bahawa SVM menunjukkan prestasi yang lebih baik dalam data set tahun 2016 dan 2018 berbanding dengan algoritma lain. Selain itu, algoritma regresi logistik menunjukkan prestasi yang lebih baik dalam data set tahun 2017. Makalah ini mengenal pasti beberapa faktor-faktor yang mempengaruhi graduan UKM dalam status sama ada bekerja atau tidak bekerja. Antara faktor ini, atribut umur, CGPA, pernahkah ikuti LI/internship/praktikal/attachment semasa di IPT dan adakah mendaftar untuk melanjutkan pengajian setelah anda menamatkan pengajian mengandungi banyak maklumat dan mempengaruhi tahap akhir graduan, iaitu kebolehpasaran status.

VI. RUJUKAN

Mishra, T., Kumar, D. & Gupta, S. 2016. Students' Employability Prediction Model through Data Mining. *International Journal of Applied Engineering Research* 11(4): 2275-2282.

Keno C.Piad, Menchita Dumlao, Melvin A. Ballera & Shaneth C. Ambat. Predicting IT Employability Using Data Mining Techniques. The Third International Conference on Digital Information Processing, Data Mining, and Wireless Communication, Moscow, Russia, 2016.

Han, J., Pei, J. & Kamber, M. 2011. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. 3rd Ed. Amsterdam. Elsevier.

Nelofar Rehman, Assistant Professor, PG Department Of Computer Science, SSM College Of Engineering and Technology, Kashmir, India. Data Mining Techniques Methods Algorithms and Tools. *IJCSMC, Vol. 6, Issue. 7, July 2017, pg.227-231*.

Gao, L. (2015). Analysis of Employment Data Mining for University Student based on Weka Platform, 2(4), 130-133.

Nor Azziaty, Kian Lam Tan & Chen Kim Lim. Predictive analysis and data mining among the employment of fresh graduate students in HEI. AIP Conference Proceedings 1891(1):020007. October 2017.

Mohd Tajul Rizal & Yuhanis Yusof. Graduates employment classification using data mining approach. AIP Conference Proceedings 1761, 020002 (2016).

Suganthi, G. & Ashok, M.V.2017. PREDICTING EMPLOYABILITY OF STUDENTS USING DATA MINING APPROACH. *International Journal of Information Research and Review 4*(2): 3798-3801.

Sheena, Kumar, K. & Kumar, G. 2016. Analysis of Feature Selection Techniques: A Data Mining Approach. International Journal of Computer Applications & 4th International Conference on Engineering & Technology, hlm. 17-21.

Lior Rokach, Ben-Gurion University of the Negev, Israel & Oded Maimon, Tel-Aviv University, Israel. Data Mining With Decision Trees Theory and Application 2nd Edition. *USA office:* 27 Warren Street, Suite 401-402, Hackensack, NJ 07601.

Sapaat, M. A., Mustapha, A., Ahmad, J., Chamili, K. & Muhamad, R. 2011. A Data Mining Approach to Construct Graduates Employability Model in Malaysia. *International Journal on New Computer Architectures and Their Applications* 1(4): 1086-1098.