Model Analitik Data Kesejahteraan Rakyat dalam Meramal Kesamaan Gender

Siti Najihah Hishamuddin[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*— Malaysia adalah sebuah negara yang sangat menitikberatkan kesejahteraan rakyat yang hidup didalam negara ini. Demi menjadi sebuah negara yang maju, pihak kerajaan telah mengenal pasti semua aspek untuk menjadikan Malaysia sebuah kerajaan yang maju. Antara aspek yang telah diambil kira ialah tahap pendapatan bagi seluruh rakyat Malaysia. Gender merujuk kepada ketetapan dan ketentuan masyarakat terhadap peranan, hubungan, personaliti, kekuasaan dan pengaruh yang dimainkan secara berbeza oleh dua jantina. Merujuk dalam kamus dewan Bahasa, kesamaan merangkumi semua konsep yang mana semua manusia lelaki dan wanita, bebas membentuk keupayaan peribadi mereka dan membuat pilihan tanpa sebarang batasan yang ditetapkan. Secara kesuluruhannya, kesamaan gender membawa maksud setiap tingkahlaku, aspirasi dan keperluan yang berbeza antara lelaki dan wanita dipertimbangkan dan dinilai secara sama rata. Kajian ini dijalankan untuk menghasilkan satu model analitik data kesejahteraan rakyat bagi mengkaji hubungan gender dengan faktor- faktor tahap pendapatan yang lain. Pendekatan perlombongan data dengan teknik pengelasan digunakan dalam kajian ini. Objektif kajian ini bertujuan untuk mengetahui dan meramal serta mengenalpasti perkaitan kesamaan gender dalam mengkelaskan tahap pendapatan isi rumah. Selain itu, kajian ini menggunakan dua kaedah klasifikasi yang digunakan ialah Pohon Keputusan (Decision Tree) dan Naive Bayes. Set data yang digunakan didalam kajian ini merupakan dataset yang diperoleh dari Jabatan Perangkaan Malaysia (DOSM). Dataset ini merupakan data–data dari Kajian Perbelanjaan Isi Rumah (HES) yang telah dijalankan pada tahun 2014.**

Keywords— Pohon Keputusan; Naïve Bayes; Support Vector Machine

I. PENGENALAN

Malaysia merupakan sebuah negara membangun yang amat menitikberatkan kesejahteraan rakyat demi menjadikan negara sebuah negara yang maju. Pelbagai usaha untuk membasmi kemisikinan telah dijalankan oleh Malaysia iaitu sentiasa memberi penekanan terhadap pembangunan Pendidikan dan kemahiran, penjanaan pendapatan, penciptaan pekerjaan dan peruntukan akses kepada keperluan asas seperti elektrik, air bersih, pengangkutan dan perumahan serta jaring keselamatan sosial. Selain itu, Malaysia Bank Data Kemiskinan Negara (eKasih) diwujudkan sebagai maklumat terpusat terperinci mengenai isi rumah yang miskin untuk memberikan kelebihan kepada pihak kerajaan tentang profil dan membantu penargetan. Di Malaysia, Kerajaan telah menetapkan tiga kelas kategori mengikut tahap pendapatan isi rumah iaitu ; Bawah 40% (B40), Tengah 40% (M40), dan Tertinggi 20% (T20). Klasifikasi tahap pendapatan berdasarkan tiga petunjuk utama; isi rumah, negeri, dan pendapatan strata.

Perlombongan data mengkaji algoritma dan paradigma komputan yang membolehkan komputer mencari corak-corak dan kekerapan dalam pangkalan data, menjalankan anggaran dan ramalan serta secra amnya, meningkatkan prestasi melalui interaksi dengan data. Pada masa kini, ia dianggap sebagai kunci elemen bagi proses am yang dikenali sebagai Penemuan Pengetahuan yang berurusan dalam mengekstrak pengetahuan yang berguna dalam data mentah. Proses Penemuan Pengetahuan ini meliputi pemilihan data, membersihkan, mengekod, menggunakan kaedah-kaedah statistikal yang berbeza, pengenalpastian corak, teknik-teknik pembelajaran mesin, melapor dan visualisasi struktur yang telah dihasilkan

II. KAJIAN KESUSASTERAAN

Kajian yang pertama merupakan daripada Ruben Thoplan (2014) yang bertajuk “Random

Forest for Poverty Classification". Kajian ini bertujuan untuk mengklasifikasikan rakyat mengikut tahap kemiskinan. Kajian ini menggunakan teknik perlombongan data untuk menangani isu klasifikasi kemiskinan di Mauritus. Algoritma hutan rawak (Random Forest) digunakan untuk data banci bagi melihat peningkatan ketepatan klasifikasi status kemiskinan. Analisis daripada kajian ini telah menunjukkan bahawa terdapat beberapa faktor penting dalam mengklasifikasikan status kemiskinan individu. Hasil kajian ini menunjukkan bahawa masih wujudnya jurang antara gender dan tahap kemiskinan ; lelaki diklasifikasikan untuk tidak miskin jika dibandingkan dengan wanita. Antara faktor – faktor yang telah diambil kira dalam kajian ini ialah jumlah waktu bekerja, umur, tahap Pendidikan dan jantina. (Ruben Thoplan, 2014)

Kajian seterusnya merupakan kajian dari beberapa pengkaji dari Universiti Kebangsaan Malaysia yang bertajuk "Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification". Kajian ini menggunakan dataset dari negara Malaysia. Kajian ini bertujuan untuk mengetahui model pembelajaran mesin yang sesuai untuk mengklasifikasi B40. Kajian ini menggunakan Naïve Bayes, Pohon Keputusan dan algorithm k-Nearest Neighbours. Hasil dari kajian menunjukkan bahawa klasifikasi B40 menunjukkan keputusan yang terbaik bila menggunakan Pohon Keputusan (J48) daripada menggunakan Naïve Bayes dan k-Nearest Neighbor. (Nor Samsiah Sani, Mariah Abdul Rahman, Azuraliza Abu Bakar, Shahnorbanun Sahran, Hafiz Mohd Sarim, 2018)

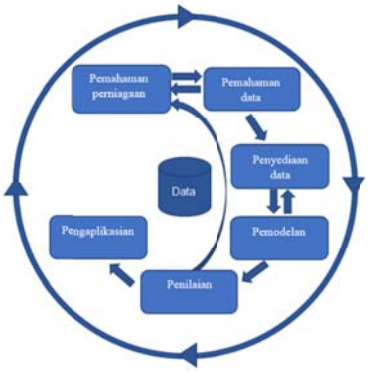
Kajian ini daripada Linden McBride dan Austin Nichols yang bertajuk "Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools" (January 2015). Kajian ini diadakan bertujuan untuk mengemukakan bukti bahawa penggunaan algoritma pembelajaran mesin untuk pembangunan PMT secara substansial dapat memperbaiki prestasi luar alat ini. Hasil membuktikan bahawa teknik pohon rawak menunjukkan keputusan yang terbaik berbanding model pembelajaran mesin yang lain. (Linden McBride dan Austin Nichols, 2015)

Jadual 1: Kajian Literasi berkaitan Pembelajaran Mesin

No	Tahun	Penulis	Tajuk
1.	2014	Ruben Thoplan	"Random Forest for Poverty Classification"
2.	September 2018	Nor Samsiah Sani, Mariah Abdul Rahman, Azuraliza Abu Bakar, Shahnorbanun Sahran, Hafiz Mohd Sarim	"Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification".
3.	January 2015	Linden McBride dan Austin Nichols	"Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools

III. KAEDAH KAJIAN

Fasa yang terlibat dalam kajian ini ialah fasa perlombongan data (CRISP DM). Bagi fasa ini, pendekatan analitik data CRISP-DM digunakan untuk memperoleh ramalan. Metodologi analisis data CRISP-DM terdiri daripada enam fasa utama iaitu pemahaman perniagaan, pemahaman data, pra-proses dan penyediaan data, pembangunan model, penilaian model dan pengaplikasian seperti Rajah 1.



Rajah 1: Model Data Mining CRISP-DM

A. Pemahaman Perniagaan

Memahami matlamat perniagaan untuk menganalisis data adalah tugas utama di dalam kajian ini dimana kita akan menggariskan matlamat dan soalan perniagaan yang berikut. Matlamat perniagaan kajian ini adalah untuk mengenalpasti perkaitan kesamaan gender dalam mengkelaskan tahap pendapatan isi rumah rakyat Malaysia dan mengetahui attribut yang boleh digunakan selain daripada jumlah pendapatan.

Dalam kajian ini, kelas pendapatan di Malaysia diklasifikasikan menggunakan satu attribute utama iaitu gender dan beberapa attribute lain yang mempunyai kaitan dengan gender.

Berdasarkan data dari Kajian Pendapatan dan Perbelanjaan Isi Rumah (HES) pada tahun 2014 yang telah dijalankan oleh Jabatan Perangkaan Malaysia (DOSM), hubungan anggota isi rumah boleh dikategorikan kepada dua belas kategori seperti yang ditunjukkan dalam Jadual 2. Berdasarkan kajian sebelum ini, mereka mengurangkan kategori asal kepada empat kategori yang mungkin diwakili sebagai RHH1, RHH2, RHH3, RHH4 seperti yang ditunjukkan dalam Jadual 2.

Jadual 2 Hubungan dengan ketua isi rumah

Original Category	Reduced Category (RHH-Relationship with Household Head)
01 Household Head 02 Head's Spouse 03 Unmarried Children	RHH1
04 Head's Married Children 05 Head's Daughter/Son In Law 06 Head's Grandchildren	RHH2
07 Head's Parent. Parent in law 08 Head's Grandparents/ grandparents in law	RHH3
09 Head's siblings/ sibling in law 10 Head's or wife's relatives 12 Other individual	RHH4

B. Pemahaman Data

Data Kajian Pendapatan dan Perbelanjaan Isi Rumah diperoleh dari Jabatan Perangkaan Malaysia. Dataset dari tiga pangkalan data adalah (i) 1,058,574 contoh perbelanjaan Isi Rumah, (ii) 64,091 kes dan 14 sifat ahli isi rumah. (iii) 14,838 kes dan 10 sifat kepala rumah dan barang. Set data diperoleh melalui persampelan bertingkat yang meliputi 14 negeri dan 3 negeri persekutuan Malaysia. Set data disatukan yang mengakibatkan dataset tunggal mengandungi 14,838 contoh dan 14 sifat. Jadual 3 dan 4 menunjukkan perihalan data isi rumah (ketua) dan penerangan atribut isi rumah (ahli) masing-masing.

Jadual 3: Attribut Data Ketua Isi Rumah

Num	Atribut	Penerangan
1.	ID Isi Rumah (numerik)	Id isi rumah
2.	No isi rumah (numeric)	
3.	Negeri (nominal)	
4.	Strata (nominal)	1 Bandar; 2 Luar Bandar
5.	Kawasan(nominal)	1Semenanjung Malaysia ; 2 Sabah and W.P.Labuan 3 Sarawak
6.	Etnik	1 Bumiputera;

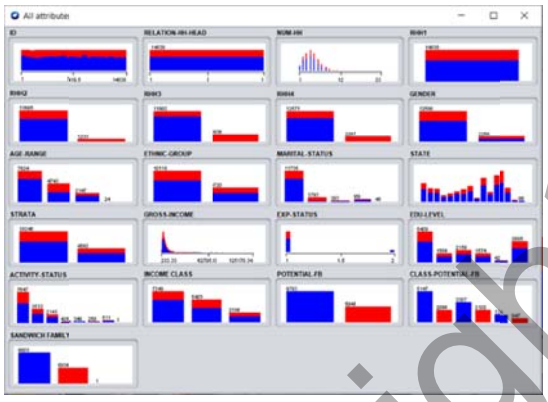
	(nominal)	2 Bukan Bumiputera
7.	Bilangan Isi Rumah (numerik)	
8.	Pemberat Data (numeric)	
9.	Jumlah Perbelanjaan Bulanan (01-12) (numerik)	Bulanan
10.	Junlah Pendapatan Kasar Bulanan (numerik)	Bulanan

Jadual 4 : Attribut Data Ahli Isi Rumah

N	Attribute	Details
1	ID Isi Rumah (numeric)	ID Isi Rumah
2	Bil. Ahli Isi Rumah (numerik)	
3	Perhubungan dengan Ketua IR (kategori)	01 Ketua isi rumah 02 Isteri / suami ketua 03 Anak ketua yang belum berkahwin 04 Anak ketua yang telah berkahwin 05 Menantu perempuan /lelaki ketua 06 Cucu Ketua 07 Bapa / ibu ketua atau isteri /suami ketua 08 Datuk / nenek ketua atau kepada isteri / suami ketua 09 Abang / kakak / adik ketua atau kepada isteri / suami ketua 10 Orang lain yang bersaudara dengan ketua atau isteri / suami ketua 11 Pembantu rumah 12 Orang lain yang tidak bersaudara dengan ketua atau isteri / suami ketua
4	Gender (nominal)	Male; Female
5	Age (numerik)	Single age
6	Etnik (nominal)	1Bumiputera; 2 Bukan-Bumiputera
7	Kewarganegaraan (nominal)	1 Warganegara; 2 Bukan-Warganegara
8	Status Perkahwinan (nominal)	1 Tidak pernah berkahwin 2 Berkahwin 3 Balu / Duda 4 Bercerai 5 Berpisah
9	Tahap Pendidikan (Kategori)	1 Tiada Pendidikan Rasmi; 2 Rendah; 3 Menengah; 4 Tertiari
10	Pekerjaan (Kategori)	01Pengurus; 02 Profesional; 03 Juruteknik dan profesional bersekutu; 04 Pekerja sokongan perkeranian; 05 Pekerja perkhidmatan dan jualan; 06 Pekerja mahir pertanian, perhutanan dan perikanan; 07 Pekerja kemahiran dan pekerja pertukangan yang berkaitan; 08 Operator loji dan mesin serta pemasangan; 09 Pekerjaan asas; 00 Pekerjaan yang tidak dikelaskan dimana-mana; 00006=Penganggur; 00007=Suri rumah/menjaga rumah; 00008=Pelajar; 00009=Pesara; 00010=Lain-lain (terangkan); 00011=Kanak-kanak tidak bersekolah
11	Sijil tertinggi yang diperoleh (Kategori)	1 Ijazah/Diploma Lanjutan; 2 Diploma / Sijil; 3 STPM; 4 SPM / SPMV; 5 PMR / SRP; 6 Tiada Sijil
12	Status Aktiviti (Kategori)	01 Majikan 02 Pekerja Kerajaan 03 Pekerja Swasta 04 Bekerja Sendiri 05 Pekerja keluarga tanpa

		gaji 06 Penganggur 07 Suri rumah/menjaga rumah 08 Pelajar 09 Pesara 10 Lain-lain 11 Kanak-kanak tidak bersekolah
13	Terima Pendapatan (nominal)	1 Ya 2 Tidak
14	Industri (nominal)	Kumpulan utama (1 digit) mengikut mengikut "Piawaian Klasifikasi Industri Malaysia (MSIC)" 2008

Rajah 2 Semua attribute dari set data HES2014



C. Penyediaan Data

Fasa penyediaan data melibatkan empat proses utama. Empat proses tersebut merupakan pembersihan data, pengurangan data, transformasi data dan pelabelan data. Tujuan fasa ini dilakukan adalah untuk memastikan kesemua set data bersih dan sedia digunakan dalam fasa yang seterusnya iaitu fasa perlombongan data. Fasa ini adalah fasa yang paling penting kerana ia sangat berguna untuk memperolehi pengetahuan yang banyak sebelum memasuki fasa perlombongan data. 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).

Fasa penyediaan data melibatkan empat proses utama. Empat proses tersebut merupakan pembersihan data, pengurangan data, transformasi data dan pelabelan data. Tujuan fasa ini dilakukan adalah untuk memastikan kesemua set data bersih dan sedia digunakan dalam fasa yang seterusnya iaitu fasa perlombongan data. Fasa ini adalah fasa yang paling penting kerana ia sangat berguna untuk memperolehi pengetahuan yang banyak sebelum memasuki fasa perlombongan data. 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).

D. Pemodelan

Fasa ini merupakan fasa pemilihan model yang akan digunakan dalam kajian ini. Terdapat beberapa teknik permodelan yang boleh dipilih. Ujian pemilihan teknik model ini akan dilakukan berulang kali untuk memastikan teknik model yang digunakan merupakan teknik model yang terbaik untuk dataset yang tersedia.

Kajian ini akan menghasilkan model klasifikasi. Terdapat pelbagai kaedah klasifikasi yang boleh digunakan. Tiga kaedah klasifikasi yang terkenal digunakan iaitu Decision Tree (DT) khususnya algoritma J48, Multilayer Perceptron (MLP), varian rangkaian neural buatan, dan Naïve Bayes, pendekatan statistik. Prestasi kaedah ini dibandingkan dalam tiga metrik, ketepatan klasifikasi, kesilapan akar min kesilapan (RMSE), dan kawasan lengkung ROC. Ketepatan klasifikasi ditentukan berdasarkan peratusan data ujian yang betul diklasifikasikan oleh model. Pemilihan model klasifikasi akan dibuat melalui ketepatan model dan ramalan dimana model yang mempunyai ketetapan yang paling tinggi akan dipiih untuk menjadi model analisis data.

Decision Tree biasanya digunakan untuk mendapatkan maklumat untuk tujuan membuat keputusan. Teknik ini bermula dengan nod akar supaya pengguna dapat mengambil tindakan. Daripada nod ini, pengguna dapat membahagi setiap nod secara rekursif menurut algoritma Decision Tree. Naïve Bayes membuat ramalan dengan menggunakan Bayes' Theorem, yang menghasilkan kebarangkalian ramalan dari bukti asas seperti yang

diperhatikan dalam data. Secara ringkas, ia menganggap bahawa kehadiran (atau ketiadaan), ciri tertentu sesuatu kelas adalah tidak berkaitan dengan kehadiran (atau ketiadaan) apa-apa ciri lain. Support Vector Machine (SVM) adalah model pembelajaran mesin yang menggunakan algoritma untuk masalah klasifikasi dua kumpulan. Setelah memberikan model SVM set data latihan berlabel untuk setiap kategori, mereka dapat mengkategorikan teks baru.

E. Penilaian

Di dalam fasa ini, kita akan menilai semula hasil untuk memastikan bahawa semua matlamat akan dicapai. Fasa ini merupakan fasa yang paling penting. Terdapat dua proses yang akan dibuat di fasa ini. Pertama, hasil penilaian ringkasan dari segi kriteria kejayaan perniagaan, termasuk kenyataan akhir mengenai sama ada projek itu sudah memenuhi objektif perniagaan awal. Seterusnya, model akan diluluskan diluluskan dalam fasa ini dimana model yang diluluskan merupakan model yang telah dinilai dan memenuhi kriteria kejayaan perniagaan.

Setelah itu, tinjauan proses akan dilakukan. Tinjauan ini bertujuan untuk merumuskan proses kajian dan menyerlahkan aktiviti yang telah terlepas dan yang perlu diulang atas sebab – sebab tertentu.

F. Pengaplikasian

Fasa ini merupakan fasa yang keenam dan yang terakhir didalam CRISP-DM. Melalui fasa ini, semua keputusan penilaian akan diambil dan strategi pengaplikasian akan diringkaskan. Hal ini termasuk langkah-langkah yang perlu dan bagaimana untuk melaksanakannya. Seterusnya, perancangan bagi pemantauan dan penyelenggaraan aplikasi. Pemantauan dan penyelenggaraan adalah isu penting jika keputusan perlombongan data menjadi sebahagian daripada perniagaan sehari-hari dan persekitarannya. Untuk memantau penggunaan hasil perlombongan data, projek ini memerlukan pelan proses pemantauan secara terperinci.

IV. KEPUTUSAN DAN PERBINCANGAN

A. Pengujian Model

Tiga kaedah klasifikasi yang terkenal digunakan iaitu, Pohon Keputusan(Decision Tree), varian rangkaian neural buatan (NN) dan Naive Bayes, pendekatan statistik. Prestasi bagi setiap kaedah klasifikasi dibandingkan berdasarkan ketetapan klasifikasi.

Jadual 5 menunjukkan perbandingan ketetapan purata bagi setiap algoritma. Berdasarkan jadual, MLP menunjukkan nilai ketetapan tertinggi iaitu 69.2277%, diikuti dengan nilai Naïve Bayes sebanyak 60.2305% dan DT sebanyak 60.1361%. Melalui keputusan ini, dapat kita rumuskan bahawa nilai ketetapan setiap algoritma adalah lebih kurang iaitu dalam lingkungan 60 peratus hingga 70 peratus.

Jadual 6 pula menunjukkan pebandingan antara beberapa metrik yang lain seperti nilai RMSE dan nilai ROC bagi setiap algoritm. Prestasi ketetapan purata yang didapati didalam kajian ini adalah disebabkan set data yang telah digunakan. Teknik DT mendapat nilai RMSE tertinggi dengan nilai 0.2976, diikuti dengan Naive Bayes sebanyak 0.29 serta MLP sebanyak 0.2662. Bagi nilai ROC, MLP menunjukkan nilai tertinggi sebanyak 0.925, diikuti dengan NB sebanyak 0.892 dan DT sebanyak 0.871. Rajah 2 menunjukkan graf perbandingan nilai ketetapan purata, nilai RMSE dan ROC.

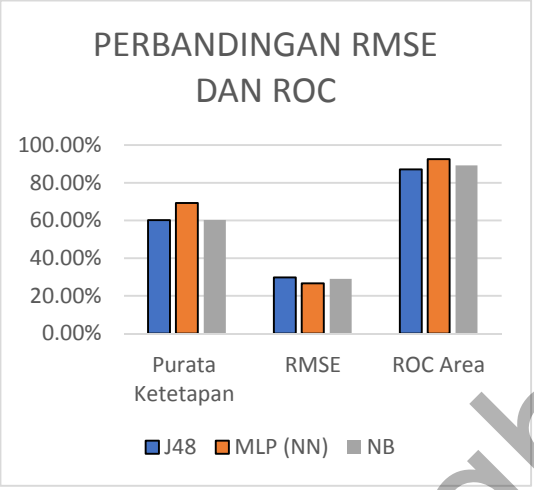
Hasil daripada keputusan kajian ini, algoritm DT (J48) telah dipilih disebabkan visual DT yang mudah difahami dan senang untuk diterjemahkan.

Jadual 5 Perbandingan ketetapan purata

Algoritm	Ketetapan purata
J48	60.1361 %
Multilayer perceptron (Neural network)	69.2277 %
Naïve bayes	60.2305 %

Jadual 6 Perbandingan nilai ROC dan RMSE

algoritm / metrik	J48	MLP (NN)	NB
Ketetapan	60.1361 %	69.2277 %	60.2305 %
RMSE	0.2976	0.2662	0.29
ROC Area	0.871	0.925	0.892



Rajah 2 Graf perbandingan nilai ketetapan purata, RMSE dan ROC

B. Analisis Pengetahuan

Untuk memastikan bahawa output model ramalan dapat ditukar menjadi wawasan yang dapat dilakukan, adalah penting untuk memahami faktor-faktor mana yang menyumbang paling banyak kepada ramalan. Dalam bahagian ini, pengetahuan yang diekstrak dari model Decision Tree dianalisis selanjutnya.

Decision Tree dikenali sebagai pohon pengelasan jika ia digunakan dalam tugas pengelasan. Pohon Keputusan juga dikenali sebagai pohon regresi jika ia digunakan untuk tugas regresi. Pohon pengelasan biasanya digunakan dalam sektor kewangan, pemasaran, kejuruteraan dan perubatan (Rokach & Maimon 2014). Kajian ini menggunakan teknik pokok keputusan kerana lebih mudah untuk memahami dan mudah mentafsir secara visual. Teknik ini mudah untuk dijelaskan kepada sesiapa terutamanya kepada eksekutif atau pembuat keputusan. Pokok keputusan juga menyediakan indikator yang jelas daripada susunan kepentingan faktor untuk membuat keputusan dan ramalan. Jadual 7 akan menunjukkan contoh-contoh petua yang berada didalam model ini.

Beberapa sub pohon telah dipilih berdasarkan attribute kesamaan gender yang telah dibincang. Semasa perbincangan, attribute kesamaan gender yang telah dipilih adalah tahap gender, umur, pekerjaan, tahap Pendidikan, strata dan status perkahwinan. Rajah yang ditunjukkan dibawah merupakan rajah yang telah diekstrak daripada pohon keputusan yang sebenar. Sub pohon keputusan yang pertama ialah contoh sub pohon keputusan untuk kelas tidak berisiko Bk bagi tahap Pendidikan SPM/SPMV. Melalui sub pohon keputusan ini, golongan ini kebanyakan dikelompokkan didalam kelas B40 bagi yang belum berkahwin dan M40 untuk golongan yang berkahwin ataupun bercerai di negeri Kedah. Manakala bagi negeri Johor, golongan yang belum berkahwin dikelompokkan kedalam kelas B40. Untuk golongan yang telah berkahwin, sub pohon telah dibahagikan mengikut etnik dan gender. Bagi gender lelaki yang bumiputera, mereka dikelaskan didalam kelas pendapatan B40 manakala bagi wanita bumiputera adalah didalam kelas T20. Bagi golongan lelaki bukan bumiputera pula, kesemuanya dikelaskan didalam kelas M40 dan wanita didalam kelas B40.

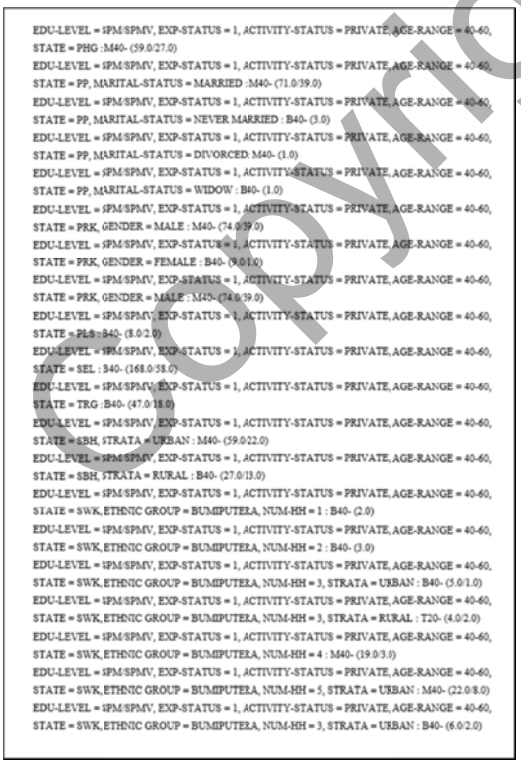
Seterusnya, contoh sub pohon keputusan untuk kelas bukan berisiko Bk bagi tahap pendidikan diploma atau sijil. Saya telah memilih julat umur dari 40 hingga 60 tahun daripada negeri Kedah, Pulau Pinang dan Sabah. Dari pemerhatian terhadap sub pohon ini, jantina lelaki di Pulau Pinang merupakan kelas M40 manakala bagi gender wanita, kebanyakannya berada di kelas B40. Bagi di Kedah pula, kelas pendapatan ditentukan berdasarkan etnik bagi yang sudah berkahwin, manakala yang bagi golongan yang tidak berkahwin, golongan ini terus dikelaskan ke dalam kelas B40. Di Sabah, dapat kita lihat golongan yang bercerai, janda dan golongan yang sudah berpisah, golongan ini telah dikelaskan dalam kelas M40. Bagi yang telah berkahwin di Sabah, ia dibahagikan lagi mengikut etnik dimana bagi etnik bumiputera, mereka dikelaskan kepada kelas M40 dan B40 bagi golongan bukan bumiputera.

Berikut merupakan sub pohon untuk kelas berisiko Bk bagi tahap Pendidikan Sarjana Muda

dan Diploma Tinggi. Sub pohon ini mempunyai petua jumlah isi rumah dan juga beberapa negeri. Melalui sub pohon ini, kebanyakan golongan yang telah berkahwin adalah didalam kelas T20 BK bagi beberapa negeri seperti Johor, Negeri Sembilan dan Pulau Pinang. Hal ini kerana negeri – negeri tersebut merupakan negeri yang mempunyai jumlah pendapatan yang tinggi dan merupakan kawasan bandar.

Contoh sub pohon keputusan yang terakhir ialah untuk kelas berisiko Bk bagi tahap pendidikan Diploma atau Sijil. Berdasarkan sub pohon, dapat diperhatikan setiap sub pohon dibahagikan mengikut status perkahwinan iaitu berkahwin, tidak berkahwin, bercerai dan janda. Selepas itu, mereka dibahagi mengikut negeri. Bagi yang telah berkahwin di negeri Perak, Perlis, Selangor, Terengganu, Sabah dan WP Kuala Lumpur, masing- masing dikelaskan ke kelas M40-BK manakala bagi negeri Sarawak dikelaskan ke kelas T20-BK. Bagi golongan yang telah bercerai dan tidak pernah berkahwin, keduanya dikelaskan kedalam kelas M40-BK. Bagi golongan janda, terdapat dua negeri yang berada di dalam kelas M40 iaitu di Kelantan dan WP Kuala Lumpur, manakala golongan janda yang berada di Sarawak berada dalam kelas T20.

Rajah 3 Contoh rule prune



Jadual 7 : Contoh jadual petua- petua

No.	Syarat	Keputusan
1.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = Pahang	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeri Pahang dikelompokkan ke dalam kelas M40 (BUKAN BK)
2.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = Pahang, Marital Status = Never Married	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeriPulau Pinang bagi yang belum berkahwin dikelompokkan ke dalam kelas B40 (BUKAN BK).
3.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = Pahang, Marital Status = Divorced	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeriPulau Pinang bagi yang bercerai dikelompokkan ke dalam kelas M40 (BUKAN BK)
4.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = Pahang, Marital Status =Widow	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeriPulau Pinang bagi golongan janda dikelompokkan ke dalam kelas

		M40 (BUKAN BK)
5.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = PRK, Gender = Male	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeri Perak bagi lelaki dikelompokkan ke dalam kelas M40 (BUKAN BK)
6.	Edu-level = SPM/SPMV, Activity status = private, age range = 40 – 60, State = PRK, Gender = Female	Ketua isi rumah yang mempunyai tahap Pendidikan SPM/SPMV dan berumur diantara 40- 60 didalam negeri Perak bagi wanita dikelompokkan ke dalam kelas B40 (BUKAN BK)

V. KESIMPULAN

Kerajaan Malaysia sentiasa menitikberatkan tahap kesejahteraan rakyat. Tahap kemiskinan di Malaysia telah menurun tapi ia masih memerlukan penambahbaikan supaya pihak kerajaan dapat mengenalpasti rakyat – rakyat yang memerlukan dan boleh mengklasifikasikan rakyat di dalam kelas pendapatan isi rumah yang betul tanpa melihat kepada pendapatan bulanan dan status starata rakyat semata. Dengan teknik perlombongan data ini, proses klasifikasi dan ramalan akan lebih berjaya selepas ini.

Setiap kajian yang dilakukan pastinya mempunyai kelebihan dan kekurangan sepanjang kajian dilakukan. Melalui kajian ini, kelebihan kajian ini ialah pihak kerajaan dapat membuat satu sasaran baru berkaitan tahap pendapatan menggunakan kesamaan gender. Gender juga merupakan satu perkaitan yang amat penting bila dikaitkan dengan tahap pendapatan seseorang dan juga tahap kemiskinan.

Kajian ini hanya akan menggunakan 3 teknik pengelasan perlombongan data sahaja iaitu Pohon Keputusan (Decision Tree) dan Naive Bayes, pendekatan statistik serta varian rangkaian neural buatan (NN) dan hanya menumpukan beberapa attribute yang telah dipilh. Teknik perlomobongan data yang juga berasaskan klasifikasi boleh dilaksanakan untuk mendapatkan banyak lagi pengetahuan yang berkaitan

Mengenai ketetapan ramalan berdasarkan model, didapati keputusan ketetapan tidak seperti yang dijanjikan. Hal ini kerana attribute yang digunakan didalam kajian ini sedikit berbeza dengan kajian sebelum ini.

Kesimpulannya, proses yang dilakukan iaitu proses perlombongan data amat penting dalam kajian ini. Dalam kajian ini, WEKA telah digunakan dalam fasa pra pemprosesan data. Ini adalah kerana Weka adalah kumpulan algoritma pembelajaran mesin untuk tugas-tugas perlombongan data. Ia mengandungi alat untuk penyediaan data, klasifikasi, regresi, kluster, peraturan pertambangan, dan visualisasi. Semua fasa yang diterangkan didalam bab ini merupakan fasa yang amat penting sebelum fasa perlombongan data diteruskan. Diakhir kajian ini, diharapkan kajian ini dapat menyelesaikan tujuan utama kajian ini dilakukan.

VI. RUJUKAN

Department of Economics and Statistics, Faculty of Social Studies and Humanities, University of Mauritius, Réduit, Mauritius

Applications- 2nd Edition. World Scientific Publishing Co. Pte. Ltd. Singapore

Arrangements and Service Expenditures on Female and Male Tasks. Social Forces, Volume 84. University of Groningen.

Casper, Lynne M.; MacLanahan, Sara S.; Garfinkel, Irwin. 1994. *The Gender Poverty Gap:*

Economic Planning Unit. June 2017. *Malaysia Sustainable Development Goals Voluntary*

Han, Jiawei , Micheline Kamber, Jian Pei. 2012. *Data mining : concepts and techniques –*

3rd ed. United States of America. Morgan Kaufmann Publishers.

learning: An application to the USAID Poverty Assessment Tools". Cornell University, New York.

Linden McBride & Austin Nichols. Jan 2015. "*Improved poverty targeting through machine*

Lior Rokach, Oded Maimon. 2014 . *Data Mining with decision trees : Theory and*

Measuring Gender Differences Using Nonmonetary Indicators". University of Oxford, Glasgow Caledonian University.

Mohd Sarim. September 2018. "*Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification*". Universiti Kebangsaan Malaysia, Malaysia.

National Review 2017. Jabatan Perdana Menteri, Putrajaya, Malaysia.

Nations Development Program.

Nilufer Cagatay. May 1998. "*Gender and Poverty*". Working Paper Series. United

Nor Samsiah Sani, Mariah Abdul Rahman, Azuraliza Abu Bakar, Shahnorbanun Sahran, Hafiz

Ruben Thoplan . 2014. *Random Forest for Poverty Classification.*

Ruijter, E. D., Treas, J. K., & Cohen, P. N. 2005. *Outsourcing the Gender Factory: Living*

Sara Cantilon, Brian Nolan. October 2015. *"POVERTY WITHIN HOUSEHOLDS:*

What Can We Learn From Other Countries?, LIS Working Paper Series, No. 112, Luxembourg Income Study (LIS), Luxembourg