

MODEL PEMBELAJARAN MESIN DALAM MERAMAL TINGKAHLAKU PERBELANJAAN KE ATAS KELAS PENDAPATAN ISI RUMAH MALAYSIA

NUR HAZWANI BINTI SHAMSUDIN

PROF. DR. AZURALIZA ABU BAKAR

*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor Darul Ehsan, Malaysia*

ABSTRAK

Kesejahteraan kewangan adalah penting bagi individu, masyarakat, dan negara, merangkumi pelbagai faktor seperti tahap pendapatan, keselamatan kerja, aspek kualiti kehidupan, dan akses kepada penjagaan kesihatan dan pendidikan untuk mencapai kestabilan kewangan. Kajian ini dibuat untuk menangani isu perbelanjaan berlebihan, yang boleh menyebabkan ketidakstabilan kewangan, terutamanya di kalangan individu dengan literasi kewangan rendah yang berisiko terdedah kepada kerentanan kewangan. Ia menekankan kepentingan peningkatan literasi kewangan dalam meredakan risiko tersebut dan memperbaiki tingkah laku kewangan. Untuk mengatasi cabaran ini, kajian ini menggunakan lima algoritma pembelajaran mesin iaitu *Decision Tree*, *Random Forest*, *Support Vector Machines*(SVM), *K-Nearest Neighbour* (kNN), dan *Naïve Bayes* untuk membangunkan model klasifikasi tingkah laku perbelanjaan isi rumah yang berbelanja secara berlebihan. Dengan menggunakan data Tinjauan Pendapatan dan Perbelanjaan Isi Rumah Malaysia 2019 yang dikutip daripada kajian yang dijalankan oleh pihak Jabatan Perangkaan Malaysia(DOSM), model ini telah dibangunkan menggunakan 12 kategori isi rumah yang merangkumi 16,354 rekod isi rumah. Model ini bertujuan untuk mengenal pasti corak perbelanjaan penting yang menyumbang kepada perbelanjaan berlebihan di kalangan kelas pendapatan yang berbeza. Dengan memanfaatkan pemodelan ramalan, pembuat dasar dan institusi kewangan dapat merancang program pendidikan kewangan yang berorientasikan serta mekanisme sokongan untuk populasi yang rentan. Model yang dihasilkan dijangkakan dapat diaplikasikan dalam sistem sokongan keputusan bagi membantu pihak berkepentingan membuat keputusan dalam situasi tingkah laku perbelanjaan isi rumah di kalangan kelas pendapatan B40, M40 dan T20 yang berbelanja berlebihan. Selain itu, hasil yang dijangkakan dari kajian ini juga dapat meningkatkan pemahaman menyeluruh tentang faktor-faktor yang menyebabkan perbelanjaan berlebihan dan impaknya terhadap kestabilan kewangan, serta dapat mengenal pasti perkara yang harus diberi tumpuan untuk intervensi dan peningkatan dalam berbelanja. Dengan menekankan kepentingan literasi kewangan dan peranannya dalam mempromosikan corak perbelanjaan yang lebih sihat, penyelidikan ini menyumbang kepada perbincangan yang lebih luas mengenai kesejahteraan kewangan dan pertumbuhan ekonomi yang mampan.

PENGENALAN

Kesejahteraan kewangan merupakan perkara penting yang dititikberatkan bagi individu, masyarakat dan negara. Secara umumnya, kesejahteraan merangkumi pelbagai aspek kualiti hidup secara keseluruhan termasuk tahap pendapatan, keselamatan kerja, kemudahan perumahan, kualiti taraf hidup, akses penjagaan kesihatan, kemudahan pendidikan, persekitaran dan ikatan sosial, dan lain-lain. Meningkatkan literasi kewangan boleh membawa kepada kesejahteraan kewangan. Literasi kewangan ataupun kecekapan kewangan adalah keupayaan untuk memahami dan menggunakan pelbagai kemahiran kewangan dengan berkesan termasuk pengurusan kewangan peribadi, memperuntukkan wang dalam belanjawan serta melabur. Tingkah laku kewangan, yang melibatkan perancangan, penjekakan kos, rutin penjimatan, dan memenuhi keperluan asas, adalah faktor lain yang menentukan literasi kewangan. Tingkah laku kewangan yang lebih baik dapat mengurangkan risiko muflis dan keputusan kewangan yang berisiko tinggi serta corak perbelanjaan yang lebih sihat semuanya boleh berlaku disebabkan oleh peningkatan literasi kewangan.

Tingkah laku kewangan termasuk meminjam, berbelanja, melabur, dan tabiat menyimpan boleh memberi impak kepada kestabilan kewangan dan kesejahteraan individu serta masyarakat. Faktor-faktor yang mempengaruhi dan membentuk tingkah laku kewangan termasuk faktor psikologi dan emosi, norma budaya dan sosial, serta akses kepada sumber-sumber kewangan dan pendidikan dapat mendorong kepada kestabilan kewangan dan pertumbuhan ekonomi. Pembelanjaan berlebihan, di sisi lain, adalah sebaliknya; ia merupakan tingkah laku kewangan yang mengakibatkan hutang dan ketidakstabilan kewangan. Pembelanjaan berlebihan adalah istilah yang digunakan untuk menggambarkan situasi di mana seseorang individu berbelanja lebih daripada pendapatan mereka. Keperluan asas keluarga boleh menjadi punca pembelanjaan berlebihan tetapi ia juga adalah hasil daripada gaya hidup yang melibatkan kekurangan literasi kewangan, sikap kritis terhadap gaya hidup, dan ketidakupayaan untuk mengukur diri sendiri untuk ukur baju di badan sendiri. Fenomena pembelanjaan berlebihan ini perlu ditangani kerana ia boleh memberi kesan kepada semua peringkat pendapatan terutamanya individu yang berpendapatan rendah lebih terkesan kerana sumber dan pendidikan kewangan yang terhad.

Menggunakan pembelajaran mesin untuk meramalkan literasi kewangan dapat memberikan maklumat yang lebih terperinci tentang perbezaan dan persamaan di kalangan kumpulan pendapatan yang berbeza. Menggabungkan pembelajaran mesin dengan model parametrik boleh membantu mengenal pasti individu literasi kewangan rendah yang berisiko terdedah kepada kerentanan kewangan. Kajian terdahulu di Malaysia telah menggunakan kaedah pembelajaran mesin, seperti *Support Vector Machines*(SVM), *Decision Tree*(J48), dan *Artificial Neural Network*(ANN) untuk memodelkan tingkah laku pembelanjaan dan beban kewangan di kalangan kelas pendapatan (Othman et al., 2020). Kajian terdahulu yang lain juga telah menggunakan kaedah pembelajaran mesin, seperti *Naïve Bayes*, *Decision Tree*(J48), dan *K-Nearest Neighbour*(KNN) untuk memodelkan tingkah laku pembelanjaan dan beban kewangan yang hanya memberi tumpuan kepada kelas pendapatan b40 (Sani et al., 2018). Model ramalan yang diperolehi akan membolehkan pembuatan dasar dan institusi kewangan untuk melabur dalam program pendidikan kewangan dan menyediakan sokongan dan sumber. Dengan itu, individu yang berisiko mengalami kerentanan kewangan boleh mengambil langkah proaktif, seperti memperuntukkan wang dalam belanjawan dan mengurangkan hutang untuk memperbaiki tingkah laku kewangan mereka.

Oleh itu, kajian ini dijalankan untuk membangunkan model klasifikasi tingkah laku perbelanjaan isi rumah yang berbelanja secara berlebihan menggunakan lima kaedah algoritma iaitu *Decision Tree*, *Random Forest* (Model rujukan), *Support Vector Machines*(SVM), *K-Nearest Neighbour*(KNN), dan *Naïve Bayes*. Model ini meramalkan faktor pembelanjaan berlebihan dalam kelas pendapatan yang berbeza di Malaysia iaitu tiga kelas pendapatan T20, M40, dan B40 serta mengenal pasti perkara-perkara perbelanjaan penting yang menyumbang kepada pembelanjaan berlebihan sama ada ia dibelanjakan untuk keperluan asas atau sebaliknya. Kajian dalam mengklasifikasikan tingkah laku pembelanjaan isi rumah dan literasi kewangan ini menggunakan data tinjauan daripada Tinjauan Pendapatan dan Perbelanjaan Isi Rumah Malaysia 2019 dari Jabatan Perangkaan Malaysia (DOSM) yang merangkumi 16,354 rekod isi rumah.

METODOLOGI KAJIAN

Model CRISP-DM (The Cross Industry Standard Process for Data Mining) merupakan metodologi yang digunakan dalam kajian ini. Metodologi CRISP-DM ini merupakan model proses yang mempunyai enam fasa yang melabelkan kitaran hidup sains data secara semula jadinya (HOTZ, 2024). Ia terdiri daripada enam fasa iaitu Pemahaman Bisnes, Pemahaman Data, Penyediaan Data, Permodelan, Penilaian dan Pelaksanaan. Metodolgi ini dipilih kerana pendekatannya urutan fasa tidak ketat memastikan kajian berjalan lancar dan teratur seta memenuhi setiap spesifikasi yang diperlukan dalam memahami corak pembelanjaan berlebihan di kalangan kelas pendapatan dengan memastikan bahawa setiap fasa dilaksanakan sepenuhnya sebelum fasa seterusnya dilaksanakan.

I. Fasa pemahaman bisnes

Pemahaman bisnes dalam kajian ini adalah untuk mengenal pasti corak pembelanjaan berlebihan di kalangan kelas pendapatan isi rumah (B40, M40, dan T20). Faktor-faktor yang menyumbang kepada matlamat ini ditentukan melalui pemilihan ciri dari data Tinjauan Pendapatan dan Perbelanjaan Isi Rumah (HIES). Selain itu, model klasifikasi tingkah laku perbelanjaan dibangunkan menggunakan lima kaedah algoritma iaitu *Decision Tree*, *Random Forest* (Model rujukan), *Support Vector Machines*(SVM), *K-Nearest Neighbour*(KNN), dan *Naïve Bayes*. Pada masa ini, Model yang dapat mengenal pasti tingkah laku kewangan yang lebih optimum amat diperlukan bagi membantu isi rumah mengurus perancangan kewangan mereka. Tambahan pula, mengenal pasti perbelanjaan yang paling banyak menyumbang kepada perbelanjaan berlebihan yang tidak diperlukan dapat membantu individu merancang perbelanjaan mereka dengan lebih baik.

II. Fasa pemahaman data

Sumber data diperoleh daripada Tinjauan Pendapatan dan Perbelanjaan Isi Rumah Malaysia 2019 yang telah dilakukan pada tahun 2019 oleh Jabatan Perangkaan Malaysia (DOSM) dikenali sebagai HES 2019. Serta di sini juga saya ingin menekankan bahawa data set HES 2019 digunakan dalam kajian ini telah diuruskan dan diproses terlebih dahulu oleh (Azmi, 2023) untuk projek akhir tahun beliau pada tahun 2022/2023, Set data merangkumi rekod perbelanjaan isi rumah termasuk rekod isi rumah, ahli isi rumah, dan perbelanjaan dengan jumlah data sebanyak 16354 baris data serta 144 parameter.

III. Fasa penyediaan data

Sebelum melalui fasa permodelan, data perlu melalui pemprosesan terlebih dahulu bagi memastikan data boleh digunakan ketika fasa pemodelan. Dengan mengintegrasikan data HES 2019 dengan Klasifikasi Penggunaan Individu mengikut Tujuan (COICOP). COICOP

membahagikan belanjawan penggunaan individu dan isi rumah mengikut dua belas bahagian dan kumpulan. Terdapat 12 kategori yang dinyatakan berdasarkan COICOP, iaitu Makanan dan minuman bukan alkohol, Minuman keras, tembakau, narkotik (ATN), Pakaian, Perumahan, Perabot, Kesihatan, Pengangkutan, Komunikasi, Rekreasi dan kebudayaan (R&C), Pendidikan, Restoran dan hotel (R&H) dan pelbagai(Lain-lain).

Seterusnya, parameter kelas baharu (binary) ditambah iaitu isi rumah yang berbelanja berlebihan yang berbelanja lebih daripada pendapatan mereka diberi nilai '1' manakala sebaliknya dilabelkan '0'. Penambahan parameter bagi mengkategorikan tiga kelas pendapatan berbeza B40, M40 dan T20 mengikut negeri berdasarkan Laporan Survei Pendapatan Garis Kemiskinan dan Pendapatan Isi Iumah Dibandingkan di Seluruh Negeri Malaysia pada tahun 2019 oleh Jabatan Perangkaan Malaysia. Serta penambahan parameter: isi rumah yang berbelanja berlebihan yang berbelanja lebih bagi setiap item perbelanjaan dalam bentuk peratus berdasarkan panduan diperoleh had penggunaan ambang dalam peratus mengikut item perbelanjaan yang dicadangkan oleh (Sabri, 2020) diberi nilai '1' manakala sebaliknya dilabelkan '0'.

Dataset dibahagikan kepada tiga set data mengikut tiga kelas kelas pendapatan B40, M40 dan T20. Beberapa parameter tidak diperlukan telah dialih keluar daripada set data. Parameter yang dialih keluar termasuk nombor pengenalan isi rumah(HID). Akhir sekali, proses pemilihan atribut atau ciri dilakukan untuk memilih ciri atau atribut terpenting yang sangat berkorelasi dengan pemboleh ubah sasaran. Jadual 1 menyenaraikan parameter (semua angka, Lebihan perbelanjaan (binary), Lebihan perbelanjaan mengikut item perbelanjaan(binary)) yang digunakan dalam kajian supaya memudahkan algoritma untuk mengenal pasti corak dan hubungan dalam data. Pra-pemprosesan melibatkan penskalaan data dan pemilihan model terbaik untuk tafsiran.

Jadual 1 Senarai atribut selepas langkah pemprosesan dan penyediaan data

No	Atribut	Penerangan
1	Jum_IR	Bilangan isi rumah
2	Negeri	1 Semenanjung, 2 Sabah dan W.P.Labuan dan 3 Sarawak
3	Kawasan	1 Johor, 2 Kedah, 3 Kelantan, 4 Melaka, 5 Negeri Sembilan, 6 Pahang, 7 Pulau Pinang, 8 Perak, 9 Perlis, 10 Selangor, 11 Terengganu, 12 Sabah, 13 Sarawak, 14 W.P. Kuala Lumpur, 15 W.P. Labuan dan 16 W.P. Putrajaya
4	Strata	1 Bandar, 2 Luar Bandar
5	Kump Etnik	1 Bumiputera, 2 Cina, 3 India, 4 Lain-lain
6	Jum_Perbelanjaan	Jumlah perbelanjaan isi rumah secara bulanan
7	Jum_Pend_Kasar	Jumlah pendapatan isi rumah secara bulanan
8	Kategori_kelas_pendapatan	Pengelasan 1 belanja berlebihan, 0 tidak belanja berlebihan
9	Perbelanjaan_berlebihan	B40 kelas pendapatan bawah, M40 kelas pendapatan tengah, T20 kelas pendapatan atas.
10	Perbelanjaan_Makanan	Pengelasan kumpulan belanja Makanan dan minuman bukan Alkohol (Makanan)
11	Perbelanjaan_ATN	Pengelasan kumpulan belanja Minuman alkohol dan tembakau(ATN)

12	Perbelanjaan_Pakaian	Pengelasan kumpulan belanja Pakaian dan kasut (Pakaian)
13	Perbelanjaan_Perumahan	Pengelasan kumpulan belanja Perumahan, air, elektrik, gas dan bahan api lain (Perumahan)
14	Perbelanjaan_Perabot	Pengelasan kumpulan belanja Hiasan, perkakasan dan penyelenggaraan isi rumah (Perabot)
15	Perbelanjaan_Kesihatan	Pengelasan kumpulan belanja Kesihatan
16	Perbelanjaan_Pengangkutan	Pengelasan kumpulan belanja Pengangkutan
17	Perbelanjaan_Komunikasi	Pengelasan kumpulan belanja Komunikasi
18	Perbelanjaan_RB	Pengelasan kumpulan belanja Perkhidmatan rekreasi dan kebudayaan (R&B)
19	Perbelanjaan_Pendidikan	Pengelasan kumpulan belanja Pendidikan
20	Perbelanjaan_RH	Pengelasan kumpulan belanja Restoran dan hotel (R&H)
21	Perbelanjaan_Lainlain	Kumpulan belanja Pelbagai barang dan perkhidmatan (Lain-lain)

IV. Fasa permodelan

Fasa pemodelan melibatkan penggunaan pelbagai teknik dan algoritma pemodelan untuk menjalankan satu kajian. Dalam proses ini, data HES 2019 telah dibahagikan kepada tiga set data yang berbeza. Kajian akan dilaksanakan dan model akan dibangunkan bagi setiap set data yang berkaitan, iaitu set B40, M40, dan T20. Setiap set data ini akan melalui proses pemilihan atribut atau ciri-ciri yang penting dan bermakna terlebih dahulu (Iranzad & Liu, 2024). Kemudian, melalui proses pemodelan bagi mendapatkan pemboleh ubah yang terbaik bagi setiap model dengan penalaan hyperparameter menggunakan pencarian grid dan pengesahan silang 5 kali ganda (Shetty et al., 2024). Teknik Persampelan Terlebih Minoriti Sintetik (SMOTE) juga digunakan untuk menangani ketidakseimbangan kelas dalam set data M40 dan T20 dengan menjana sampel sintetik untuk kelas minoriti. Pencarian grid dan Teknik Persampelan Terlebih Minoriti Sintetik adalah salah satu perpustakaan yang terdapat dalam aplikasi Python. Algoritma yang digunakan bagi kajian meramal tingkah laku perbelanjaan terhadap tiga kelas pendapatan di Malaysia adalah lima kaedah algoritma iaitu iaitu Decision Tree, Random Forest (Model rujukan), Support Vector Machines (SVM), K-Nearest Neighbour (KNN), dan Naïve Bayes.

V. Fasa penilaian

Fasa penilaian ini akan melibatkan penilaian ke atas model pengelasan untuk mengukur sejauh mana pengelasan yang dicadangkan menggunakan model sepadan dengan pengelasan sebenar kes tersebut melalui parameter penilaian seperti ketepatan, kejituan, kekhususan (specificity), kepekaan (sensitivity), F-ukuran (F1-score), dan dapatan semula (recall) sering digunakan dalam penilaian klasifikasi. Oleh itu, kajian ini juga menggunakan parameter penilaian tersebut dalam menilai pendekatan yang diusulkan. Untuk memahami bagaimana metrik ini dikira, pengetahuan tentang matriks kekeliruan (confusion matrix) diperlukan. Matriks kekeliruan dalam masalah pengelasan dengan dua kelas boleh dirujuk dalam Jadual 2. Ia boleh disimpulkan daripada jadual bahawa terdapat empat kemungkinan ramalan keputusan yang berbeza. Hasil yang benar-benar positif (True Positive) dan benar-benar negatif (True Negative) ialah pengelasan yang betul, manakala hasil positif palsu (False Positive) dan negatif palsu (False Negative) ialah dua jenis ralat yang mungkin berlaku.

Jadual 2 Matriks kekeliruan.

		Kelas ramalan	
		Positif	Negatif
Kelas sebenar	Positif	<i>True Positive</i> (TP)	<i>False Negative</i> (FP)
	Negatif	<i>False Positive</i> (FN)	<i>True Negative</i> (TF)

Ketepatan mengukur peratusan ramalan yang betul daripada keseluruhan ramalan yang dibuat oleh model. Ia memberi gambaran umum tentang prestasi model secara keseluruhan. Persamaan diberikan oleh Persamaan 1:

$$Ketepatan = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Kejituan adalah nisbah ramalan positif yang betul berbanding jumlah keseluruhan ramalan positif yang dibuat oleh model. Ia menunjukkan berapa banyak ramalan positif yang sebenarnya betul, ia diberikan oleh Persamaan 2:

$$Kejituan = \frac{TP}{TP + FP} \quad (2)$$

Kekhususan(specificity) mengukur kemampuan model untuk mengenal pasti negatif yang betul daripada keseluruhan kes negatif yang sebenar. Ia penting dalam konteks di mana mengelakkan positif palsu adalah kritikal. Persamaan diberikan oleh Persamaan 3:

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Kepekaan(sensitivity), mengukur kemampuan model untuk mengenal pasti positif yang betul daripada keseluruhan kes positif yang sebenar. Ia penting untuk menangkap semua kes positif. Persamaan diberikan oleh Persamaan 4:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

F-Ukuran(F1-score) adalah ukuran keseimbangan antara kejituan dan kepekaan, dihitung sebagai purata harmonik dari kedua-duanya. Ia berguna apabila kita ingin mengimbangi kedua-dua metrik tersebut. Persamaan diberikan oleh Persamaan 5:

$$F1 \text{ score} = 2 \times \frac{Kejituan \times Recall}{Kejituan + Recall} \quad (5)$$

Dapatan semula (Recall) adalah sinonim kepada kepekaan, ia mengukur nisbah ramalan positif yang betul berbanding jumlah keseluruhan kes positif yang sebenar. Ia memastikan semua kes positif dikenalpasti. Persamaan diberikan oleh Persamaan 6:

$$Recall = \frac{TP}{TP + FN}$$

(6)

Pemilihan kaedah penilaian bergantung pada masalah perniagaan dan jenis data. Matlamatnya ialah mengenal pasti model dengan ketepatan pengelasan tertinggi. Selain ukuran ketepatan, kejituan, kekhususan (specificity), kepekaan (sensitivity), F-ukuran (F1-score), dan dapatan semula (recall), ukuran seperti Ralat Kuasa Dua Min (MSE), ralat mutlak purata (MAE), ralat min mutlak (RMSE) juga diambil kira dalam menilai prestasi model.

VI. Fasa pelaksanaan

Fasa pelaksanaan ialah fasa terakhir dalam pembangunan model ramalan tingkah laku perbelanjaan isi rumah mengikut kelas pendapatan. Keputusan ramalan kini siap untuk disebarkan dan digunakan oleh orang ramai. Hasil model bagi setiap kelas pendapatan di Malaysia akan dipaparkan melalui sistem web, membolehkan pihak berkepentingan memahami dan meneroka hasil analisis dengan mudah. Pemantauan dan penyelenggaraan berkala diperlukan untuk memastikan ketepatan ramalan dan mengelakkan penyebaran maklumat palsu.

KEPUTUSAN DAN PERBINCANGAN

Kajian ini menghasilkan beberapa perkara utama, termasuk parameter yang dipilih dan model ramalan yang terbaik. Pemilihan ciri memainkan peranan penting dalam analitik data kerana ia membantu mengenal pasti parameter yang paling relevan yang boleh mempengaruhi hasil model, menghasilkan ramalan yang lebih tepat. Membandingkan model ramalan juga adalah penting untuk menentukan kaedah terbaik untuk set data tertentu dan untuk mengelakkan model daripada terlalu atau kurang sesuai.

I. Pemilihan atribut

Pemilihan atribut adalah proses memilih atribut yang relevan dan penting daripada set atribut yang lebih besar dalam set data. Kajian ini menggunakan Random Forest yang berfungsi membina beberapa pokok keputusan pada subset rawak set data dan kemudian mengagregatkan ramalan setiap pokok untuk mencapai ramalan akhir. Semasa proses ini, *Random Forest* mengira kepentingan setiap parameter atau atribut dalam set data (Iranzad & Liu, 2024). Serta penggunaan kaedah *SelectFromModel* dalam perpustakaan *sci-kit-learn* Python, yang menggunakan model pembelajaran mesin untuk memilih parameter yang paling penting. Fungsi ini menggunakan ambang purata atau nilai ambang untuk memilih parameter dengan skor kepentingan yang lebih besar daripada purata semua skor.

Selepas membina model *Random Forest*, kepentingan ciri boleh diekstrak dan disenaraikan mengikut skor mereka. Parameter dengan skor lebih tinggi daripada nilai ambang yang telah ditetapkan sebagai 0.015 dianggap lebih relevan untuk ramalan parameter kelas. Cara ini dilakukan untuk mengatasi masalah dimensi set data yang tinggi berlaku. Proses pemilihan atribut ini mengurangkan kerumitan model dan meningkatkan prestasi model dengan memberi tumpuan kepada atribut yang paling relevan. Atribut yang dipilih bagi setiap set data dikumpulkan dan digunakan untuk menguji model pengelasan atau klasifikasi. Jadual 3 menunjukkan atribut penting yang dipilih untuk setiap set data kelas pendapatan B40, M40 dan T20, dan Jadual 4 menyediakan penerangan mengenai parameter atau atribut tersebut.

Jadual 3 Atribut penting yang dipilih bagi set data B40, M40 dan T20

Keutamaan atribut	Atribut pilihan		
	B40	M40	T20
1	Jum_Pend_Kasar	Jum_Perbelanjaan	Jum_Perbelanjaan
2	Jum_Perbelanjaan	Jum_Pend_Kasar	Jum_Pend_Kasar
3	Negeri	Negeri	Negeri
4	Jum_IR	Jum_IR	Jum_IR
5	Perbelanjaan_Perabot	Perbelanjaan_Perabot	Perbelanjaan_Perabot
6	Kump_Etnik	Kump_Etnik	Perbelanjaan_RH
7	Perbelanjaan_Komunikasi	Perbelanjaan_Makanan	Kump_Etnik
8	Strata	Perbelanjaan_Pengangkutan	Perbelanjaan_Lainlain
9	Perbelanjaan_RB	Perbelanjaan_RH	Kawasan
10	Perbelanjaan_Lainlain	Strata	
11	Perbelanjaan_Perumahan	Perbelanjaan_Komunikasi	
12	Perbelanjaan_Pakaian	Perbelanjaan_Lainlain	
13	Perbelanjaan_Pengangkutan		

Jadual 4 Penerangan atribut

Atribut	Penerangan
Jum_IR	Bilangan isi rumah
Kawasan	1 Semenanjung, 2 Sabah dan W.P.Labuan dan 3 Sarawak
Negeri	1 Semenanjung, 2 Sabah dan W.P.Labuan dan 3 Sarawak
Strata	1 Bandar, 2 Luar Bandar
Kump_Etnik	1 Bumiputera, 2 Cina, 3 India, 4 Lain-lain
Jum_Perbelanjaan	Jumlah perbelanjaan isi rumah secara bulanan
Jum_Pend_Kasar	Jumlah pendapatan isi rumah secara bulanan
Perbelanjaan_Makanan	Pengelasan kumpulan belanja Makanan dan minuman bukan Alkohol (Makanan)
Perbelanjaan_Pakaian	Pengelasan kumpulan belanja Pakaian dan kasut (Pakaian)
Perbelanjaan_Perumahan	Pengelasan kumpulan belanja Perumahan, air, elektrik, gas dan bahan api lain (Perumahan)
Perbelanjaan_Perabot	Pengelasan kumpulan belanja Hiasan, perkakasan dan penyelenggaraan isi rumah (Perabot)
Perbelanjaan_Pengangkutan	Pengelasan kumpulan belanja Pengangkutan
Perbelanjaan_Komunikasi	Pengelasan kumpulan belanja Komunikasi
Perbelanjaan_RB	Pengelasan kumpulan belanja Perkhidmatan rekreasi dan kebudayaan (R&B)
Perbelanjaan_RH	Pengelasan Kumpulan belanja Restoran dan hotel (R&H)
Perbelanjaan_Lainlain	Kumpulan belanja Pelbagai barang dan perkhidmatan (Lain-lain)

II. Model pengelasan

Atribut yang dipilih daripada algoritma *Random Forest* digunakan untuk membina model ramalan, termasuk *Decision Tree*, *Random Forest* (RF), *Support Vector Machines* (SVM), *K-Nearest Neighbour* (KNN), dan *Naïve Bayes* (NB). Skor metrik keseluruhan bagi setiap model merentasi set data B40, M40 dan T20 ditunjukkan dalam Jadual 5. Keputusan menunjukkan bahawa *Support Vector Machine* (SVM) mengatasi kaedah lain dalam semua ukuran dan hanya sedikit lebih rendah daripada NB dalam ukuran recall. Perlu diingat bahawa *Random Forest* pada mulanya dipilih sebagai model rujukan kerana keteguhan dan popularitinya dalam mengendalikan pelbagai jenis dataset. Walau bagaimanapun, selepas penilaian model yang teliti dan berdasarkan nilai-nilai metrik terbaik, algoritma yang dipilih bagi kajian ini sebagai model pengelasan adalah SVM kerana didapati bahawa model *Support Vector Machine* (SVM) menghasilkan metrik prestasi yang lebih baik berbanding model *Random Forest*.

Jadual 5 Metrik ukuran keseluruhan bagi setiap model merentasi set data B40, M40 dan T20

Metrik ukuran	Model				
	NB	SVM	KNN	DT	RF
ketepatan	0.943	0.999	0.994	0.983	0.990
Kejituan	0.371	1.000	0.933	0.678	0.589
Kekhususan	0.939	1.000	0.998	0.989	0.997
Kepekaan	1.000	0.933	0.907	0.791	0.618
F1 score	0.520	0.963	0.918	0.727	0.600
Recall	1.000	0.933	0.907	0.791	0.618
MSE	0.057	0.001	0.006	0.017	0.010
MAE	0.057	0.001	0.006	0.017	0.010
RMSE	0.234	0.015	0.068	0.128	0.101

III. Perbandingan dapatan model dengan kajian lepas

Bahagian ini akan membandingkan penemuan model kajian ini dengan kajian lepas (Abdul, 2023) kerana kedua-dua kajian mempunyai persamaan dalam metodologi kajian, pembelajaran mesin yang digunakan dan matlamat kajian adalah meramalkan tingkah laku perbelanjaan kelas pendapatan berbeza di Malaysia. Namun begitu, kajian ini menggunakan set data yang berlain dengan kajian lepas, iaitu set data Penyiasatan Perbelanjaan Isi Rumah (HES) yang dijalankan pada tahun 2016 dan tahun 2017 yang akan dirujuk sebagai HES 2016. Set data yang digunakan oleh kajian lepas mempunyai rekod isis rumah, ahli isi rumah dan perbelanjaan dengan jumlah data sebanyak 14551 baris data dengan 163 atribut. Perbandingan ciri pilihan model dengan kajian lepas boleh dirujuk dalam Jadual 6 sehingga Jadual 8.

Jadual 6 menunjukkan perbandingan ciri model bagi set data B40 dengan kajian lepas. Dalam kajian ini, bagi kumpulan pendapatan B40, tiga ciri perbelanjaan yang berkepentingan tinggi adalah perabot, komunikasi serta rekreasi dan kebudayaan. Ini membuktikan bahawa ketiga-tiga ciri ini jelas sekali menyimpulkan bahawa antara perbelanjaan terbesar bagi isi rumah B40 lebih kepada barang bukan keperluan asas. Ia juga menjelaskan impak ciri-ciri tersebut kepada model yang dibina. Manakala bagi kajian lepas pula, tiga ciri perbelanjaan yang berkepentingan tinggi adalah perumahan,

pengangkutan dan makanan. Ini membuktikan bahawa ketiga-tiga ciri ini jelas sekali merangkumi antara perbelanjaan terbesar bagi isi rumah B40 untuk data HES 2016 yang lebih kepada barang keperluan asas.

Jadual 6 Perbandingan kepentingan ciri model bagi set data B40 dengan kajian lepas

B40		
Kepentingan ciri	Kajian	Kajian lepas
1	Jumlah Pendapatan	Julat Pendapatan
2	Jumlah Perbelanjaan	Perumahan
3	Negeri	Pengangkutan
4	Jumlah Isi Rumah	Makanan
5	Perabot	Lain-lain
6	Kumplan Etnik	RdH
7	Komunikasi	Pakaian
8	Strata	Perabot
9	RdB	Kesihatan
10	Lainlain	RdB
11	Perumahan	Negeri
12	Pakaian	
13	Pengangkutan	

Bagi kumpulan pendapatan M40 pula, empat ciri perbelanjaan yang berkepentingan tinggi adalah perabot, makanan, pengangkutan serta restoran dan hotel. Ini membuktikan bahawa keempat-empat ciri ini jelas sekali merangkumi antara perbelanjaan terbesar bagi isi rumah M40. Ia juga menjelaskan impak ciri-ciri tersebut kepada model yang dibina. Manakala bagi kajian lepas pula, empat ciri perbelanjaan yang berkepentingan tinggi adalah perumahan, pengangkutan, makanan dan, restoran dan hotel (RdH). Ini membuktikan bahawa keempat-empat ciri ini jelas sekali merangkumi antara perbelanjaan terbesar bagi isi rumah M40 untuk data HES 2016. Berdasarkan perbandingan set data ini, terdapat persamaan ciri berkepentingan tinggi, iaitu pengangkutan. Ini menunjukkan bahawa seiring dengan peredaran masa, pengangkutan tetap menjadi aspek penting dalam isi rumah M40. Perbandingan kepentingan ciri model bagi set data M40 dengan kajian lepas boleh dirujuk dalam Jadual 7.

Jadual 7 Perbandingan kepentingan ciri model bagi set data M40 dengan kajian lepas

M40		
Kepentingan ciri	Kajian	Kajian lepas
1	Jumlah Perbelanjaan	Perumahan
2	Jumlah Pendapatan	Pengangkutan
3	Negeri	Negeri
4	Jumlah Isi Rumah	Makanan

5	Perabot	RdH
6	Kumplan Etnik	RdB
7	Makanan	Lain-lain
8	Pengangkutan	Perabot
9	RdH	Pakaian
10	Strata	Kesihatan
11	Komunikasi	ATN
12	Lain-lain	

Bagi kumpulan pendapatan T20 pula, tiga ciri perbelanjaan yang berkepentingan tinggi adalah perabot, restoran dan hotel dan lain-lain. Ini membuktikan bahawa ketiga-tiga ciri ini jelas sekali merangkumi antara perbelanjaan terbesar bagi isi rumah T20. Ia juga menjelaskan impak ciri-ciri tersebut kepada model yang dibina. Manakala bagi kajian lepas pula, empat ciri perbelanjaan yang berkepentingan tinggi adalah perumahan, restoran dan hotel(RdH) dan pengangkutan. Ini membuktikan bahawa keempat-empat ciri ini jelas sekali merangkumi antara perbelanjaan terbesar bagi isi rumah T20 untuk data HES 2016. Berdasarkan perbandingan set data ini, terdapat persamaan ciri berkepentingan tinggi, iaitu restoran dan hotel. Ini menunjukkan bahawa seiring dengan peredaran masa, isi rumah T20 memiliki kebebasan perbelanjaan yang lebih besar berbanding B40 dan M40 dengan berbelanja berlebihan barang kepada bukan keperluan asas. Mereka cenderung berbelanja untuk kemewahan seperti makan di luar dan hotel. Perbandingan kepentingan ciri model bagi set data T20 dengan kajian lepas boleh dirujuk dalam Jadual 8.

Jadual 8 Perbandingan kepentingan ciri model bagi set data T20 dengan kajian lepas

T20		
Kepentingan ciri	Kajian	Kajian lepas
1	Jumlah Perbelanjaan	Perumahan
2	Jumlah Pendapatan	RdH
3	Negeri	Pengangkutan
4	Jumlah Isi Rumah	Makanan
5	Perabot	Pakaian
6	RdH	Perabot
7	Kumplan Etnik	ATN
8	Lain-lain	Lain-lain
9	Kawasan	

IV. Papan antara muka interaktif

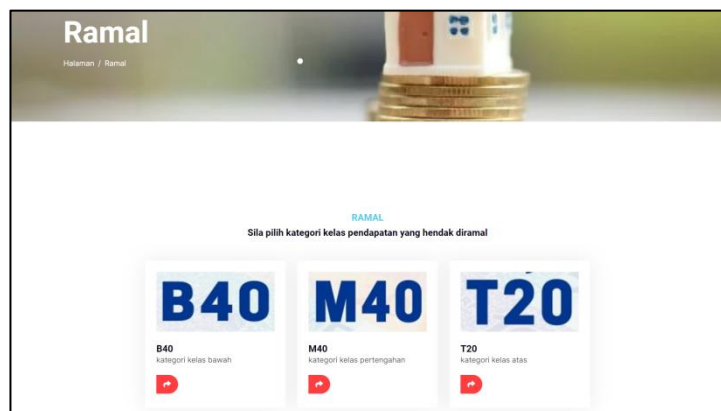
Setelah pemprosesan bahagian belakang berjaya dibina(BACK-END), model pengelasan yang dipilih perlu dibentangkan di mana sebuah sistem web dibina sebagai papan antara muka interaktif yang digunakan untuk meramal tingkah laku perbelanjaan setiap kelas

pendapatan berbeza di Malaysia yang melibatkan interaksi antara pengguna dan sistem. Aplikasi utama yang digunakan dalam pembangunan sistem ini termasuk Flask, yang digunakan untuk membuat laluan web dan mengendalikan permintaan HTTP (Khalid, 2024). Pandas digunakan untuk manipulasi dan analisis data, manakala Plotly digunakan untuk mencipta tampakan saling tindak. HTML dan CSS digunakan untuk membangunkan antara muka hadapan sistem (Powell, 2010).

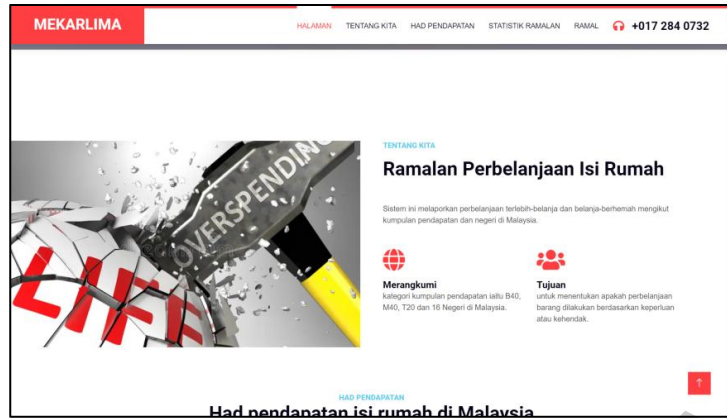
Apabila memasuki sistem, pengguna akan disambut dengan paparan antara muka bagi laman utama sistem ramalan tingkah laku perbelanjaan isi rumah bagi semua kelas pendapatan seperti yang ditunjuk pada Rajah 1. Untuk mulakan ramalan, pengguna perlu menekan butang 'Ramal' pada bar navigasi atau butang 'mulakan ramalan'. Ini akan membawa pengguna ke paparan antara muka ramal seperti yang ditunjuk pada Rajah 2. Pengguna dapat melihat paparan info mengenai sistem apabila pengguna menatal paparan antara muka laman utama ke bahagian 'Tentang Kita' seperti Rajah 3. Seterusnya, Pengguna dapat melihat info had pendapatan isi rumah mengikut kelas pendapatan di Malaysia iaitu B40, M40 dan T20 apabila pengguna menatal paparan antara muka laman utama ke bahagian 'Had Pendapatan' seperti Rajah 4. Pengguna juga dapat melihat statistik ramalan terlebih belanja mengikut negeri bagi setiap kelas pendapatan di Malaysia apabila pengguna menatal paparan antara muka laman utama ke bahagian 'Statistik ramalan' seperti Rajah 5.



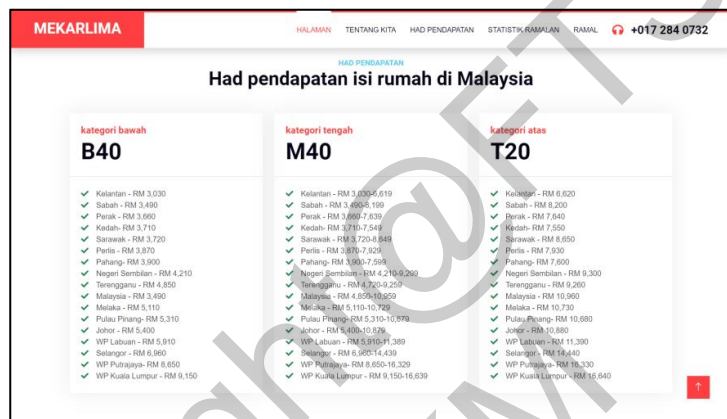
Rajah 1 Paparan reka bentuk laman utama



Rajah 2 Paparan Reka Bentuk Ramal



Rajah 3 Paparan Tentang Kita



Rajah 4 Paparan Had Pendapatan Isi Rumah

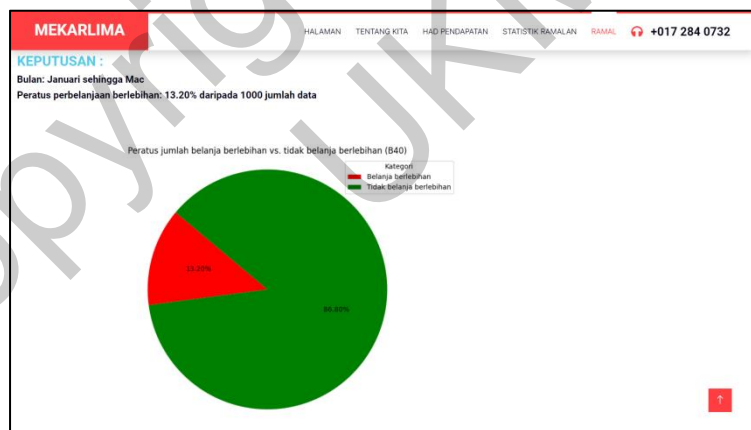


Rajah 5 Paparan Statistik Ramalan mengikut negeri

Apabila pengguna hendak mulakan ramalan, pengguna perlu navigasi ke paparan antara muka ramal dengan menekan butang ‘Ramal’ pada bar navigasi atau butang ‘mulakan ramalan’. Paparan antara muka ramal untuk mulakan ramalan adalah seperti yang ditunjukkan dalam Rajah 2, di mana pengguna perlu memilih untuk mu meramal bagi kategori pendapatan B40, M40 atau T20 dengan menekan butang ‘Mula’ yang akan muncul setelah penunjuk berada di simbol butang panah pada kategori kelas pendapatan yang akan diramal.

Apabila pengguna memilih untuk meramal kelas pendapatan B40, sistem akan memaparkan paparan antara muka ramal B40 seperti yang ditunjukkan dalam Rajah 6. Untuk meramal, pengguna perlu muat naik fail data mengikut kategori kelas pendapatan iaitu B40 dalam bentuk .csv, dan memilih bulan mula dan akhir yang dikehendaki terlebih dahulu sebelum menekan butang 'Ramal'. Seterusnya, sistem akan memaparkan keputusan ramalan iaitu laporan statistik terhadap perbelanjaan dan laporan statistik perbelanjaan mengikut kategori item perbelanjaan berdasarkan bulan tertentu disertai dengan pernyataan sama ada kelas pendapatan ini berbelanja berlebihan terhadap keperluan ataupun tidak seperti yang ditunjukkan dalam Rajah 7.

Rajah 6 Paparan Reka Bentuk Ramal B40

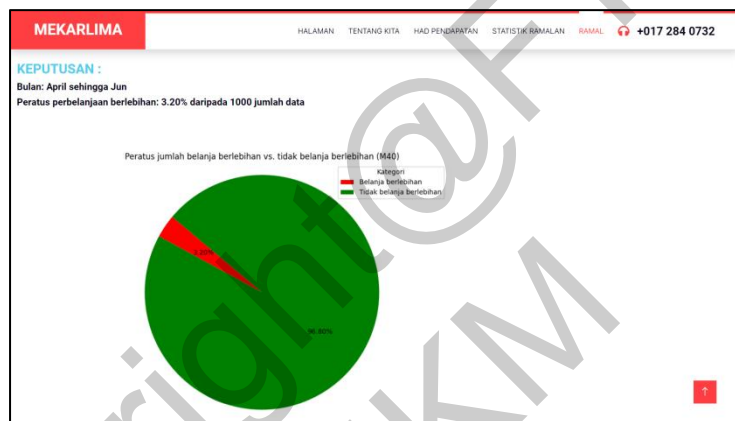


Rajah 7 Paparan Keputusan Ramalan B40

Apabila pengguna memilih untuk meramal kelas pendapatan M40, sistem akan memaparkan paparan antara muka ramal M40 seperti yang ditunjukkan dalam Rajah 8. Untuk meramal, pengguna perlu muat naik fail data mengikut kategori kelas pendapatan iaitu M40 dalam bentuk .csv, dan memilih bulan mula dan akhir yang dikehendaki terlebih dahulu sebelum menekan butang 'Ramal'. Seterusnya, sistem akan memaparkan keputusan ramalan iaitu laporan statistik terhadap perbelanjaan dan laporan statistik perbelanjaan mengikut kategori item perbelanjaan berdasarkan bulan tertentu disertai dengan pernyataan sama ada kelas pendapatan ini berbelanja berlebihan terhadap keperluan ataupun tidak seperti yang ditunjukkan dalam Rajah 9.

The screenshot shows a web interface for 'RAMAL M40' (Forecasting M40). The title is 'Ramal kategori kelas bawah'. Below the title, there is a note: 'Sila pastikan fail yang hendak dimuat naik adalah data dalam kategori kelas M40 dan fail dalam bentuk csv.' (Please ensure the file to be uploaded is data in the M40 category and the file is in CSV format). There is a 'Sila pilih fail:' (Please select file) section with a 'Choose File' button and 'No file chosen' text. Below this are two dropdown menus: 'Pilih Bulan Mula' (Select Start Month) and 'Pilih Bulan Akhir' (Select End Month). At the bottom of the form is a red 'Ramal' button.

Rajah 8 Paparan Reka Bentuk Ramal M40



Rajah 9 Paparan Keputusan Ramalan M40

Apabila pengguna memilih untuk meramal kelas pendapatan T20, sistem akan memaparkan paparan antara muka ramal T20 seperti yang ditunjukkan dalam Rajah 10. Untuk meramal, pengguna perlu muat naik fail data mengikut kategori kelas pendapatan iaitu T20 dalam bentuk .csv, dan memilih bulan mula dan akhir yang dikehendaki terlebih dahulu sebelum menekan butang 'Ramal'. Seterusnya, sistem akan memaparkan keputusan ramalan iaitu laporan statistik terhadap perbelanjaan dan laporan statistik perbelanjaan mengikut kategori item perbelanjaan berdasarkan bulan tertentu disertai dengan pernyataan sama ada kelas pendapatan ini berbelanja berlebihan terhadap keperluan ataupun tidak seperti yang ditunjukkan dalam Rajah 11.

MEKARLIMA HALAMAN TENTANG KITA HAD PENDAPATAN STATISTIK RAMALAN RAMAL +017 284 0732

RAMAL T20

Ramal kategori kelas bawah

Sila pastikan fail yang hendak dimuat naik adalah data dalam kategori kelas T20 dan fail dalam bentuk *csv*.

Sila pilih fail:

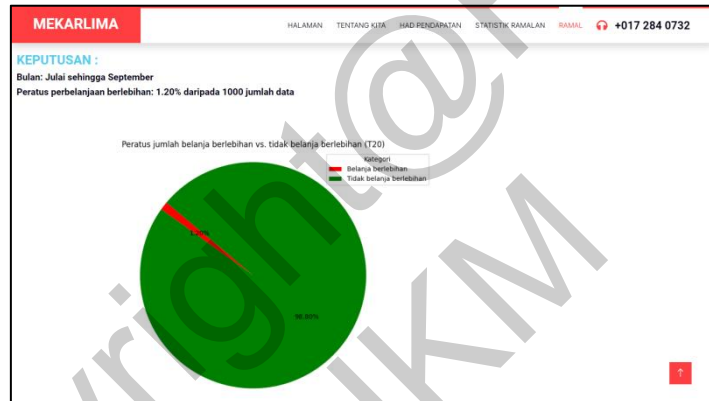
Choose File | No file chosen

Pilih Bulan Mula
Pilih Bulan Mula

Pilih Bulan Akhir
Pilih Bulan Akhir

Ramal

Rajah 10 Paparan Reka Bentuk Ramal T20



Rajah 11 Paparan Keputusan Ramalan T20

Cadangan Penambahbaikan

Pembangunan model pembelajaran mesin untuk meramal tingkah laku perbelanjaan isi rumah di Malaysia memberikan gambaran mengenai tahap kecelikan kewangan bagi setiap kategori pendapatan B40, M40, dan T20. Untuk analisis masa hadapan, kajian boleh dijalankan menggunakan data terbaru dari DOSM bagi tahun seterusnya. Selain meramal perbelanjaan, kajian perlu mengkaji faktor penyebab dan kesan perbelanjaan berlebihan untuk membantu isi rumah bagi setiap kelas pendapatan menguruskan perbelanjaan mereka dengan lebih efisien. Cadangan penambahbaikan kajian melibatkan langkah penyediaan, pra-pemprosesan, dan pemodelan data, kerana data yang diberikan sudah lengkap dengan maklumat yang diperlukan.

KESIMPULAN

Secara keseluruhan, Projek ini melibatkan penggunaan algoritma-algoritma pembelajaran mesin untuk meramal perlakuan perbelanjaan berlebihan isi rumah B40, M40 dan T20 di Malaysia. Matlamat bisnes yang telah dinyatakan pada perancangan awal projek ini telah dicapai melalui pembangunan model pilihan. Set data yang digunakan dalam projek ini merupakan set data yang diperoleh dari Jabatan Statistik Negara (DOSM) iaitu set data HES 2019 yang telah melalui pemprosesan data pada fasa penyediaan data untuk menjamin kualiti data supaya tidak mempengaruhi keputusan ramalan. Model yang dibangunkan, terutamanya model SVM, telah menunjukkan prestasi yang unggul dalam meramalkan tingkah laku perbelanjaan berlebihan, memberikan cerapan berharga tentang kecelikan kewangan. Penemuan ini penting untuk pendidikan kewangan dan pembangunan dasar, membolehkan pengenalpastian individu yang memerlukan bantuan dalam meningkatkan kecelikan dan tingkah laku kewangan mereka.

I. Kekuatan Sistem

Dari segi kekuatan, sistem ini menunjukkan ketepatan yang tinggi dalam ramalan tingkah laku perbelanjaan berlebihan dengan penggunaan model SVM, yang memberikan cerapan penting untuk meningkatkan kecelikan kewangan. Penggunaan pelbagai algoritma pembelajaran mesin seperti *Decision Tree*, *Random Forest*, SVM, kNN, dan Naïve Bayes membolehkan perbandingan dan pemilihan model terbaik. Selain itu, penggunaan set data HES 2019 yang telah diproses dengan teliti memastikan kualiti data yang tinggi, mengurangkan risiko keputusan yang salah.

II. Kelemahan Sistem

Namun begitu, terdapat beberapa kelemahan dalam sistem ini. Ketidakseimbangan dalam set data B40, M40, dan T20, di mana data yang tidak terlebih belanja jauh melebihi data terlebih belanja, menyukarkan untuk mendapatkan ukuran metrik yang tepat dan boleh dipercayai. Selain itu, Keutamaan tinggi perlu diberikan untuk memastikan data diproses dengan tepat, kerana kesilapan dalam pemprosesan boleh membawa kepada keputusan yang salah. Terdapat juga kekurangan dokumentasi yang mencukupi mengenai cara menggunakan Python Flask untuk berinteraksi dengan elemen-elemen dinamik atau boleh ditekan seperti output pyLDAvis, memerlukan sumber rujukan tambahan untuk mencari panduan yang tepat.

Projek ini membuktikan potensi penggunaan algoritma pembelajaran mesin dalam meramalkan tingkah laku perbelanjaan berlebihan dan memberikan pandangan yang penting untuk meningkatkan kecelikan kewangan dalam kalangan isi rumah di Malaysia. Cadangan untuk penambahbaikan pada masa hadapan termasuk penggunaan data terbaru dan kajian faktor penyebab dan kesan perbelanjaan berlebihan.

PENGHARGAAN

Penulis kajian ini ingin ucapkan setinggi-tinggi penghargaan dan jutaan terima kasih kepada Prof. Dr. Azuraliza Abu Bakar, penyelia penulis kajian ini yang telah memberi tunjuk ajar serta bimbingan untuk menyiapkan projek ini dengan jayanya.

Penulis kajian ini juga ingin mengucapkan terima kasih kepada semua pihak yang membantu secara langsung mahupun tidak langsung dalam menyempurnakan projek ini. Segala bantuan yang telah dihulurkan amatlah dihargai kerana tanpa bantuan mereka, projek ini tidak dapat dilaksanakan dengan baik. Semoga tuhan merahmati dan memberikan balasan yang terbaik.

RUJUKAN

- Abdul, L. I. (2023). Model Ramalan Tingkah Laku Kewangan Berdasarkan Tabiat Berbelanja Bagi Kelas Pendapatan Isi Rumah Malaysia Menggunakan Pendekatan Pembelajaran Mesin.
- Azmi, A. N. (2023). Kaedah Perlombongan Petua Sekutuan Untuk Meramalkan Tingkah Laku Perbelanjaan Dalam Kalangan Kelas Pendapatan B40.
- DOSM. (2019). Kumpulan Utama Klasifikasi Penggunaan Individu Mengikut Tujuan (COICOP) Malaysia.
- Iranzad, R., & Liu, X. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics*. <https://doi.org/10.1007/s41060-024-00509-w>
- Khalid, Z. (2024, March 20). Handling Requests in Flask. Medium.
- Powell, T. (2010). *HTML & CSS: The Complete Reference, Fifth Edition (5th ed.)*. McGraw-Hill, Inc.
- Sabri, M. F. (2020). *Mengurus Bajet 2019 Anda Dan Keluarga*. INSTITUTE FOR MATHEMATICAL RESEARCH.
- Shetty, A. M., Aljnid, M. F., Manjaiah, D. H., & Shaik Afzal, A. M. S. (2024). Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis (pp. 451–474). https://doi.org/10.1007/978-981-99-7814-4_36

Nur Hazwani Binti Shamsudin (A192968)

Prof. Dr. Azuraliza Abu Bakar

Fakulti Teknologi & Sains Maklumat

Universiti Kebangsaan Malaysia