RAMALAN PENGELASAN PENYENARAIAN HARTANAH PALSU BERDASARKAN K-MIN DAN MESIN SOKONGAN VEKTOR

Shaema Binti Mohd Hamim[1], Nor Samsiah binti Sani[2]

[1]Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor Darul Ehsan, Malaysia.

P107027@ukm.edu.my, norsamsiahsani@ukm.edu.my

ABSTRAK

Pasaran pelaburan hartanah melibatkan urusan jual beli hartanah sangat bergantung kepada pengiklanan atau penyenaraian hartanah bagi menyebarluas maklumat penawaran jual beli kepada orang ramai. terdapat banyak agensi hartanah yang menyediakan platform pengiklanan penyenaraian hartanah secara dalam talian di Malaysia. Perkhidmatan penyenaraian hartanah yang ditawarkan kebanyakannya adalah percuma. Orang ramai mahupun ejen-ejen hartanah dapat menjalankan urusan jual beli dengan mudah dan pantas dengan adanya perkhidmatan ini. Industri pelaburan hartanah Malaysia berkembang pesat dengan kehadiran teknologi dan pembangunan ekonomi yang pantas. Walau bagaimanapun, industri hartanah telah dicemari dengan pemain industri yang tidak bertanggungjawab. Penyelewengan dan penipuan telah berlaku dengan penyenaraian hartanah palsu seperti mengiklankan hartanah dengan gambar yang bukan gambar sebenar atau pun harga penawaran yang murah dari harga sebenar. Kajian ini bertujuan untuk meramal penyenaraian hartanah palsu dari data hartanah yang sebenar menggunakan teknik pembelajaran mesin tanpa selia. Model pengelompokan k-min telah dibangunkan bagi mengelompok penyenaraian hartanah palsu. Seterusnya, model pengelasan penyenaraian hartanah palsu dibangunkan menggunakan tiga algoritma pengelasan: algoritma pohon keputusan, mesin sokongan vektor dan juga algoritma hutan rawak. Model pengelasan dibangunkan berdasarkan set data pengelompokan k-min serta set data penyenaraian hartanah asal yang diperolehi daripada industri penyenaraian hartanah. Keputusan eksperimen menunjukkan model pengelasan set data pengelompokan secara puratanya, telah mengatasi model pengelasan set data asal. Algoritma mesin sokongan vektor pula dipilih sebagai model terbaik yang dapat mengelas penyenaraian hartanah palsu dengan ketepatan 99.73%.

Kata Kunci: Ramalan Penyenaraian Hartanah Palsu, Hartanah Palsu, Pengelompokan, K-Min, Pembelajaran Mesin, Pembelajaran Mesin Tanpa Selia, Mesin Sokongan Vektor, Hutan Rawak, Pohon Keputusan.

I. PENGENALAN

Kuantiti data yang semakin meningkat dalam setiap persekitaran yang bersaing seperti industri hartanah bukan sahaja membuka peluang penyelidikan dalam mengekstrak maklumat berguna malah ia memberi cabaran untuk memproses jumlah data yang besar dengan berkesan. Perlombongan data merupakan proses mengkestrak maklumat yang sahih dan boleh difahami daripada pangkalan data yang besar, dimana maklumat itu tidak diketahui sebelum ini. Perkara utama yang perlu dibuat sebaik sahaja proses

mengkestrak data dibuat adalah mencari alat penganalisis data (*data analyzing tools*) yang cekap bagi mengubah data menjadi maklumat dan pengetahuan. Maklumat tersebut digunakan bagi membuat keputusan dengan menggabungkan kecerdasan buatan serta statistik (Hand et al. 2001).

Perlombongan data semakin banyak digunakan di dalam industri hartanah. Ia dapat membantu pihak berwajib dan pemain industri mendapatkan maklumat penting daripada data yang terkumpul. Sebagai contoh, pihak kerajaan dapat memahami status perkembangan industri hartanah dan dapat menggubal dasar yang sesuai bagi menjaga kepentingan industri hartanah di Malaysia. Pemaju hartanah juga mendapat manfaat dengan membuat perancangan projek hartanah serta mengetahui peluang komersial berdasarkan maklumat daripada teknologi perlombongan data.

Hartanah merujuk kepada tanah dan sebarang seni bina kekal, sama ada semula jadi atau buatan, termasuk rumah, pangsapuri, bangunan bukan kediaman dan pagar Yang et al. (2022). Industri hartanah mengalami kenaikan yang sangat pantas disebabkan oleh kehadiran teknologi dan pembangunan ekonomi yang pesat. Di dalam pasaran pelaburan hartanah, para pelabur menyasarkan untuk memaksimumkan keuntungan pendapatan dan keuntungan modal mereka (Li et al. 2021). Justeru, pengiklanan hartanah adalah sangat penting kepada pelabur hartanah mahupun individu bagi membuat urusan jual beli hartanah menjadi lebih pantas. Sebagai faktor utama dalam industri hartanah, pengiklanan dibuat dengan matlamat menyebarkan maklumat berkaitan pasaran hartanah kepada orang awam seterusnya menyasarkan kepada pelanggan yang berpotensi (Cheng et al. 2016).

Terdapat beberapa laman web penyenaraian hartanah di Malaysia yang terkenal seperti iProperty (https://www.iproperty.com.my/), *Propertyguru* (https://www.propertyguru.com.my/), dan beberapa laman web yang lebih dipercayai seperti *EdgeProp* (https://www.edgeprop.my/), *Prop Social* (https://www.propsocial.my/), *DurianProperty* (https://www.m.durianproperty.com.my/) serta banyak lagi. Laman-laman web ini menyediakan kemudahan mengiklan hartanah untuk dijual atau disewa. Perkhidmatan pengiklanan tersebut kebanyakannya adalah secara percuma, memudahkan orang ramai terutama sekali ejen-ejen hartanah bagi membuat penawaran hartanah. Ejen hartanah, pembeli dan penjual memainkan peranan penting di dalam pasaran hartanah. Jika pemilik rumah ingin menjual rumah, mereka boleh diwakili oleh ejen hartanah yang bertauliah.

Walau bagaimanapun, penyelewengan dan penipuan kerap berlaku dalam pasaran hartanah di Malaysia kerana kekurangan undang-undang yang ketat mahupun mekanisma yang boleh mengenalpasti penipuan penyenaraian hartanah. Salah satu penipuan yang sering dilakukan adalah gambar yang diiklankan adalah bukan gambar hartanah yang sebenar, sebagai kaedah untuk menarik perhatian daripada bakal pelanggan baharu. Antara kaedah penipuan lain adalah agensi menerbitkan penyenaraian rumah yang tidak benar atau tidak tersedia, pada harga yang murah, juga bertujuan untuk

menarik perhatian dan minat pelanggan. Keadaan ini menyebabkan prestasi agensi hartanah akan tercalar kerana pelanggan akan hilang kepercayaan kepada ejen dan agensi tersebut. Pelanggan juga akan mengalami kesusahan yang berulang bagi mencari hartanah yang sesuai dengan kemahuan dan keperluan mereka.

Pakar bidang hartanah (*domain expert*) yang dirujuk memaklumkan bahawa penyenaraian hartanah yang mengandungi kenyataan berikut; "*Pictures shown are for illustration purposes only*" atau kenyataan-kenyataan lain yang memberi maksud yang sama adalah merupakan pengiklanan palsu. Ini adalah kerana gambar hartanah yang disertakan di dalam sesuatu iklan hartanah tersebut bukan gambar hartanah yang sebenar ataupun asli sebaliknya gambar tersebut hanyalah bertujuan sebagai ilustrasi sahaja. Keadaan ini menyebabkan pelanggan menghadapi kesulitan bagi meninjau gambar hartanah yang sebenar. Sehubungan itu, penyenaraian hartanah yang mempunyai petikan perkataan '*illustration'* akan dilabel sebagai palsu bagi kajian ini.

A. Pernyataan Masalah

Masalah yang perlu diambil perhatian dalam kajian ini adalah tiada atau kekurangan kajian ramalan penyenaraian hartanah palsu menggunakan kaedah pembelajaran mesin tanpa selia iaitu teknik pengelompokan terhadap data penyenaraian hartanah oleh para pengkaji. Dapat dilihat kajian-kajian sebelum ini bagi domain hartanah tertumpu kepada pembangunan model ramalan harga hartanah menggunakan pembelajaran mesin (Abdulaziz & Zeki 2020; Li et al. 2021; Park & Kwon Bae 2015; Yang et al. 2022; Yu et al. 2021). Kebanyakan kajian juga hanya menggunakan algoritma pengelasan dan kurangnya kajian yang menggunakan algoritma pengelompokan dalam membuat ramalan. Sehubungan itu, kajian-kajian bagi profil palsu, penyenaraian palsu atau penipuan di dalam pelbagai domain turut dianalisa.

Pada masa kini, pelbagai kajian dalam pembelajaran mesin telah dijalankan untuk mengenal ciri-ciri atau atribut yang mempengaruhi harga rumah. Walau bagaimanapun, kajian berdasarkan pembelajaran mesin untuk mengenal ciri-ciri atau atribut penyenaraian hartanah palsu adalah terhad. Ini penting kerana ia dapat memberi petunjuk yang tepat dalam mengklasifikasikan penyenaraian hartanah palsu. Sehubungan itu, kajian ini dijalankan bagi mengenalpasti ciri-ciri tersebut seterusnya menghasilkan model yang dapat mengklasifikasi penyenaraian hartanah palsu.

Meramal penyenaraian hartanah palsu adalah sangat penting tetapi kurang diberi perhatian. Sejak dunia dilanda wabak COVID-19, keadaan ekonomi menjadi tidak menentu. Industri hartanah juga tidak terkecuali menerima kesan ini. Kempen Pemilikan Rumah atau *Home Ownership Campaign*

(HOC) yang berlangsung dari 1 Januari 2019 hingga 30 Jun 2020 adalah inisiatif kerajaan yang dirancang untuk menyokong pembeli rumah yang ingin membeli hartanah. Ia juga adalah bertujuan mendorong penjualan hartanah yang tidak terjual dalam pasaran perumahan Malaysia untuk merancakkan kembali industri hartanah Malaysia. Pengecualian penuh duti stem adalah antara faedah dari kempen ini. Justeru, pengiklanan yang tulin, tidak dicemari dengan penipuan-penipuan adalah sangat diperlukan. Mekanisma bagi membuat ramalan penyenaraian hartanah palsu adalah penting di dalam pasaran hartanah.

Sehubungan itu, bagi menyelesaikan isu-isu di atas, kajian ini akan membangunkan model pengelompokan penyenaraian hartanah palsu menggunakan algoritma k-Min yang optimum. Set data yang digunakan adalah data penyenaraian hartanah sebenar daripada laman web pengiklanan agensi hartanah *EdgeProp* (EP) dan *DurianProperty* (DP).

B. Matlamat Dan Objektif Kajian

Matlamat kajian ini adalah untuk meramal penyenaraian hartanah palsu berdasarkan pembelajaran mesin. Justeru, bagi memastikan matlamat tersebut dapat dicapai, tiga objektif kajian telah dirangka dan perlu dilaksanakan seperti berikut:

i. Membangunkan model pengelompokan k-Min yang optimum untuk ramalan penyenaraian hartanah palsu.

ii. Mengenalpasti kelompok penyenaraian hartanah palsu dan atribut penting dalam setiap kelompok melalui analisis kelompok yang diperoleh daripada model pengelompokan.

iii. Mengenal model pengelasan ramalan penyenaraian hartanah palsu berdasarkan ciri penting dari hasil pengelompokan.

Kertas kerja ini disusun mengikut bahagian seperti berikut: Bahagian II menyajikan kajian penyelidikan yang berkait domain hartanah, penyenaraian palsu, profil palsu dan penipuan dalam pelbagai domain dan teknik pengelompokan yang akan digunakan dalam kajian ini. Bahagian III menerangkan dengan terperinci metodologi kajian yang akan dilaksanakan termasuk kaedah pra-pemprosesan data, analisis deskriptif data, proses pra-pemprosesan data serta kaedah pelaksanaan eksperimen. Bahagian IV menganalisis hasil pengelompokan bagi mengenalpasti atribut-atribut yang mempengaruhi setiap kelompok yang dihasilkan. Model pengelasan penyenaraian hartanah menggunakan algoritma pengelasan juga akan dibangunkan dan dibincangkan. Bahagian V, menyimpulkan keseluruhan hasil temasuk kekangan, sumbangan kajian dan cadangan perluasan kajian pada masa hadapan.

II. KERJA PENYELIDIKAN YANG BERKAITAN

A. Kajian Penyelidikan Pembelajaran Mesin dan Perlombongan Data

Pembelajaran mesin merupakan sub-bidang kepada kecerdasan buatan. Matlamat utama pembelajaran mesin adalah untuk membolehkan komputer belajar secara automatik tanpa diberi arahan atau bantuan dari manusia dan ia akan menyesuaikan tindakan sewajarnya. Pembelajaran mesin akan meneroka pengkajian, pembinaan algoritma dan pembinaan model yang boleh dipelajari serta peramalan data bagi membuat keputusan berdasarkan data-data secara analitis. Jika komputer boleh meningkatkan cara ia melaksanakan tugas tertentu berdasarkan pengalaman lalu, maka mesin itu telah belajar. Oleh itu, pembelajaran mesin adalah tentang menggunakan sistem automatik dan belajar untuk melakukan yang lebih baik pada masa hadapan berdasarkan apa yang dialami pada masa lalu (Sani et al. 2018).

Menurut Bartschat et al. (2019), perlombongan data mempunyai asas yang kukuh dalam statistik, kecerdasan buatan, pembelajaran mesin dan penyelidikan pangkalan data. Defini asal perlombongan data merujuk kepada aplikasi yang sangat berguna bagi mengekstrak corak tersembunyi daripada data. Ia merupakan salah satu langkah dalam proses penerokaan pengetahuan daripada pangkalan data yang melibatkan analisis data. Terdapat lima fasa yang terlibat dalam proses penerokaan pengetahuan iaitu fasa penyediaan data, fasa pra-pemprosesan data, fasa transformasi data, fasa perlombongan data serta fasa interpretasi pengetahuan iaitu bagaimana corak pengetahuan ditafsir dan dipersembahkan dalam bentuk yang lebih difahami (Fayyad et al. 1996).

B. Kajian Terdahulu Bagi Domain Hartanah

Kebanyakan kajian terdahulu berkaitan hartanah yang dibuat adalah ramalan harga hartanah menggunakan kaedah pembelajaran mesin. Menurut kajian yang dibuat oleh Li et al. (2021) satu model ramalan harga yang berketepatan tinggi telah dicadangkan. Ia bertujuan untuk membina sistem sokongan bagi pengumpulan maklumat dan analisa hartanah yang menguntungkan secara automatik dalam pasaran pelaburan hartanah berdasarkan perlombongan data dan pembelajaran mesin. Kaedah Ralat Peratusan Purata Min (*Mean Average Percentage Error*) digunakan bagi membuat penilaian keputusan ramalan harga hartanah yang telah dibuat. Keputusan eksperimen kajian mereka menunjukkan algoritma *Light GBM* menunjukkan nilai ralat peratusan purata min lebih rendah berbanding menggunakan algoritma rangkaian neural.

Kajian berkenaan ramalan hari dalam pasaran atau *days on market* (DOM) bagi mengoptimumkan strategi jualan hartanah telah dibuat oleh Castelli et al. (2020). Kajian ini menyatakan bahawa bilangan hari sesuatu penyenaraian hartanah diterbitkan dalam talian mempunyai hubungan yang kuat dengan kebarangkalian penyenaraian tersebut adalah palsu. Atas sebab ini, mereka membina

sebuah model ramalan yang tepat untuk meramalkan bilangan hari penyenaraian yang diterbitkan akan berada dalam talian bagi membantu untuk menyelesaikan tugas mengenal pasti penyenaraian palsu. Sebanyak empat algoritma pembelajaran mesin telah digunakan iaitu *Lasso*, *Ridge*, *Elastic Net*, dan *Artificial Neural Networks.* Model ramalan dapat dibangunkan dengan menggunakan algoritma *Lasso Regression*, berupaya meramalkan hari dalam talian dengan tepat.

C. Teknik Pengelompokan k-Min

Algoritma pembahagian yang paling popular dan banyak digunakan adalah algoritma k-Min yang telah diperkenalkan oleh MacQueen pada tahun 1967. Algoritma pengelompokan pembahagian merupakan algoritma tanpa selia dimana data yang digunakan tidak mempunya label. Matlamat algoritma ini adalah untuk mencari kumpulan di dalam sesuatu data di mana pembolehubah *k* mewakili jumlah kumpulan. Setiap titik data akan diumpukkan kepada salah satu kumpulan berdasarkan ciri-ciri kesamaan data. Kelebihan algoritma k-Min adalah ia merupakan algoritma yang mempunyai masa pelaksanaan yang rendah serta kecekapan perkomputeran yang tinggi. Algoritma ini masih menjadi salah satu teknik pengelompokan paling popular kerana kecekapannya, mudah dilaksanakan serta efisien dari segi masa pelaksanaan (Oyelade et al. 2019).

D. Kajian Terdahulu Teknik Pengelompokan Bagi Domain Penipuan/Palsu

Memandangkan kajian terdahulu berkenaan penyenaraian hartanah palsu adalah sangat sedikit, rujukan yang dibuat adalah berdasarkan penyenaraian pengiklanan palsu, profil palsu atau penipuan pelbagai domain. Seifollahi et al. (2017) telah membuat kajian tentang analisa aktiviti pancingan data emel. Tiga algortima pengelompokan yang dioptimumkan (*Optimized based clustering algorithm*) telah dicadangkan iaitu *Multi-Start Modified Global K-Means, Incremental Nonsmoothed Optimization Clustering Algorithm* (INCA) dan algoritma berasaskan perbezaan perwakilan cembung bagi fungsi pengelompokan (DCClust). Hasil kajian mereka membuktikan algortima pengelompokan yang dioptimumkan menghasilkan kelompok yang lebih bermakna dan lebih mudah untuk ditafsir bagi set data pancingan data emel berbanding algoritma k-min.

Menurut kajian Kigerl (2016), algoritma k-Min digunakan bagi mengelompok jenis jenayah siber dan aktiviti yang berkaitan dengannya berdasarkan negara. Sebanyak 190 negara telah dikaji dengan tujuh jenis ukuran jenayah siber. Penemuan kajian telah menentukan bahawa sesebuah negara boleh dikategorikan kepada empat kategori yang berbeza berdasarkan aktiviti jenayah siber iaitu negara jenayah siber yang rendah, negara jenayah siber yang tidak serius, negara jenayah penipuan yuran pendahuluan, dan negara penipuan pancingan data.

Sangalli et al. (2020) telah membuat kajian bagi mengenalpasti pelajar yang membuat penipuan semasa menjalani kursus secara dalam talian. Terdapat dua cara pelajar berinteraksi melalui modul kursus iaitu dengan membaca bahan kursus dan menyelesaikan latihan. Walau bagaimanapun, terdapat kelakuan pelajar yang berlainan iaitu hanya menyelesaikan latihan tanpa membaca bahan kursus. Ini menunjukkan terdapat dua kemungkinan iaitu pelajar yang berkongsi jawapan dengan pelajar lain dan juga pelajar menggunakan akaun palsu untuk mendapatkan jawapan yang betul bagi latihan tersebut. Algoritma k-Min telah digunakan dan dapat mengelompokan dua kemungkinan tersebut menggunakan data yang terdiri daripada log aktiviti pelajar seperti interaksi dengan bahan kursus, cubaan penyerahan latihan dan penyiaran siaran di forum.

Kamil & Pharmasetiawan (2019) dalam kajiannya telah mencadangkan algoritma pengelompokan DBSCAN bagi mengesan pemalsuan di dalam data kehadiran pekerja menggunakan pengesanan cap jari. Pengiraan jarak yang ketat telah dibuat bagi membuang data sahih memandangkan data sahih secara semulajadinya adalah unik. Pengiraan jarak diantara objek diterangkan melalui fungsi ketaksamaan iaitu menggunakan purata Euclidean (AED), purata Manhattan (AMA) dan jarak Chebyshev (DMAX). Hasil kajian menunjukkan DMAX mengatasi teknik yang lain dengan nilai F-Measure sebanyak 84%.

Bagi mengesan serangan profil klon di *Facebook*, Kiruthiga et al. (2014) telah membuat kajian berdasarkan tempoh masa tindakan pengguna dan corak klik pengguna untuk mencari persamaan antara profil klon dan profil sebenar. Algoritma pengelompokan k-Min digunakan bagi mengumpulkan objek pengguna iaitu seperti objek dari sekolah atau kolej yang sama manakala *Clone Spotter* digunakan bagi mengesan pengklonan di *Facebook.* Kajian ini mengaplikasikan algoritma *Naive Bayes* untuk mengelaskan maklumat lengkap setiap pengguna. Kajian ini telah berjaya mengenalpasti serangan profil klon pengguna daripada Universiti California.

Hot & Popović-Bugarin (2016) telah membuat kajian pengelompokan data tanah berdasarkan ciri-ciri kimia tanah menggunakan algoritma k-Min serta *fuzzy* k-Min. Matlamat kajian ini adalah bagi mengetahui keupayaan perlombongan data dalam mengelaskan jenis tanah dan seterusnya menggambarkan tanah secara visual supaya lebih mudah difahami oleh masyarakat. Keputusan yang diperolehi menggunakan k-Min dipersembahkan di aplikasi *Google Map* dan juga *dynamic Open Street Map* Montenegro.

Selain daripada itu, Zhang et al. (2019) telah menjalan kajian tentang mengesan penipuan transaksi dalam talian dengan menggunakan kelas data yang tidak seimbang. Kelas data yang tidak seimbang terdiri daripada empat faktor iaitu pengagihan kelas yang tidak seimbang, saiz sampel, pemisahan kelas dan konsep dalaman kelas. Teknik pohon pengelompokan di gunakan bagi mengesan

masalah pemprosesan ketidakseimbangan kelas dan dibandingkan dengan model pengelasan yang lain iaitu *AdaBoosting*, pohon keputusan, hutan rawak, *SVM*, dan regresi logistik. Dengan menggunakan set data transaksi kad kredit daripada syarikat kewangan di Cina, pohon pengelompokan telah memberi keputusan yang terbaik.

III. METODOLOGI KAJIAN

A. Rangka Kerja

Rangka kerja bagi kajian ini merangkumi tiga fasa iaitu fasa input, fasa pembangunan model dan fasa output.

Fasa input bagi kajian ini merangkumi proses mengenalpasti masalah kajian setelah beberapa perkara dipertimbangkan. Seterusnya, penetapan objektif, skop dan kepentingan kajian akan dilakukan berdasarkan permasalahan kajian. Kajian yang lebih mendalam dilakukan keatas kajian-kajian yang telah diterbitkan dalam bidang pembelajaran mesin dan penyenaraian pengiklanan palsu, profil palsu atau penipuan pelbagai domain.

Fasa pembangunan model pengelompokan pula mengandungi langkah-langkah pra-pemprosesan data dan diikuti dengan pemilihan atribut. Kemudian, analisis deskriptif secara statistik akan dilakukan untuk melihat ciri dan nilai setiap atribut yang telah dipilih. Seterusnya, model pengelompokan akan dibangunkan dan penilaian prestasi model pengelompokan akan dilakukan. Fasa terakhir iaitu fasa output mengandungi proses analisis kelompok menggunakan teknik pengekstrakan ciri dan diikuti dengan pembangunan model pengelasan untuk mengelas pelajar berdasarkan prestasi yang telah dihasilkan. Prestasi model pengelasan ini juga akan dinilai bagi mengenalpasti teknik yang terbaik.

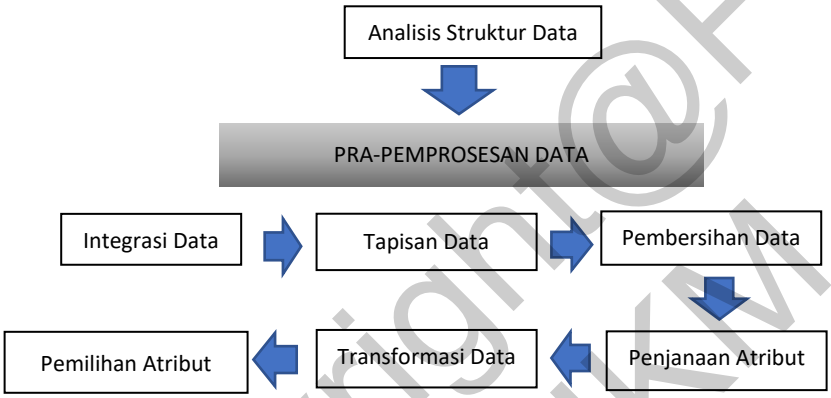
Fasa terakhir di dalam rangka kerja ini adalah fasa output yang merangkumi proses analisis keputusan, mengenal pasti korelasi di antara metrik-metrik kekompleksan dan mengenal pasti model pengelasan yang memberi ketepatan yang paling tinggi.

B. Fasa Pembangunan Model Pengelompokan

i. Data

Data mentah penyenaraian hartanah dikumpul secara rawak dan diekstrak daripada laman web agensi hartanah tempatan iaitu *EdgeProp* (EP) dan *DurianProperty* (DP) yang menyediakan perkhidmatan

penyenaraian hartanah. Sebanyak 12,914 data mentah yang diperolehi daripada set data DP dengan bilangan 28 atribut manakala 111,821 data mentah EP dengan 29 atribut. Jumlah keseluruhan data mentah yang terlibat di dalam kajian ini adalah sebanyak 124,735 data penyenaraian hartanah dengan atribut keseluruhan sebanyak 57. Tiada data peribadi pengguna atau ejen dikumpulkan bagi menjaga keselamatan data peribadi mereka. Kesemua data tersebut adalah dalam format *comma separated values* (CSV) dan merupakan data bagi Oktober 2021 (DP) manakala bagi data EP, maklumat tarikh penyenaraian data yang dinyatakan adalah tahun 2021. Sebaik sahaja data mentah ini diterima, aktiviti menganalisis struktur data dibuat. Data ini seterusnya akan melalui proses pra-pemprosesan data bagi mempersiapkan data sebelum proses pengelompokan bermula. Seperti yang ditunjukkan di dalam Rajah 1, proses pra-pemprosesan data melibatkan aktiviti integrasi data, tapisan data, pembersihan data, penjanaan atribut, transformasi data, serta pemilihan atribut.



Rajah 1 Rangka Kerja Fasa Data

Jadual 1 menunjukkan senarai set data akhir yang terpilih selepas kesemua pra-pemprosesan data dibuat. Set data akhir ini secara keseluruhannya mengandungi 15 atribut dengan bilangan data sebanyak 90,462 rekod.

Jadual 1 Set Data Akhir

Bil.	Atribut	Bil.	Atribut
1.	*Bathroom*	9.	*Price_PSLand*
2.	*Bedroom*	10.	*Car_Park_new*
3.	*Built_Up_SF*	11.	*Property_Type*
4.	*Furnishing*	12.	*Unit_Type*
5.	*Land_Size*	13.	*Tenure*
6.	*Occupancy*	14.	*Area*
7.	*Place*	15.	*Illustration*
8.	*Price*		

Setelah proses pra-pemprosesan dan pemilihan atribut telah selesai, data-data tersebut dianalisis secara visual bagi memeriksa taburan nilainya. Analisis deskriptif data dilakukan untuk melihat dengan lebih terperinci tentang ciri-ciri dan nilai setiap atribut yang telah dipilih.

ii. Model Pengelompokan

❖ Algoritma K-Min

Kajian ini menggunakan algoritma k-Min sebagai algoritma pengelompokan tanpa selia bagi mengelompok set penyenaraian hartanah palsu. Kelebihan algoritma ini adalah tahap perkomputeran yang efisien dan mudah digunakan. Algoritma ini akan membahagi kepada beberapa bilangan kelompok (k) dimana setiap titik data dimiliki oleh kelompok dengan bilangan min yang terdekat.

Saraswathi & Sheela (2014) dan Xu & Tian (2015) telah membuat kajian perbandingan terhadap pelbagai teknik pengelompokan dalam perlombongan data dan berupaya menyenaraikan kelebihan dan kekurangannya bagi setiap algoritma. Jadual 2 di bawah menyenaraikan dan menerangkan secara ringkas setiap kelebihan dan kekurangan tersebut mengikut teknik pengelompokan yang dipilih untuk digunakan dalam kajian ini.

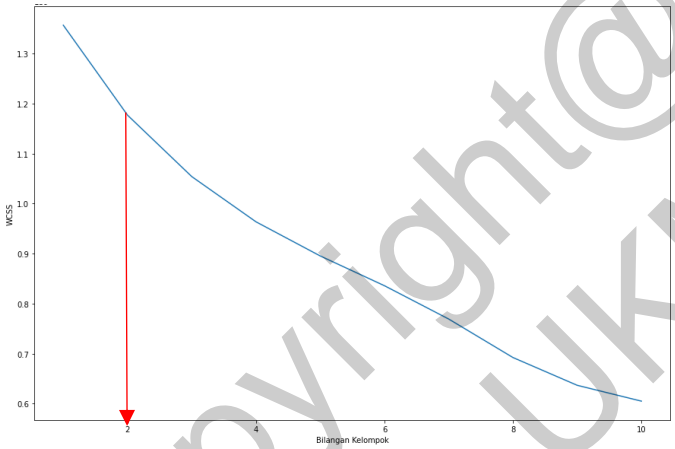
*Jadual **Error! No text of specified style in document.** Kelebihan Dan Kekurangan Algoritma Pengelompokan Mengikut Kategori*

Kategori Pengelompokan	**Contoh Algoritma**
Pembahagian	K-Min K-*Medoids* *Clustering for Large Applications* (CLARA) *Partition Around Medoids* (PAM)
Hirarki	Pemusatan (*Agglomerative*): *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH) *Clustering using Representatives* (CURE) *Robust Clustering using Links* (ROCK) *Chameleon* Pemecahbelahan (*Divisive*): DIVCLUS-T
Ketumpatan	*Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) *Ordering Points to Identify the Clustering Structure* (OPTICS) *Mean-shift*
Grid	*Statistical Information Grid-Based* (STING) *The Classical High-Dimensional Algorithm* (CLIQUE) *WaveCluster*

Model	*Self-Organizing Map* (SOM) COBWEB *Adaptive Resonance Theory* (ART)

- Penetapan nilai *k* terbaik

Teknik yang dipilih bagi menentukan nilai *k* yang optimum untuk kajian ini adalah teknik *elbow* (*elbow method*) iaitu satu teknik yang memerlukan pengamatan pada graf yang diplot. Graf tersebut merupakan plot diantara nilai kelompok *k* dan *Within Cluster Sum of Square* (WCSS) yang dikira dengan mendapatkan jarak purata antara setiap titik dalam kelompok dengan titik sentroid. Apabila graf ini ditunjukkan di dalam bentuk visual, ia dilihat seperti bentuk 'siku'. Nilai *k* terbaik adalah apabila titik graf menunjukkan nilai permulaan mendatar (Bholowalia & Kumar 2014). Julat nilai *k* yang di tentukan bagi penetapan teknik *elbow* ini adalah 2 hingga 10 dengan mengaplikasikan algoritma k-Min bagi setiap nilai *k* yang ditetapkan. Rajah 2 menunjukkan graf yang diplot bagi penentuan nilai *k*.



Rajah 1 Teknik Elbow Bagi Penetapan Bilangan Kelompok

Rajah 2 menunjukkan terdapat lengkukan WCSS yang paling ketara pada nilai *k* bersamaan 2. Walau bagaimanapun, dapat juga dilihat lengkukan WCSS pada nilai *k*=3 dan k=4. Oleh itu, bagi mengenalpasti nilai *k* terbaik untuk model k-Min, analisis tambahan iaitu penetapan parameter dan penilaian *sillhouette* dijalankan untuk menyokong dapatan nilai *k* terbaik.

- Penetapan Hiperparameter

Algoritma pengelompokan k-Min diuji dengan nilai *k*=2. Beberapa parameter perlu ditetapkan nilainya sebelum proses pengelompokan dijalankan seperti yang ditunjukkan di dalam Jadual 3 dibawah:

Jadual 3 Penetapan Hiperparameter Algoritma K-Min

Pembolehubah	**Tetapan Nilai**
k	2
Max Runs	10
Max Optimization Steps	100
Measure Type	*Numerical Measure*
Numerical Measure	*Euclidean Distance* *Correlation Similarity* *Cosine Similarity*

❖ Teknik Pengiraan Jarak

K-Min menggunakan ukuran berasaskan jarak untuk mengira nilai kesamaan setiap titik data dengan sentroid. Jarak yang paling minimum diantara titik data dan sentroid merupakan nilai yang paling optimum. Oleh itu, pengiraan jarak memainkan peranan yang sangat penting dalam proses pengelompokan. Teknik ini diperlukan bagi mengenal pasti bagaimana sesuatu data itu saling berkaitan, dan bagaimana data-data tersebut berbeza atau serupa di antara satu sama lain (Singh et al. 2013). Kajian ini menggunakan tiga teknik pengiraan jarak iaitu teknik *Euclidean Distance, Correlation Similarity* dan *Cosine Similarity*. Prestasi bagi ketiga-tiga teknik akan dibandingkan bagi memilih teknik pengiraan jarak yang terbaik.

- ***Euclidean Distance***

Teknik pengiraan jarak *Euclidean* mudah untuk dilaksanakan dan menunjukkan hasil yang hebat dalam banyak penggunaan. Walau bagaimanapun, teknik ini tidak mesra kepada pelbagai skala, justeru data perlu melalui proses normalisasi bagi mendapatkan skala yang seragam (Bora & Gupta 2014).

- ***Correlation Similarity***

Teknik pengiraan jarak *Correlation Similarity* pula, jarak dikira diantara vektor secara rawak (Bora & Gupta 2014).

- ***Cosine Similarity***

Cosine Similarity adalah jarak sudut kosinus antara vektor. Persamaan yang terhasil adalah daripada −1 (tiada persamaan), hingga 1 (sangat mirip), manakala nilai di antaranya menunjukkan kemiripan atau ketidaksamaan pertengahan.

C. Penilaian Model Pengelompokan

Penilaian model pengelompokan adalah sangat penting kerana di dalam fasa ini, kualiti pengelompokan akan dinilai dan disahkan. Proses pengelompokan adalah pencarian persamaan diantara objek tanpa mempunyai pengetahuan tentang pengagihan data atau kelompok yang betul. Hal ini mendorong penyelidikan dalam bidang penilaian dan pengesahan pengelompokan dibuat lebih daripada bidang penilaian klasifikasi (Hassani & Seidl 2017). Secara umumnya, penilaian atau pengesahan pengelompokan terdiri daripada dua jenis iaitu pengesahan dalaman (*internal validation*) dan pengesahan luaran (*external validation*).

Di dalam kajian ini, pengesahan luaran dibuat dengan kaedah membandingkan hasil kelompok dengan kelas label yang telah diklasifikasikan. Di dalam kajian ini, kelas label yang dikenalpasti adalah atribut '*Illustration'* yang menjadi indikator kepada label pengelasan hartanah palsu. Hasil kelompok dianggap baik apabila keputusan perbandingan diantaranya adalah sama (Hassani & Seidl 2017).

Bagi pengesahan dalaman, ia menggunakan maklumat dalaman proses pengelompokan itu sendiri untuk menilai kebaikan struktur pengelompokan tanpa merujuk kepada maklumat luaran. Ia juga boleh digunakan untuk menganggar bilangan kelompok dan algoritma pengelompokan yang sesuai tanpa sebarang data luaran. Kajian ini menggunakan tiga jenis metrik penilaian pengesahan dalaman iaitu indeks *Davies Bouldin, Average within Centroid Distance* dan *Sum of Squares.*

i. Indeks *Davies Bouldin* (DB)

Metrik pengukuran indeks *Davies Bouldin* (DB) adalah ukuran untuk menilai jarak antara kelompok ke-*i* dan ke-*j;* dimana jarak maksimum diperlukan bagi jarak antara kelompok (*inter-cluster*) dan jarak didalam kelompok (*intra-cluster*) yang minimum. Kelompok yang padat dan mempunyai jarak yang jauh diantara kelompok merupakan indikator kepada pengelompokan yang baik (Hofmann & Klinkenberg 2013).

ii. *Average within Centroid Distance* (AWCD)

Metrik pengukuran *Average within Centroid Distance* (AWCD) adalah jarak purata setiap titik di dalam kelompok dengan titik sentroid. Jarak purata dikira dengan mencari purata jarak berpasangan antara titik dalam setiap kelompok. Klinkenberg & Hofmann (2014) menyatakan semakin padat sesuatu kelompok, semakin kurang nilai ukuran jarak, begitu juga jika bilangan kelompok meningkat, purata jarak akan meningkat menjadikan metrik ini agak sukar untuk dijelaskan.

iii. *Sum of Square* (SS)

Sum of Square (SS) merupakan metrik penilaian ukuran yang membahagikan bilangan titik data di dalam sesuatu kelompok dengan jumlah titik data di dalam semua kelompok yang ada. Metrik ini mempunyai persamaan dengan kepadatan (*density*) kerana ia menghasilkan ukuran yang bersamaan dengan jarak setiap titik dalam kelompok (Klinkenberg & Hofmann 2014). Nilai SS yang lebih kecil menunjukkan kelompok yang lebih padat dan baik (Baarsch & Celebi 2012).

D. Keputusan Eksperimen

i. Teknik Pengiraan Jarak Terbaik

Keputusan eksperimen bagi ketiga-tiga teknik pengiraan jarak dengan tetapan $k = 2$ menggunakan tiga metrik penilaian iaitu *Davies Bouldin*, *Average within Centroid Distance* dan *Sum of Squares* telah direkodkan seperti di Jadual 4.

Jadual ***Error! No text of specified style in document.*** *Prestasi Pengelompokan-Teknik Pengiraan Jarak*

Teknik Pengiraan Jarak	***Davies Bouldin (DB)***	***Average within Centroid Distance (AWCD)***	***Sum of Squares (SS)***
Euclidean Distance	**1.849**	**12.203**	0.506
Correlation Similarity	1.897	12.284	**0.500**
Cosine Similarity	1.863	12.211	0.502

Kaedah penetapan skor dibuat dengan memberi kedudukan (*rank*) terhadap prestasi teknik penilaian DB, AWCD dan SS. Nilai prestasi terendah iaitu nilai terbaik akan diberi kedudukan 1. Jumlah skor yang terendah merupakan teknik pengiraan jarak terbaik seperti yang ditunjukkan di dalam Jadual 5.

Jadual 5 Skor Kedudukan Teknik Pengiraan Jarak

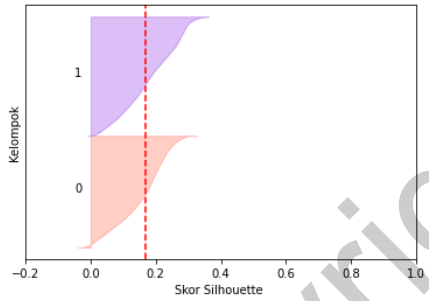
Teknik Pengiraan Jarak	***Davies Bouldin (DB)***	***Average within Centroid Distance (AWCD)***	***Sum of Squares (SS)***	**Jumlah Skor Kedudukan**	**Kedudukan Akhir**
Euclidean Distance	1	1	3	**5**	**1**
Correlation Similarity	3	3	1	**7**	2
Cosine Similarity	2	2	1	**5**	**1**

Kedudukan akhir menunjukkan teknik pengiraan jarak *Euclidean Distance* dan *Cosine Similarity* berkongsi kedudukan dengan jumlah skor bersamaan 5. *Correlation Similarity* pula mencatatkan jumlah skor 7 dan berada di kedudukan terakhir. Teknik pengiraan jarak *Euclidean*

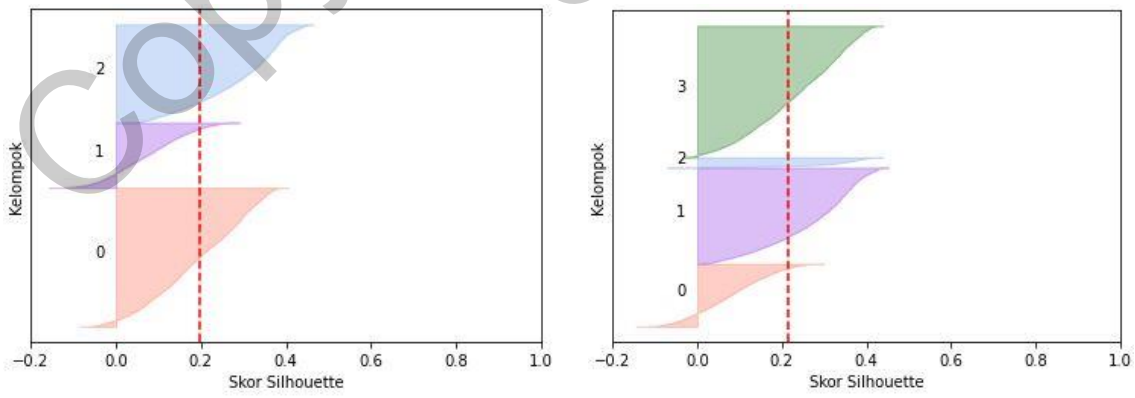
mendapat skor 1 sebanyak dua kali iaitu bagi nilai DB dan AWCD, berbanding teknik *Cosine Similarity* yang mempunyai skor 2 bagi kedua-dua metrik penilaian. Justeru, bagi kajian ini, teknik *Euclidean Distance* dipilih sebagai teknik pengiraan jaran terbaik.

ii. Penilaian *Silhouette* bagi Pengelompokan K-Min

Penilaian s*ilhouette* merupakan analisis tambahan bagi memeriksa prestasi pengelompokan. Hasil pengelompokan k-Min dengan teknik pengiraan *Euclidean Distance* dengan nilai *k*=2 dianalisis mengunakan plot nilai pekali s*ilhouette* dengan membandingkan nilai purata pekali *silhoutte* keseluruhan kelompok dengan nilai skor s*ilhouette* setiap data. Seperti yang ditunjukkan dalam Gambarajah 3, nilai purata *silhoutte* yang diperolehi bagi keseluruhan kelompok adalah 0.17, ditanda dengan garisan merah dan didapati nilai skor s*ilhouette* bagi setiap kelompok yang terhasil melebihi garis nilai purata s*ilhouette.* Keadaan ini menandakan pengelompokan yang terhasil adalah baik dan optimal.



Rajah 2 Plot Silhouette Bagi Pengelompokan k=2



Rajah ***Error! No text of specified style in document.*** *Plot Silhouette Bagi Pengelompokan k=3 dan k=4*

Analisis penilaian *silhouette* ini menunjukkan prestasi pengelompokan k-Min dengan nilai *k*=2, menggunakan teknik pengiraan jarak *Euclidean Distance* menandakan pengelompokan yang terhasil adalah baik dan optimal. Analisis ini juga dibuat bagi membandingkan prestasi pengelompokan apabila

nilai *k* atau kelompok ditetapkan kepada nilai 3 dan 4 seperti yang ditunjukkan di Rajah 4. Didapati turun naik (*fluctuation*) saiz plot *silhouette* adalah besar dan ketara bagi kedua-dua nilai k=3 dan k=4. Ketebalan plot *silhouette* yang mewakili setiap kelompok juga dilihat tidak seragam dan menunjukkan prestasi pengelompokan yang tidak memuaskan. Sehubungan itu, pembangunan model pengelompokan k-Min untuk meramal penyenaraian hartanah palsu akan menggunakan nilai *k*=2.

IV. ANALISIS KELOMPOK DAN MODEL PENGELASAN PEYENARAIAN HARTANAH PALSU

A. Saiz Kelompok

Jadual 6 menunjukkan hasil kelompok yang diperolehi menggunakan algoritma k-min. Dapat dilihat daripada jadual tersebut pembahagian kelompok adalah agak seimbang dimana saiz kelompok 0 menunjukkan saiz 45.06% (40,251 senarai) manakala kelompok 1 mencatatkan 54.94% (50,211 senarai).

Jadual 6 Hasil Kelompok Algoritma K-Min

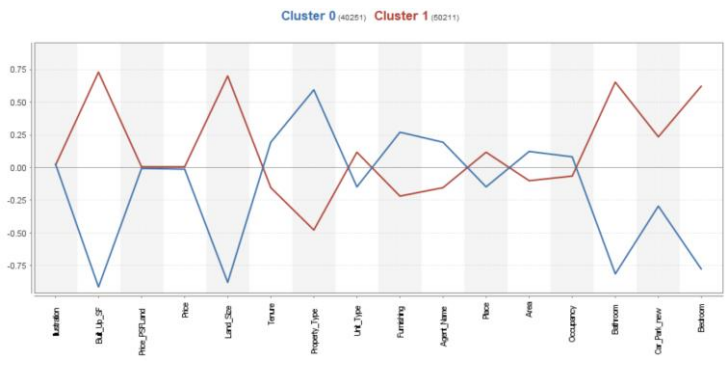
Kelompok	Teknik Pengiraan Jarak	Bil. Penyenaraian Hartanah	Saiz Kelompok (%)
0	*Euclidean Distance*	40,251	45.06
1		50,211	54.94

B. Jadual Sentroid, Carta Sentroid, Analisis *Heatmap* dan Taburan Kelompok

Jadual sentroid menunjukkan nilai purata titik sentroid bagi setiap atribut dalam setiap kelompok. Atribut-atribut ini menunjukkan atribut yang mempunyai ciri penting yang mempengaruhi pembentukan kelompok. Atribut dengan nilai purata sentroid yang tinggi menggambarkan jarak yang hampir diantara atribut tersebut dengan titik sentroid dan begitu juga sebaliknya, seperti yang ditunjukkan di Jadual 7.

Jadual 7 Jadual Sentroid Pengelompokan

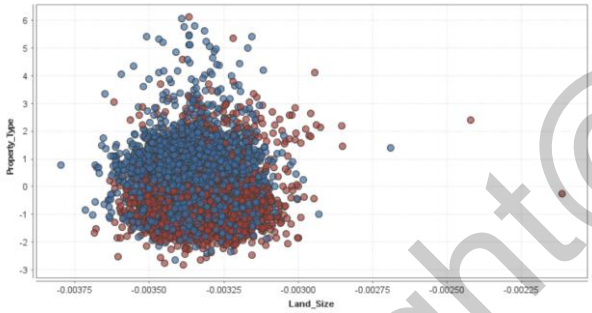
Atribut	Kelompok	
	0	1
Illustration	0.029	0.018
Built_Up_SF	-0.912	0.731
Price_PSFLand	-0.004	0.004
Price	-0.010	0.008
Land_Size	-0.876	0.702
Tenure	0.192	-0.154
Property_Type	0.592	-0.475
Unit_Type	-0.148	0.119
Furnishing	0.270	-0.216



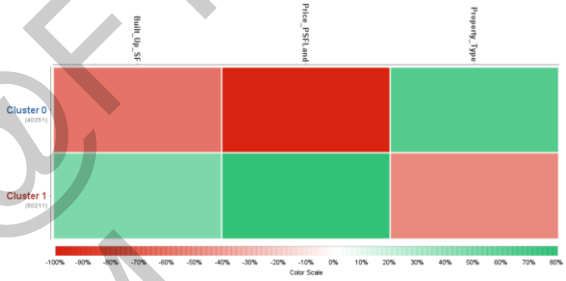
Rajah 5 Carta Sentroid Pengelompokan

Agent_Name	0.194	-0.156
Place	-0.150	0.120
Area	0.123	-0.099
Occupancy	0.082	-0.066
Bathroom	-0.816	0.654
Car_Park_new	-0.294	0.235
Bedroom	-0.776	0.622

Selain daripada jadual sentroid, nilai purata titik sentroid setiap atribut bagi setiap kelompok dapat juga dipersembahkan melalui carta sentroid. Grafik ini lebih jelas menunjukkan perbezaan nilai di antara dua kelompok. Di dalam Rajah 5, dapat dilihat atribut *Price_PSFLand*, *Land_Size*, *Property_Type*, *Bathroom*, C*ar_Park_new*, *Built_Up_SF* dan *Bedroom* dibezakan dengan sangat baik iaitu perbezaan nilai purata sentroid yang jauh diantara kelompok yang terhasil.



Rajah 6 Plot Taburan Kelompok



Rajah 7 Rajah Heat Map Pengelompokan

Fungsi plot taburan (*scatter plot*) kelompok yang disediakan *Rapidminer* ini memaparkan plot taburan dua atribut terpenting bagi kelompok yang terhasil. Rajah 6 menunjukkan plot taburan bagi kelompok 0 dan 1 dan dua atribut yang penting bagi kelompok iaitu *Property_Type* dan *Land_Size*. Didapati pasangan atribut ini menunjukkan hubungan atau korelasi yang tinggi.

Rajah *heat map* boleh disamakan dengan sebuah hamparan *excel* (*excel spreadsheet)* yang mengandungi warna, dan bukannya nombor. Rajah 7 menunjukkan tiga atribut bagi kelompok 0 dan 1 dengan indikator skala warna hijau dan merah. Warna hijau menggambarkan nilai data yang tinggi manakala warna merah menggambarkan nilai data yang rendah di dalam set data bagi sesuatu atribut. Warna hijau yang paling pekat adalah purata nilai data yang paling tinggi dan warna merah yang paling pekat adalah purata nilai data yang terendah dan begitu juga sebaliknya. Daripada rajah tersebut dapat dilihat tiga atribut penting yang mempengaruhi pembentukan kelompok 0 dan 1 iaitu atribut *Built_Up_SF*, *Property_Type* dan *Price_PSFLand*.

C. Analisis Diskriptif dan Penetapan Label Kelas

Setelah analisis diskriptif statistik kelompok dibuat, pengekstrakan ciri-ciri kelompok dibuat bagi membuat penetapan kelas label berdasarkan analisis diskriptif yang telah dijalankan. Jadual 8 merupakan ringkasan analisis diskriptif statistik kelompok.

Jadual 8 Jadual Ringkasan Analisis Diskriptif Statistik

Kelompok	*Illustration*	%	*Property_Type*	%	*Area*	%	*Place*	%	*Price*	%	*Price_PSFLand*	%	*Built_Up_SF*	%	*Land_Size*	%
0	0	43.06	*Condominium/Apartment/Serviced Residence*	34.77	Ampang	17.56	KLCC	36.63	200,243 – 400,216	18.77	< 536.380	13.65	< 919.500	19.73	< 930.50	19.94
	1	1.31														
1	0	53.97	*Terrace House*	41.85	Bangi	17.33	X2 Residency	12.83	400,216 – 600,189	18.65	< 213.105	17.81	> 2760.500	19.98	> 2780.500	19.80
	1	1.00														

Bersambung..

..sambungan

Kelompok	*Bathroom*	%	*Bedroom*	%	*Car_Park_new*	%	*Furnishing*	%	*Occupancy*	%	*Tenure*	%	*Unit_Type*	%	*Area*	*Agent_Name*	%
0	2	21.38	3	24.16	2	39.58	Partly Furnished	17.17	Vacant	41.13	Freheold	22.8	Intermediate Lot	43.01	188 Suites Service Apartment KLCC	Alex Cheah	17.13
1	4	29.52	4	41.78	2	51.68	Unfurnished	33.52	Vacant	53.87	Freeheold	38.55	Intermediate Lot	50.03	16 Quartz, Taman Melati	Christina Tan	16.49

Daripada Jadual 8, dapat diperhatikan bagi atribut '*Illustration'*, nilai '0' iaitu 'benar' mempengaruhi kedua-dua kelompok dengan peratusan tertinggi adalah di kelompok 1 iaitu sebanyak 53.97%. Nilai '1' ataupun 'palsu' juga mempengaruhi kedua-dua kelompok tetapi peratusan tertinggi nilai ini adalah milik kelompok 0 iaitu 1.31% berbanding 1% (kelompok 1). Berdasarkan pengalaman dan maklumat daripada pakar hartanah yang dirujuk, perkataan '*illustration'* yang terdapat di penyenaraian hartanah merupakan ciri penyenaraian hartanah palsu. Label kelas '*Illustration'* set data asal yang terhasil daripada pengekstrakan perkataan '*illustration'* daripada atribut '*Desc'*, mengandungi nilai bilangan penyenaraian hartanah palsu sebanyak 2095 dan bilangan penyenaraian hartanah benar sebanyak 88367. Maklumat ini memberi gambaran awalan yang penting bagi penentuan kelas palsu.

Bagi atribut *Property_Type*, nilai '*Condominium/Apartment/Serviced Residence'* mempengaruhi kelompok 0 dengan peratusan penyenaraian hartanah adalah 34.77%. Manakala kelompok 1 di pengaruhi oleh nilai '*Terrace House'* dengan peratusan yang tinggi iaitu 41.85%.

Atribut '*Price'* menunjukkan nilai dalam julat harga ke tiga (urutan menaik) iaitu 400,216 – 600,189 mempengaruhi kelompok 1 dan 0 dengan kelompok 1 mencatatkan peratus tertinggi (18.65%) manakala kelompok 0 adalah 12.78%. Kelompok 0 juga didominasi oleh nilai harga 200,243 – 400,216 iaitu 18.77%.

Walau bagaimanapun, didapati atribut '*Price_PSFLand'* tidak menunjukkan corak pembahagaian kelompok yang sama dengan atribut '*Price'*. Nilai julat harga tertinggi iaitu >536.380 mendominasi kelompok 0 dan julat harga terendah iaitu <213.105 mendominasi kelompok 1.

Nilai 'Ampang' bagi atribut '*Area'* telah mempengaruhi kedua-dua kelompok, dengan peratusan tertinggi nilai ini adalah milik kelompok 0 iaitu 17.56% berbanding 10.26%. Bangi mempengaruhi kelompok 1 dengan peratusan yang tinggi iaitu sebanyak 17.33%. Nilai 'Bukit Jalil', 'Bandar Sunway' dan 'Bandar Baru Ampang' mencatatkan peratusan di kelompok 0 melebihi kelompok 1. Keadaan ini menunjukkan kontradik dengan ciri kelompok 0 yang lain seperti nilai bagi atribut '*Price'* yang menunjukkan ciri harga dalam julat yang rendah. Hartanah di kawasan Ampang, Bandar Sunway, Bandar Baru Ampang, Bukit Jalil dan Bandar Sunway adalah kawasan yang mempunyai nilai hartanah yang tinggi dan ini menunjukkan ciri-ciri penyenaraian hartanah yang palsu.

Nilai di dalam atribut '*Place'* menunjukkan corak pembahagian kelompok yang sama dengan atribut '*Area'*. Kesemua nilai yang tinggi peratusannya adalah dimiliki oleh kelompok 0. 'KLCC' mempengaruhi kelompok 0 dengan nilai 36.63%, dan keadaan ini adalah selari dengan nilai 'Ampang' di atribut '*Area'.* Ini juga merupakan ciri yang bertentangan dengan ciri kelompok 0 yang lain.

Nilai-nilai yang dimiliki oleh kedua-dua kelompok bagi atribut '*Bathroom'*, '*Bedroom'*, '*Car_Park_new'*, '*Built_Up_SF* 'dan '*Land_Size'* menunjukkan ciri yang bersesuaian dengan ciri nilai '*Property_Type'* kelompok 0 dan 1. Sebagai contoh, nilai 2 bagi '*Bathroom'* merupakan nilai tertinggi bagi kelompok 0, dan kelompok 0 juga didominasi oleh hartanah jenis '*Condominium/Apartment/Serviced Residence'.* Kelompok 1 juga dipengaruhi oleh nilai yang bersesuaian bagi atribut-atribut tersebut.

Berdasarkan analisis diskriptif dan pengekstrakan ciri yang telah dibuat, dapat disimpulkan dua label kelas daripada kelompok yang terhasil adalah kelompok penyenaraian hartanah palsu dan kelompok penyenaraian hartanah benar seperti yang diterangkan di Jadual 9 dibawah:

Jadual 9 Kelas Label Berdasarkan Pengelompokan

Kelompok	Kelas Label	Bilangan Data
0	Penyenaraian Hartanah Palsu	40,251
1	Penyenaraian Hartanah Benar	50,211

D. Model Pengelasan Penyenaraian Hartanah Palsu

Kajian ini akan membangunkan model pengelasan menggunakan algoritma Pohon Keputusan, Mesin Sokongan Vektor dan algoritma Hutan Rawak menggunakan perisian *Rapidminer*. Bilangan data yang terlibat adalah sebanyak 90,462 data dengan 16 atribut yang akan menjadi input kepada model pengelasan.

iii. Keputusan Eksperimen

Prestasi keputusan bagi ketiga-tiga algoritma pengelasan diukur dengan nilai ketepatan (*accuracy*) dan dijalankan keatas set data pengelompokan serta set data asal. Jadual 10 menunjukkan perbandingan purata prestasi ketepatan bagi kedua-dua set data menggunakan ketiga-tiga algoritma pengelasan.

Jadual 10 Perbandingan Keputusan Prestasi Model Pengelasan

	Set Data Pengelompokan	Set Data Asal
Algoritma Pengelasan	**Ketepatan (%)**	
Mesin Sokongan Vektor	**99.37**	97.68
Pohon Keputusan	96.97	97.70
Hutan Rawak	97.45	97.72
Purata	**97.93**	**97.70**

Dapat dilihat dari Jadual 10, prestasi pengelasan set data pengelompokan algoritma mesin sokongan vektor mencatatakan peratus nilai ketepatan yang lebih tinggi (99.37%) daripada nilai ketepatan set data asal (97.68%). Sebaliknya, bagi algoritma pengelasan pohon keputusan dan hutan rawak, didapati nilai peratus ketepatan bagi set data asal melebihi set data pengelompokan. Dapat dilihat disini bahawa purata peratus ketepatan bagi set data pengelompokan melebihi daripada purata peratus ketepatan set data asal iaitu 97.93% mengatasi 97.70%. Algoritma pengelasan mesin sokongan vektor bukan sahaja mempunyai nilai ketepatan tertinggi bagi set data pengelompokan, malah nilai peratus ketepatannya melebihi daripada ketepatan set data asal. Justeru, dapat diputuskan bahawa model pengelasan algoritma mesin sokongan vektor adalah model pengelasan terbaik bagi pengelasan penyenaraian hartanah palsu.

iv. Ujian Statistik *t* (*Statistical t-Test*)

Ujian statistik-*t* yang dipilih bagi mengetahui prestasi algoritma pengelas adalah ujian-t berpasangan. Ujian ini bertujuan untuk menilai prestasi sesuatu algoritma pengelas bagi menentukan perbezaan nilai ketepatan yang signifikan secara statistik (Jovanovic & Vukicevic 2014). Algoritma mesin sokongan vektor dijadikan asas perbandingan berpasangan kepada algoritma pohon keputusan dan hutan rawak dengan nilai keyakinan ditetapkan kepada 0.05.

A	B	C	D
	0.994 +/- 0.001	0.970 +/- 0.002	0.975 +/- 0.002
0.994 +/- 0.001		0.000	0.000
0.970 +/- 0.002			0.000
0.975 +/- 0.002			

Rajah 8 Keputuan Ujian-T Berpasangan

Rajah 8 menunjukkan keputusan ujian-t berpasangan yang diperolehi. 0.994, 0.970 dan 0.975 merupakan nilai ketepatan ketiga-tiga algoritma iaitu SVM, DT dan RF masing-masing. Nilai yang mempunyai latarbelakang berwarna adalah nilai yang signifikan kerana nilai *p* (0.001 dan 0.002) adalah kurang daripada 0.05. Dapat dilihat dari ujian ini terdapat perbezaan yang signifikan diantara pasangan SVM dan DT serta SVM dan RF. Ini membuktikan bahawa algoritma SVM adalah pengelas terbaik dengan keputusan ujian-t berpasangan yang signifikan secara saintifik.

E. Perbincangan

Algoritma pengelompokan k-Min telah diuji keatas 90,462 rekod data penyenaraian hartanah menggunakan tiga teknik pengiraan jarak iaitu pengiraan jarak iaitu *Euclidean Distance*, *Correlation Similarity* dan *Cosine Similarity*. Penetapan nilai *k* terbaik dibuat menggunakan teknik '*elbow'* dengan

nilai *k* bersamaan 2. Penilaian prestasi pengelompokan telah dibuat dengan menggunakan tiga teknik penilaian iaitu *Davies Bouldin* (DB), *Average Within Centroid Distance* (AWCD) dan *Sum of Squares* (SS) dan teknik *Euclidean Distance* dipilih sebagai teknik pengiraan jarak terbaik. Analisis tambahan penilaian dibuat dengan ujian plot nilai pekali s*ilhouette* dan plot nilai pekali s*ilhouette* yang terhasil menggambarkan dua kelompok yang terbentuk adalah kelompok yang baik apabila kedua-duanya melebihi nilai purata indeks pekali s*ilhouette*. Ujian plot nilai pekali s*ilhouette* juga dianalisis dengan nilai *k*=3 dan 4, dan didapati prestasi kelompok keduanya adalah tidak memuaskan.

Analisis kelompok dijalankan bagi mengetahui perbezaan ciri setiap kelompok. Saiz kelompok, jadual dan carta sentroid dan plot taburan kelompok dianalisa begitu juga analisis diskriptif statistik setiap atribut dijalankan. Ciri-ciri atau atribut penting bagi kedua-dua set data pengelompokan dan set data asal bagi penyenaraian hartanah palsu dapat dikenalpasti. Set data asal Jadual 11 menerangkan ciri-ciri atau atribut penting bagi kedua-dua set data pengelompokan dan set data asal.

Jadual 11 Ciri Penting Penyenaraian Hartanah Palsu

Set Data	**Penjanaan Kelas Label**	**Ciri/Atribut Penting Penyenaraian Hartanah Palsu**
Set Data Asal	Berdasarkan atribut '*Illustration'*	Perkataan '*illustration'*
Set Data Pengelompokan k-Min	Berdasarkan pengelompokan k-Min	Atribut *Illustration, Property_Type, Area, Price, Place, Built_UP_SF*

Model pengelasan ramalan set data penyenaraian hartanah dijalankan menggunakan algoritma pengelasan pohon keputusan, mesin sokongan vektor dan hutan rawak keatas kedua-dua set data pengelompokan dan set data asal. Keputusan prestasi model pengelasan mendapati purata ketepatan ramalan set data pengelompokan mengatasi purata ketepatan set data asal iaitu 97.93% mengatasi 97.70%. Set data asal hanya bergantung kepada atribut '*Illustration'* sahaja dalam penentuan penyenaraian hartanah palsu manakala set data pengelompokan mempunyai banyak atribut lain yang mempengaruhi kelas label hartanah palsu. Menurut Tanwani et al. ciri-ciri penting bagi sebuah set data adalah atributnya, kelas label serta bilangan data (Tanwani et al. 2009). Justeru, set data yang mempunyai atribut yang lebih banyak daripada set data yang bergantung dengan hanya satu atribut adalah lebih baik dari segi ketepatan ramalan.

Nilai pekali satah hiper (*hyperplane coefficient*) bagi algoritma SVM telah dikira bagi melihat perkaitan atau korelasi setiap atribut. Nilai tersebut merupakan nilai pemberat (*weight*) setiap fitur atau atribut yang dapat menentukan atribut utama yang digunakan dalam pengelasan SVM. Enam atribut

yang penting yang dikenalpasti daripada set data pengelompokan telah dikira nilai pekali seperti yang ditunjukkan di Jadual 12.

Jadual 12 Nilai Pekali Satah Hiper SVM

Bil	Atribut	Nilai pekali SVM
1	*Property_Type*	2.621
2	*Area*	1.164
3	*Price*	-0.023
4	*Place*	-0.541
5	*Price_PSFLand*	0.000
6	*Built_Up_SF*	-3.789

Daripada Jadual 12 dapat dilihat atribut *Property_Type* dan *Area* mempunyai nilai pekali yang tinggi iaitu 2.621 dan 1.164 dalam mengelas penyenaraian hartanah palsu. Atribut *Price*, *Place* dan *Built_Up_SF* mempunyai nilai pemberat yang negatif dimana nilai negatif ini telah mengecilkan nilai atribut tersebut. Persamaan x^i x θ^i menerangkan keadaan ini dimana x adalah nilai pemberat dan θ adalah nilai input atau atribut, menjadikan nilai atribut tersebut juga adalah negatif.

Algoritma mesin sokongan vektor dipilih sebagai model terbaik bagi penyenaraian hartanah palsu dengan nilai ketepatan paling tinggi bagi kedua-dua set data. SVM mempunyai kelebihan tersendiri iaitu algoritma yang dinamik yang boleh menyelesaikan pelbagai masalah pengelasan. SVM juga memainkan peranan penting dalam bidang pengelasan dengan fungsinya yang mengubah setiap item data asal kepada titik dalam ruang fitur berdimensi tinggi (Pranita & Gadhe 2013). Set data penyenaraian hartanah ini dapat di kelaskan dengan baik oleh algoritma SVM. Algoritma SVM memberikan hasil yang sangat baik dari segi ketepatan apabila data boleh dipisahkan secara linear atau tidak linear dimana garis satah hiper memaksimumkan margin pemisahan antara kelas (Afifi et al. 2013).

Berbanding algoritma pohon keputusan yang mempunyai nilai ketepatan terendah, ia tidak seperti pendekatan pengelasan lain yang menggunakan set ciri secara bersama untuk melaksanakan pengelasan dalam satu langkah keputusan. Algoritma pohon keputusan jarang memberikan ketepatan ramalan yang optimum dengan data yang ada (Hastie 2009). Ini mungkin disebabkan sifat algoritma ini yang 'tamak' yang menyebabkan perubahan keputusan yang agak signifikan jika pertambahan atau pengurangan satu baris data dibuat. Walau bagaimanapun, ketepatan algoritma pohon keputusan bagi penyenaraian hartanah palsu ini tetap dianggap baik dan tidak menunjukkan perbezaan ketara berbanding dua algoritma yang lain.

V. KESIMPULAN DAN CADANGAN

Bab ini memberi penceritaan menyeluruh terhadap kajian yang dilaksanakan. Ia merangkumi rumusan kajian, kekangan yang dihadapi sepanjang proses kajian, sumbangan kajian terhadap domain penyenaraian hartanah palsu dan cadangan perluasan kajian pada masa hadapan.

A. Rumusan

Penyenaraiaan hartanah palsu merupakan masalah yang dihadapi dalam sektor hartanah tetapi ia tidak diberi perhatian yang mendalam disebabkan beberapa kekangan. Kajian ini dilaksanakan bertujuan untuk membangunkan model ramalan penyenaraian hartanah palsu menggunakan pembelajaran mesin. Tiga objektif utama yang telah ditetapkan pada awal kajian telah dicapai melalui kajian ini.

Objektif pertama iaitu membangunkan model pengelompokan k-Min yang optimum bagi penyenaraian hartanah palsu. Sebelum pembangunan model pengelompokan dibuat, beberapa aktiviti kajian dibuat iaitu mengenalpasti masalah, menentukan objektif dan skop kajian serta menjalankan kajian literasi. beberapa aktiviti pra-pempresasan data dijalankan semasa penyediaan data dibuat dan seterusnya model pengelompokan dibangunkan. Hasilnya, satu model pengelompokan menggunakan algoritma pengelompokan k-min yang optimum berjaya dibangunkan.

Bagi mencapai objektif yang kedua iaitu mengenalpasti kelompok penyenaraian hartanah palsu dan atribut penting dalam setiap kelompok, analisis kelompok dibuat dengan terperinci terhadap kelompok yang terhasil. Menerusi analisis kelompok yang dibuat dapat dikenalpasti atribut-atribut yang menyumbang kepada pembentukan kedua-dua kelompok tersebut.

Seterusnya objektif ketiga adalah membangunkan model pengelasan ramalan penyenaraian hartanah palsu berdasarkan ciri penting dari hasil pengelompokan. Model pengelasan ini dibangunkan menggunakan tiga algoritma pengelasan iaitu algoritma pohon keputusan, mesin sokongan vektor dan hutan rawak dan prestasi diukur melalui nilai ketepatan. Kajian mendapati nilai peratus ketepatan algoritma mesin sokongan vektor bagi set data pengelompokan mengatasi nilai peratus ketepatan bagi set data asal. Sehubungan itu, model pengelasan mesin sokongan vektor dipilih sebagai model terbaik di dalam ramalan penyenaraian hartanah palsu.

B. Batasan Kajian

Model pengelompokan di dalam kajian ini dibangunkan menggunakan dua set data penyenaraian hartanah daripada agensi penyenaraian hartanah *DurianProperty* dan *EdgeProp*. Kedua-dua set data ini

mempunyai atribut serta rekod tiada nilai yang sangat banyak seperti yang diterangkan di Bab 3. Penerokaan pengetahuan dibuat bagi menggantikan rekod tiada nilai tersebut melalui hanya satu sumber atribut sahaja iaitu atribut '*Desc'* yang mengandungi teks yang tidak berstruktur. Model pengelompokan mungkin akan memberikan hasil yang lebih baik jika rekod data lebih berstruktur dan jumlah atribut serta rekod tiada nilai ini tidak terlalu banyak.

Algoritma pengelompokan yang digunakan di dalam kajian ini adalah algoritma k-Min sahaja dengan mengaplikasikan tiga teknik pengiraan jarak iaitu *Euclidean Distance*, *Cosine Similarity* dan *Correlation Similarity*. Banyak lagi algoritma pengelompokan serta teknik pengiraan jarak lain yang boleh diterokai. Kajian ini menghadapi kesukaran dalam mengaplikasikan beberapa algoritma yang lain apabila set data didapati tidak sesuai digunakan semasa eksperimen dijalankan.

Selain itu, pembangunan dan analisis model pengelompokan yang dilaksanakan kebanyakannya menggunakan perisian *Rapidminer* dan sebahagian kecilnya menggunakan platfom *Google Colab* berasaskan pengaturcaraan *python.* Model pengelasan dibangunkan menggunakan perisian *Rapidminer*. Terdapat banyak lagi aplikasi dan perisian pembelajaran mesin di pasaran bagi pembinaan model dan perbandingan prestasi pemodelan yang boleh diterokai untuk mendapatkan keputusan analisis dan penerokaan pengetahuan yang lebih mendalam.

C. Cadangan Perluasan Kajian

Berdasarkan batasan kajian yang telah dibincangkan, beberapa cadangan dikemukakan bagi tujuan perluasan kajian. Cadangan pertama adalah sumber data daripada agensi hartanah yang lain dengan data yang lebih lengkap. Terdapat banyak lagi agensi hartanah yang menyediakan platform pengiklanan dalam talian dimana mungkin data-data ini lebih lengkap dan sesuai digunakan bagi tujuan kajian ramalan penyenaraian hartanah palsu.

Kajian ini dapat dijadikan sebagai kajian asas dalam membuat ramalan penyenaraian hartanah palsu menggunakan kaedah pembelajaran mesin. Sehubungan itu, kajian ini boleh diperluaskan lagi dengan menggunakan pelbagai algoritma pengelompokan yang lain samada daripada jenis pemetaan yang sama dengan k-min seperti k-mediods mahupun pendekataan yang lain seperti pendekatan hirarki, model, grid dan densiti seperti algoritma BIRCH, DBSCAN dan SOM. Bagi model pengelasan, kajian boleh diperluaskan lagi menggunakan selain daripada tiga algoritma yang digunakan seperti algoritma Naive Bayes mahupun Deep Learning.

PENGHARGAAN

Penulis ingin mengucapkan ribuan terima kasih kepada Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia atas segala bimbingan, pengetahuan dan peluang yang diberikan kepada penulis untuk menyelesaikan kajian ini.

RUJUKAN

Abdulaziz, M.H. & Zeki, A.M. 2020. Prediction of Real Estate Land Prices in the Kingdom of Bahrain. *2020 International Conference on Decision Aid Sciences and Application, DASA 2020* 1220–1223.

Afifi, A., Zanaty, E.A. & Ghoniemy, S. 2013. Improving the Classification Accuracy Using Support Vector Machines (Svms) With New Kernel. *Journal of Global Research in Computer Science 4*(2): 1–7.

Baarsch, J. & Celebi, M.E. 2012. Investigation of internal validity measures for K-means clustering. *Lecture Notes in Engineering and Computer Science 2195*: 471–476.

Bartschat, A., Reischl, M. & Mikut, R. 2019. Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9*(4): 1–14.

Bholowalia, P. & Kumar, A. 2014. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications 105*(9): 975–8887.

Bora, D.J. & Gupta, A.K. 2014. Effect of Different Distance Measures on the Performance of K-Means Algorithm : An Experimental Study in Matlab. *International Journal of Computer Science and Information Technologies 5*(2): 2501–2506.

Castelli, M., Dobreva, M., Henriques, R. & Vanneschi, L. 2020. Predicting Days on Market to Optimize Real Estate Sales Strategy. *Complexity 2020*

Cheng, X., Yuan, M., Xu, L., Zhang, T., Jia, Y., Cheng, C. & Chen, W. 2016. Big data assisted customer analysis and advertising architecture for real estate. *2016 16th International Symposium on Communications and Information Technologies, ISCIT 2016* 312–317.

Das, A. & Rad, P. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey 1–24. http://arxiv.org/abs/2006.11371.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Int Conf on Knowledge Discovery and Data Mining* 82–88.

Hand, D., Mannila, H. & Smyth, P. 2001. *Principles of Data Mining Cambridge*. *MIT Press* Vol. 2001 http://link.springer.com/10.1007/978-1-4471-4884-5.

Hassani, M. & Seidl, T. 2017. Using internal evaluation measures to validate the quality of diverse

stream clustering algorithms. *Vietnam Journal of Computer Science 4*(3): 171–183.

Hastie, T. et. all. 2009. The Elements of Statistical Learning. *The Mathematical Intelligencer 27*(2): 83–85. http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf.

Hofmann, M. & Klinkenberg, R. 2013. *Rapid Miner Data Mining Use Cases and Business Analytics Applications*.

Hot, E. & Popović-Bugarin, V. 2016. Soil data clustering by using K-means and fuzzy K-means algorithm. *Telfor Journal 8*(1): 56–61.

Jovanovic, M.Z. & Vukicevic, M. 2014. Chapter 24 Using RapidMiner for Research : (October)

Kamil, I. & Pharmasetiawan, B. 2019. Fingerprint presence fraud detection using tight clustering on employee's presence and activity data. *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019 6*: 480–483.

Kigerl, A. 2016. Cyber crime nation typologies: K-means clustering of countries based on cyber crime rates. *International Journal of Cyber Criminology 10*(2): 147–169.

Kiruthiga, S., Kola Sujatha, P. & Kannan, A. 2014. Detecting cloning attack in Social Networks using classification and clustering techniques. *2014 International Conference on Recent Trends in Information Technology, ICRTIT 2014* 1–6.

Klinkenberg, R. & Hofmann, M. 2014. *Rapidminer Data Mining Use Cases and Business Analytics Applications*.

Li, T., Akiyama, T. & Wei, L. 2021. Constructing a highly accurate price prediction model in real estate investment using LightGBM. *Proceedings - 4th International Conference on Multimedia Information Processing and Retrieval, MIPR 2021* (1): 273–276.

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., Olaniyan, D. & Olawole, O. 2019. Data Clustering: Algorithms and Its Applications. *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019* (ii): 71–81.

Park, B. & Kwon Bae, J. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications 42*(6): 2928–2934. http://dx.doi.org/10.1016/j.eswa.2014.11.040.

Pranita, N.K. & Gadhe, D.. 2013. Er （救急外来） の特徴 Er （救急外来） でのアプローチ Er でのアプローチ Er でのアプローチのポイント Er でのアプローチ （非緊急）. *International Journal of Engineering Research & Technology 2*(2): 1–8.

Sangalli, V.A., Martinez-Munoz, G. & Canabate, E.P. 2020. Identifying cheating users in online courses. *IEEE Global Engineering Education Conference, EDUCON 2020-April*: 1168–1175.

Sani, N.S., Shamsuddin, I.I.S., Sahran, S., Rahman, A.H.A. & Muzaffar, E.N. 2018. Redefining selection of features and classification algorithms for room occupancy detection. *International*

Journal on Advanced Science, Engineering and Information Technology 8(4–2): 1486–1493.

Saraswathi, S. & Sheela, M.I. 2014. A Comparative Study of Various Clustering Algorithms in Data Mining. *IJCMS 3*(11): 422–428.

Seifollahi, S., Bagirov, A., Layton, R. & Gondal, I. 2017. Optimization Based Clustering Algorithms for Authorship Analysis of Phishing Emails. *Neural Processing Letters 46*(2): 411–425.

Singh, A., Rana, A. & Pradesh, U. 2013. K-means with Three different Distance Metrics. *International Journal of Computer Applications 67*(10): 13–17.

Tanwani, A.K., Afridi, J., Shafiq, M.Z. & Farooq, M. 2009. Guidelines to select machine learning scheme for classification of biomedical datasets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5483 LNCS*: 128–139.

Xu, D. & Tian, Y. 2015. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science 2*(2): 165–193.

Yang, C.H., Lee, B. & Lin, Y. Da. 2022. Effect of Money Supply, Population, and Rent on Real Estate: A Clustering Analysis in Taiwan. *Mathematics*

Yu, Y., Lu, J., Shen, D. & Chen, B. 2021. Research on real estate pricing methods based on data mining and machine learning. *Neural Computing and Applications 33*(9): 3925–3937. https://doi.org/10.1007/s00521-020-05469-3.

Zhang, Y., Liu, G., Zheng, L. & Yan, C. 2019. A hierarchical clustering strategy of processing class imbalance and its application in fraud detection. *Proceedings - 21st IEEE International Conference on High Performance Computing and Communications, 17th IEEE International Conference on Smart City and 5th IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2019* 1810–1816.