RAMALAN KEJAYAAN FILEM MENGGUNAKAN KAEDAH PERLOMBONGAN DATA

DYANA NAJIHA BT MOHD NASIR

PROF.MADYA. DR SUHAILA BT ZAINUDIN

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Kajian ini dijalankan bertujuan untuk membangunkan model ramalan kejayaan filem berdasarkan skor IMDB filem menggunakan kaedah perlombongan data dan algoritma pembelajaran mesin menggunakan perisian RStudio. Kajian ini menggunakan teknik pengelasan iaitu Pohon Keputusan(DT), *Random Forest* dan *K-Nearest Neighbour*(KNN). Pembangunan model ini memfokuskan kepada 5043 data filem yang merangkumi 100 tahun dan 66 negara. Selain itu, metodologi CRISP-DM telah digunakan berdasarkan data filem terdahulu untuk mengurangkan tahap ketidakpastian yang tertentu terhadap hasil skor IMBD filem. Beberapa kriteria dalam mengira kejayaan filem termasuk pelakon, pengarah, skor imdb , bilangan kritikan dan bilangan ulasan filem . Oleh kerana pembuatan filem melibatkan pelaburan yang besar, ramalan filem memainkan peranan penting dalam industri filem. Hasil keputusan dari model ini membantu pembikin filem mengetahui dan memahami kriteria yang penting dalam menjana kejayaan sesebuah filem. Projek ini juga menunjukkan kepentingan analitik data ramalan dan preskriptif kepada sistem maklumat untuk membantu industri filem.

1 PENGENALAN

Pembikinan filem adalah industri yang berisiko tinggi. Selepas melabur sejumlah wang yang besar sehingga jutaan ringgit, syarikat filem hanya akan memperolehi pulangan pelaburannya dalam masa setahun atau dua tahun selepas filem tersebut ditayangkan di pawagam. Risiko pelaburan yang tinggi juga disebabkan kesukaran syarikat untuk mengetahui kehendak pasaran. Skor IMDB dalam beberapa tahun kebelakangan ini dipengaruhi oleh banyak faktor yang menjadikan ramalan skor yang tepat untuk filem baru yang dikeluarkan menjadi sukar. Namun faktor yang mempengaruhi sambutan terhadap sesebuah filem ini sukar dijangkakan sehingga filem tersebut telah selesai seratus-peratus. Faktor ini antara sebab risiko perniagaan filem semakin bertambah tinggi.

2 PENYATAAN MASALAH

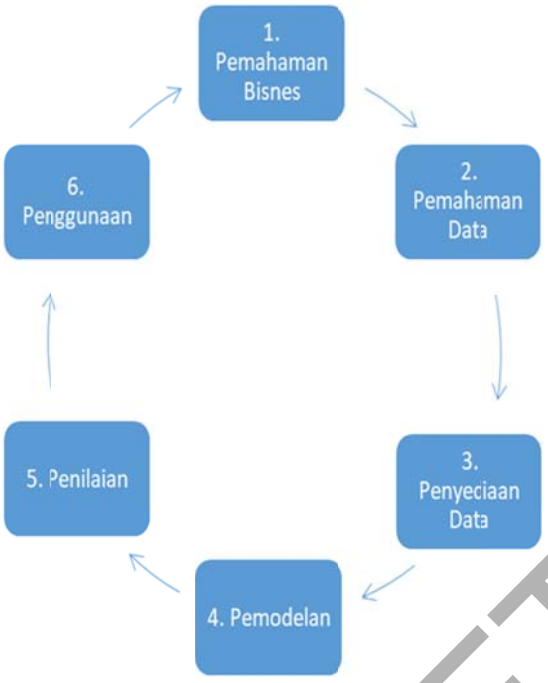
Pendapatan filem bergantung kepada pelbagai komponen seperti pelakon yang berlakon dalam filem, anggaran untuk pembuatan filem, ulasan kritikan filem, penarafan untuk filem, tahun pelepasan filem, dan lain-lain. Dengan begitu banyak perkara yang dipertaruhkan , ia adalah kepentingan komersial untuk membangunkan model yang boleh meramalkan kejayaan sesebuah filem dengan mengukur kejayaan filem menggunakan skor IMDb . Oleh itu, timbulnya masalah bagaimana untuk menganalisa faktor-faktor yang mempengaruhi kejayaan sesebuah filem dengan efektif agar pelbagai pihak berkepentingan seperti dalam industri filem boleh menggunakan ramalan ini untuk membuat keputusan yang lebih tepat tentang faktor-faktor yang menjadikan sesuatu filem itu lebih berjaya.

3 OBJEKTIF KAJIAN

Kajian yang dijalankan ini bertujuan untuk mencapai beberapa objektif iaitu , membangunkan model berdasarkan teknik perlombongan data yang boleh membantu meramalkan kejayaan filem dan seterusnya mengurangkan tahap ketidakpastian tertentu, mengenalpasti faktor-faktor yang mempengaruhi kejayaan sesebuah filem dan akhir sekali, melaksanakan teknik pengelasan.

4 METOD KAJIAN

Metodologi yang digunakan adalah CRISP-DM (*cross-industry process for data mining*) . CRISP-DM adalah metodologi perlombongan data yang komprehensif yang boleh digunakan sama ada orang yang baru terlibat dengan perlombongan data sehinggalah kepada pakar-pakar dengan satu pelan lengkap untuk menjalankan projek perlombongan data. CRSIP-DM bertujuan membuat projek perlombongan data yang lebih besar, jimat kos, lebih dipercayai, lebih terkawal dan lebih cepat. Metodologi kajian memainkan peranan penting dalam memastikan kajian ini berjalan dengan lancar dan teratur. Fasa kajian CRISP-DM dipaparkan dalam Rajah 1.



Rajah 1 Fasa Kajian CRISP-DM

4.1 FASA PEMAHAMAN BISNES

Fasa awal memberi tumpuan kepada memahami objektif dan keperluan projek dari perspektif perniagaan, dan kemudian menukar pengetahuan ini ke dalam definisi masalah data dan membuat perancangan awal bagi mencapai objektif. Pendapatan filem bergantung kepada pelbagai komponen seperti pelakon yang berlakon dalam filem, anggaran untuk pembuatan filem, ulasan kritikan filem, penarafan untuk filem, tahun pelepasan filem, dan lain-lain. Dengan begitu banyak perkara yang dipertaruhkan , ia adalah kepentingan komersial untuk membangunkan model yang boleh meramalkan kejayaan sesebuah filem dengan mengukur kejayaan filem menggunakan skor IMDb . Oleh itu, timbulnya masalah bagaimana untuk menganalisa faktor-faktor yang mempengaruhi kejayaan sesebuah filem dengan efektif agar pelbagai pihak berkepentingan seperti dalam industri filem boleh menggunakan ramalan ini untuk membuat keputusan yang lebih tepat tentang faktor-faktor yang menjadikan sesuatu filem itu lebih berjaya.

4.2 FASA PEMAHAMAN DATA

Fasa pemahaman data bermula dengan pengumpulan data dan seterusnya memahami data , untuk mengenal pasti masalah data,untuk mencari pandangan pertama ke dalam data , atau untuk mengesan subset menarik untuk membentuk hipotesis daripada maklumat yang

tersembunyi . Proses pengumpulan dan pemahaman data merupakan fasa kedua dalam metodologi pembangunan model yang digunakan dalam kajian ini.Kajian ini menggunakan data *TMDB Movie Dataset* yang dimuat turun dari Kaggle sebagai input. Terdapat 5043 data dalam set data yang dikumpul. Atribut dataset yang diperolehi di terangkan secara terperinci di dalam Jadual 1.

Jadual 1 Atribut Dataset

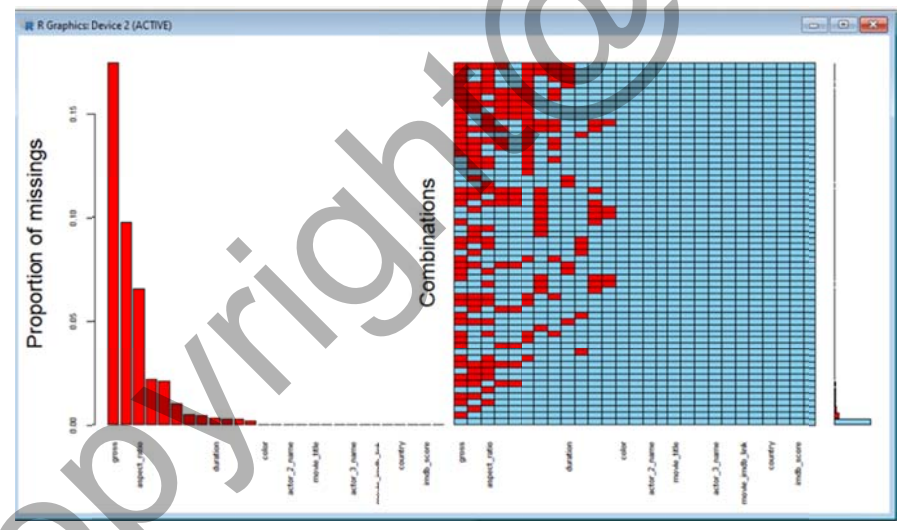
Atribut	**Penerangan**
movie_title	Tajuk filem
Duration	Durasi filem dalam minit
director_name	Nama pengarah filem
director_facebook_like	Bilangan 'likes' di laman Facebook pengarah
actor_1_name	Pelakon utama filem
actor_1_facebook_likes	Bilangan 'likes' di laman Facebook pelakon utama
actor_2_name	Pelakon sampingan dalam filem
actor_2_facebook_likes	Bilangan 'likes' di laman Facebook pelakon sampingan
actor_3_name	Pelakon sampingan dalam filem
actor_3_facebook_likes	Bilangan 'likes' di laman Facebook pelakon sampingan
num_user_for_reviews	Bilangan pengguna yang memberi ulasan
num_critic_for_reviews	Bilangan kritikan dalam ulasan
num_voted_users	Bilangan pengguna yang mengundi untuk filem
cast_total_facebook_likes	Bilangan keseluruhan 'likes' di laman Facebook kesemua pelakon filem
movie_facebook_likes	Bilangan likes di laman Facebook filem
plot_keywords	Kata kunci yang menceritakan plot keseluruhan filem
facenumber_in_poster	Bilangan pelakon yang terdapat di dalam poster tayangan filem
Color	Pewarnaan filem . Contoh : hitam/putih atau berwarna
Genre	Genre filem
title_year	Tahun filem di tayangkan
Language	Bahasa Contoh: English , Arabic , Chinese, French.
Country	Negara dimana filem dikeluarkan
content_rating	Rating sesebuah filem
aspect_ratio	Nisbah aspek fielm dibikin
movie_imdb_link	Pautan IMDB filem
Gross	Pendapatan kasar filem dalam Dollar
Budget	Bajet pembikinan filem dalam Dollar
Imdb_score	Skor IMDB filem di IMDB

4.3 FASA PENYEDIAAN DATA

Tujuan fasa ini dilakukan adalah untuk memastikan set data adalah bersih dan sedia di gunakan dalam fasa perlombongan data. Perisian RStudio telah digunakan dalam kajian ini

kerana mudah diterapkan pada set data. Jenis fail yang digunakan dalam kajian ini untuk menganalisis data dalam RStudio ialah jenis fail dalam format *Comma-Separated Values* (CSV).

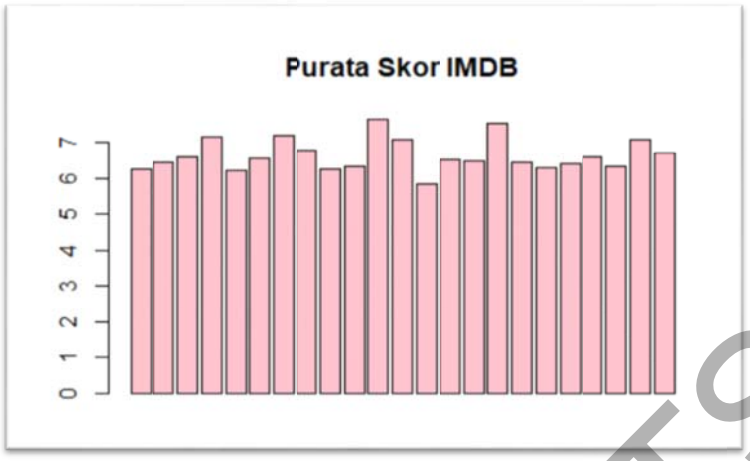
Proses pembersihan data adalah langkah awal dalam prapemprosesan dan bertujuan untuk menyingkirkan atau membetulkan data yang mempunyai ralat, tidak konsisten, hilang, rekod bertindih dan mengenalpasti outliers. Dalam dataset yang diperolehi , terdapat beberapa rekod yang bertindih dan perlu di singkirkan dan hanya rekod yang unik sahaja akan disimpan untuk proses perlombongan data. Jumlah rekod yang bertindih adalah 45 rekod dan 45 rekod tersebut juga telah disingkirkan. Selain itu , di dapati bahawa terdapat beberapa nilai yang hilang di dalam beberapa kolum dan fungsi colSums() digunakan untuk memaparkan rekod NA*(not available*) dalam setiap kolum. *Heatmap* telah digunakan untuk proses visualisasi rajah terhadap nilai-nilai hilang. Rajah 2 memaparkan *Heatmap* nilai hilang terhadap keseluruhan data.



Rajah 2 Heatmap Nilai Hilang

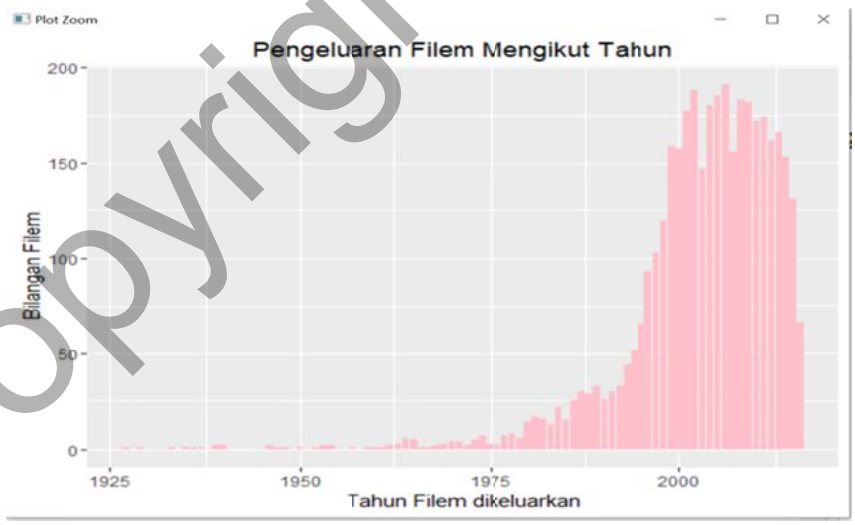
Seterusnya, pengeluaran data adalah proses untuk membuang atribut atau data daripada set data setelah mengkaji kepentingannya sama ada akan memberi impak atau tidak kepada hasil kajian. Pengeluaran data terbahagi kepada dua iaitu pengurangan jumlah data dan pengurangan jumlah atribut. Atribut genre telah disingkirkan datanya kerana data genre tidak berkaitan dengan skor sesebuah filem. Ini dapat dilihat dari rajah di bawah di mana purata

skor imdb bagi setiap genre adalah hampir sama dalam julat yang sama iaitu 6-8 seperi yang dapat dilihat pada Rajah 3.



Rajah 3 Carta Bar Purata Skor IMDB

Selain itu , atribut tahun pengeluaran filem juga telah di analisa kepentingannya dan divisualisasi melalui histogram dan didapati bahawa produksi filem mula meningkat dengan tinggi selepas tahun 1990. Melalui Rajah 4 **Error! Reference source not found.** juga dapat dilihat bahawa rekod pengeluaran filem sebelum tahun 1980 adalah sangat sedikit untuk di analisa dan data-data filem yang dikeluarkan sebelum tahun 1980 telah disingkirkan .



Rajah 4 **Error! Reference source not found.**

Setelah membuat kesemua proses pengeluaran data , atribut yang sebelum ini adalah 28 kini mempunyai 15 atribut yang penting dan 3711 data maklumat. Hal ini kerana attribut yang dibuang adalah yang telah dikaji tidak memberi kesan yang besar kepada keputusan kajian . Perlombongan data ke atas set data yang lebih kecil akan menjadikan proses pemodelan

berlaku dengan lebih lancar. Rajah 5 (A) memaparkan atribut sebelum pemilihan fitur dan **Error! Reference source not found.** 5(B) memaparkan senarai atribut selepas pemilihan fitur.

(A)

(B)

Rajah 5 Atribut Pemilihan Fitur

Selain itu , proses pendiskretan juga dilakukan untuk memastikan semua data yang berbentuk berterusan ditukarkan kepada nominal dan dibahagikan kepada skala tertentu . Pendiskritan data adalah bertujuan menyediakan data yang memudahkan proses perlombongan dilakukan dengan meringkaskan proses menganalisis data . Jadual 2 menunjukkan penukaran data terhadap attribut *imdb_score.*

Jadual 2 Pendiskretan Data

Atribut	*Imdb_score*			
Nama	'Flop'	'Average'	'Hit'	'Blockbuster'
Skala	0-4	4-6	6-8	8-10

4.4 FASA PEMODELAN

Proses perlombongan data dengan menggunakan teknik pengelasan akan dilakukan setelah prapemprosesan telah dilakukan . Dalam proses ini, tiga kaedah pengelasan akan digunakan iaitu Pohon Keputusan (*Decision Tree*), *Random Forest* dan juga KNN untuk membandingkan keputusan dan mencari model yang paling sesuai untuk meramal kejayaan sesebuah filem. Terdapat tiga peringkat dalam tugas pengelasan iaitu latihan, ujian dan pengesahan. Data ujian digunakan untuk menganggarkan ketepatan ramalan. Model yang dihasilkan digunakan untuk meramal kejayaan sesebuah filem berdasarkan skor IMDB dan menganalisa faktor-faktor kejayaan sesebuah filem. Perisian RStudio digunakan untuk membandingkan ketiga-tiga teknik pengelasan ini. Setelah mendapat model yang terbaik , model tersebut akan digunakan untuk memastikan fitur utama untuk model tersebut.

4.5 FASA PENILAIAN

Penalaan parameter dijalankan pada algoritma Pohon Keputusan untuk meningkatkan peratusan ketepatan pengelasan. Kaedah di jalankan adalah dengan trial and error dengan menala parameter yang terdapat pada Pohon Keputusan di dalam RStudio. Sebagai contoh , parameter CP adalah singkatan untuk *Complexity Parameter* untuk sesebuah pohon. Dalam algoritma pengelasan CP yang di ambil kira adalah CP yang mempunyai nilai ralat pengesahan yang terkecil iaitu xerror (*cross validation error*). Penalaan parameter juga di jalankan pada parameter *minsplit* yang berfungsi untuk menetapkan bilangan pemerhatian minimum di dalam *node* sebelum algoritma melakukan pembahagian. Analisis data bagi kajian ini juga menggunakan set data latihan , pengesahan dan ujian menggunakan *percentage split* 6:2:2 iaitu 60% set latihan, 20% set ujian dan 20% set pengesahan.

Jadual 3 Experimen 1

Percentage Split	**Parameter**	**Ketepatan Algoritma**

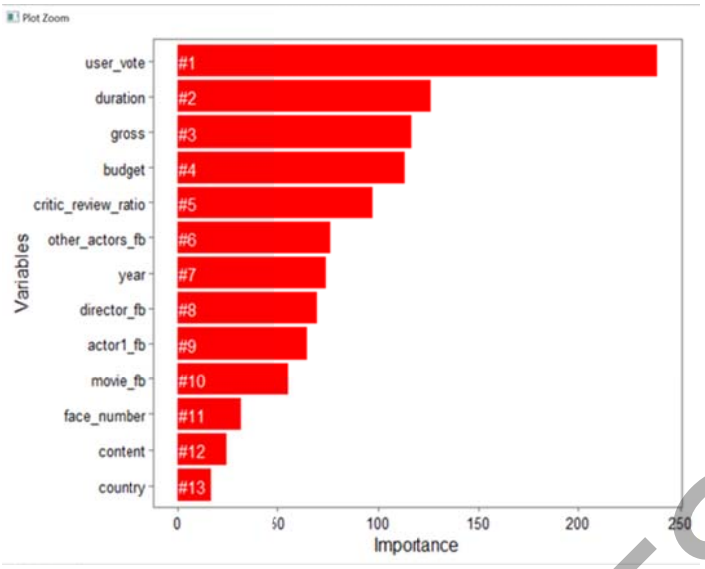
	minsplit	cp	xval	Set Latihan	Set Pengesahan	Set Ujian
6 : 2 : 2	5	0.0001	5	0.7812	0.7426	0.7022
8 : 1 : 1	5	0.0001	5	0.7844	0.69	0.7212

Jadual 4 Experimen 2

Percentage Split	Parameter			Ketepatan Algoritma		
	minsplit	cp	xval	Set Latihan	Set Pengesahan	Set Ujian
6 : 2 : 2	10	0.001	10	0.7812	0.7332	0.6887
8 : 1 : 1	10	0.001	10	0.7679	0.69	0.7256

K-Nearest Neighbour (KNN) adalah teknik klasifikasi yang termudah untuk digunakan apabila terdapat sedikit atau tiada pengetahuan awal mengenai data. Dalam kaedah ini setiap sampel harus di kelaskan sama dengan sampel di sekitarnya. Ini kerana sekiranya klasifikasi sampel tidak diketahui , maka ketepatan boleh diramalkan dengan mempertimbangkan pengelasan dengan mempertimbangkan nilai jiran(*neighbouring*) terdekat. Dalam kajian ini , 'K' akan ditetapkan sebagai 1 hingga 20 dan 20 model yang berbeza akan dibina. Ketepatan klasifikasi setiap model akan dikira dan 'K' yang terbaik adalah berdasarkan ketepatan yang tertinggi . 'K' yang mempunyai ketepatan yang tertinggi adalah apabila k=10 dengan ketepatan 0.72237.

Teknik klasifikasi *Random Forest* telah terbukti berjaya untuk beberapa aplikasi dan dianggap menghasilkan hasil prestasi yang baik serta kemudahan pengunaan dengan mempertimbangkan kepentingan atribut. Kepentingan pemboleh ubah relatif boleh dilihat dengan menggunakan ggplot2 di dalam perisian R untuk menunjukkan kepentingan relatif pemboleh ubah. Berdasarkan Rajah 6 atribut *user_vote* adalah yang paling penting manakala atribut *country* merupakan atribut yang paling rendah kepentinggannya.

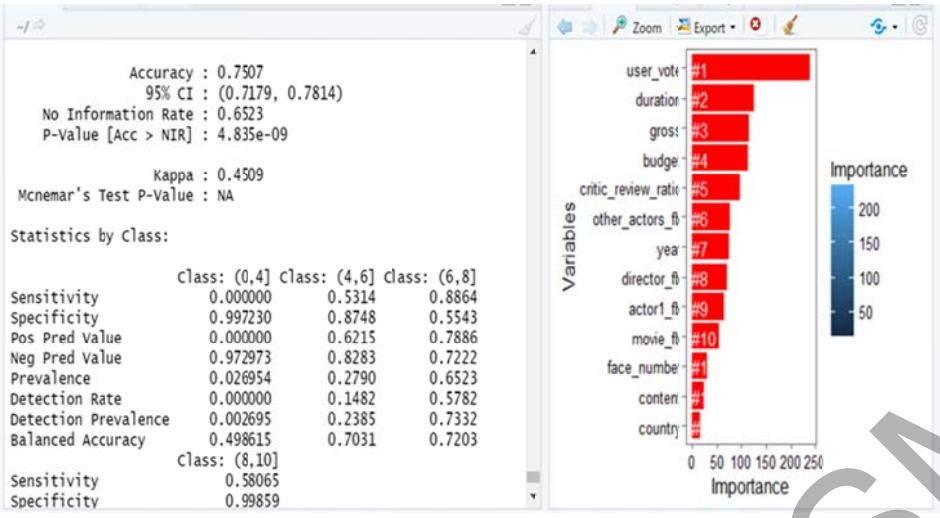


Rajah 6 Kepentingan Atribut

Ketepatan bagi set pengesahan adalah 0.7803 dan ketepatan bagi set ujian adalah 0.7507. Ketepatan bagi klasifikasi *Random Forest* ini merupakan yang tertinggi berbanding ketiga-tiga model. Parameter "mtry" dalam algoritma ini menentukan berapa banyak pemboleh ubah yang digunakan untuk model untuk membahagi pohon keputusan. Fungsi *caret* di dalam R akan memilih nilai parameter 'mtry' yang paling tepat dalam pengesahan bersilang.

Jadual 5 Ketepatan Random Forest

Nilai parameter *mtry*	**Ketepatan**	
	Set pengesahan	**Set Ujian**
5	0.7701	0.7412
10	0.7803	0.7507



Rajah 7 Ketepatan model RF terhadap set ujian

4.6 FASA PENGGUNAAN

Ini adalah fasa terakhir dalam CRISP-DM. Bergantung kepada keperluan . Fasa ini boleh menjadi seringkas membuat laporan atau menjadi kompleks seperti melakukan proses perlombongan data berulangkali.

5 HASIL KAJIAN

Hasil daripada dapatan kajian ini , dapat disimpulkan bahawa algoritma Random Forest memberi ketepatatan yang tertinggi sebanyak 75% berbanding model lain yang memberi ketepatan sebanyak 72%(Pohon Keputusan) dan 68%(KNN). Selain itu , dapat dilihat bahawa faktor-faktor yang banyak menyumbang dalam kejayaan sesebuah filem dan digunakan dalam setiap model adalah bilangan pengguna yang mengundi untuk filem di laman IMDB , durasi sesebuah filem , pendapatan kasar filem ,bajet pembikinan filem dan bilangan kritikan dalam ulasan sesebuah filem.

6 KESIMPULAN

Kesimpulannya , proses perlombongan data merupakan proses yang amat penting dalam kajian ini. Metrik prestasi yang terdapat dalam algoritma haruslah dikenalpasti untuk mendapat keputusan yang baik supaya dapat mencari model yang paling sesuai untuk meramalkan kejayaan sesebuah filem. Setiap fasa perlu dilaksanakan untuk memastikan

sebuah model yang dibina adalah tepat . Spesifikasi yang diperlukan adalah penting untuk membolehkan kajian ini dapat beroperasi dengan baik. Namun begitu terdapat beberapa cadangan masa hadapan untuk meningkatkan prestasi kajian ini iaitu seperti menggumpulkan dataset dari filem-filem tempatan dan mempelbagaikan sumber data, membuat lebih banyak perbandingan model dengan teknik yang berbeza dan membangunkan sistem untuk meramal skor IMDB berdasarkan model yang sedia ada. Sistem ini boleh di bangunkan dengan menggunakan perisan Microsoft Visual Studio.

7 RUJUKAN

Ahmad, J., Duraisamy, P., Yousef, A. & Buckles, B. 2017. Movie success prediction using data mining. *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017* 6(4): 198–203. doi:10.1109/ICCCNT.2017.8204173

Ahmed, M., Jahangir, M., Afzal, H., Majeed, A. & Siddiqi, I. 2015. Using crowd-source based features from social media and conventional features to predict the movies popularity. *Proceedings - 2015 IEEE International Conference on Smart City, SmartCity 2015, Held Jointly with 8th IEEE International Conference on Social Computing and Networking, SocialCom 2015, 5th IEEE International Conference on Sustainable Computing and Communic* (December): 273–278. doi:10.1109/SmartCity.2015.83

Cirp, P., Huber, S., Wiemer, H., Schneider, D. & Ihlenfeldt, S. 2018. ScienceDirect ScienceDirect DMME : Data mining methodology for engineering applications – a holistic extension the CRISP-DM model architecture of A new methodology to analyze the to functional and physical a holistic extension to the CRISP-DM model products for an assembly oriented product family identification DMME : Data mining methodology engineering applications –. *Procedia CIRP* 79: 403–408. doi:10.1016/j.procir.2019.02.106

Cook, C. & Tilcock, K. (n.d.). Predicting Blockbuster Success.

Imandoust, S. B. & Bolandraftar, M. 2017. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background (January 2013).

Kumar, S. 2019. Movie Success Prediction using Data Mining For Data Mining and Business Intelligence (ITA5007) Master of Computer Application (April).

Md.Dawam. 2007. emasaran Filem di Malaysia: Konsep Produk, Segmentation dan Positioning dalam Filem. *Pemasaran Filem di Malaysia: Konsep Produk, Segmentation dan Positioning dalam Filem*.

Quader, N, Gani, M. O. & Chaki, D. 2017. Performance evaluation of seven machine learning classification techniques for movie box office success prediction. *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, hlm. 1–6. doi:10.1109/EICT.2017.8275242

Quader, Nahid, Gani, M. O., Chaki, D. & Ali, M. H. 2018. A machine learning approach to predict movie box-office success. *20th International Conference of Computer and Information Technology, ICCIT 2017* 2018-Janua: 1–7. doi:10.1109/ICCITECHN.2017.8281839

Wirth, R. 2000. CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (24959): 29–39. doi:10.1.1.198.5133