PENENTUAN MODEL PEMBELAJARAN MESIN DALAM PENGELASAN RATING FILEM DAN RANCANGAN TELEVISYEN DI PLATFORM PENSTRIMAN NETFLIX

Nur Izyani Ahmad, Siti Aishah Hanawi, Ruzzakiah Jenal, Mohd Syazwan Baharuddin

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM

Bangi, Selangor Darul Ehsan

Abstrak

Dewasa ini, industri hiburan memainkan peranan penting dalam pembangunan ekonomi. Kemajuan pesat industri mengakibatkan revolusi dalam industri hiburan di mana platform penstriman dalam talian seperti Netflix, Amazon TV, Iflix dan sebagainya menjadi platform utama pilihan penonton. Pertambahan platform penstriman dalam talian atau juga dikenali sebagai Over The Top (OTT) memberi persaingan sengit di antara platform ini bagi memastikan kandungan yang dikeluarkan menepati cita rasa pengguna. Dalam konteks hiburan, prestasi sesebuah filem atau rancangan televisyen (TV) adalah bergantung kepada rating yang diberi oleh penonton. Pelbagai faktor yang akan menyumbang kepada prestasi ini. Terdapat beberapa kajian literasi yang menggariskan pelbagai faktor seperti genre, musim, pengarah, durasi dan sebagainya mempengaruhi rating IMDb.Namun, kepelbagaian serta keluasan atribut perlu sentiasa disemak dengan mengetengahkan kaedah kecerdasan buatan iaitu pembelajaran mesin dalam meramal dan mengkelaskan rating rancangan TV dan filem. Oleh itu, kajian bertujuan memberikan cadangan model rating rancangan TV dan filem yang ditayangkan oleh platform penstriman dalam talian dengan menggunakan algoritma pembelajaran mesin. Beberapa model pembelajaran mesin seperti Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), K-nearest Neighbours (KNN), Gradient Boosting (GB) dan Ada Boosting (AB) diguna dalam kajian . Hasil kajian menunjukkan model pembelajaran mesin yang mendapat ketepatan tinggi dalam membuat ramalan rating skor IMDb ialah model GB dengan memperoleh 89% ketepatan.

1 PENGENALAN

Dewasa ini, industri hiburan memainkan peranan penting dalam pembangunan ekonomi. Perkembangan pesat industri ini membawa kepada revolusi dalam industri hiburan di mana platform penstriman dalam talian seperti Netflix, Amazon TV, Iflix dan sebagainya menjadi platform utama pilihan penonton. Dalam konsep hiburan, prestasi sesebuah filem atau rancangan televisyen (TV) adalah bergantung kepada rating yang diberi oleh penonton. Terdapat pelbagai cara digunakan untuk mengukur kualiti kandungan sesebuah filem atau rancangan TV dan salah satu metrik yang terkenal dan sering digunakan secara meluas ialah rating IMDb (Pangkalan Data Filem Internet).

IMDb merupakan laman sesawang pangkalan data filem dan rancangan TV terkemuka dan terkenal seluruh dunia. Rating IMDb memberi impak besar bagi industri hiburan kerana rating ini mencerminkan pendapat pengguna bagi mengukur populariti sesuatu filem dan rancangan TV. Rating mempengaruhi keputusan filem atau rancangan TV yang ditonton serta menjadi penanda aras bagi penyedia platform dalam talian untuk menilai kualiti dan kejayaan sesuatu kandungan. Proses perlombongan data hiburan melibatkan teknik perlombongan data boleh digunakan untuk menganalisis dan menemui kumpulan atau pakej baharu atau saluran berbeza yang dipilih oleh pelanggan (Sharma et al. 2019).

Sektor hiburan merupakan sektor yang kompetitif dan dinamik. Cabaran utama bagi penyedia platform penstriman dalam talian adalah untuk memastikan kandungan yang disediakan memenuhi kehendak penonton. Rating IMDb yang berasaskan ulasan dan undian pengguna adalah sangat subjektif dan sering dipengaruhi oleh pelbagai faktor seperti cita rasa individu, latar belakang budaya dan juga pengalaman peribadi. Hal ini menyukarkan lagi proses ramalan dan membawa kepada perkembangan teknik perlombongan data dalam industri hiburan.

Perlombongan data adalah satu teknik mengekstrak maklumat berguna daripada data yang berskala besar. Perlombongan data bermaksud menganalisis data daripada perspektif yang berbeza dan meringkaskannya kepada maklumat berguna yang boleh digunakan untuk membuat keputusan kritikal. Antara contoh teknik di dalam perlombongan data adalah meneroka, menganalisis dan mengesan corak dalam jumlah data yang besar. Matlamat utama bagi perlombongan data adalah untuk membuat klasifikasi data atau ramalan data.

Pembelajaran mesin merupakan salah satu cabang kepada teknologi kecerdasan buatan. Pembelajaran mesin merupakan disiplin kecerdasan buatan yang menggunakan data untuk melatih mesin membuat keputusan tertentu (Azahari 2022). Pembelajaran mesin berupaya untuk meneroka, mentafsir, mempelajari struktur data dan membina algoritma yang boleh meramal keputusan dan membuat keputusan di luar kebolehan manusia. Di samping itu, teknik ini juga bersifat adaptif dan fleksibel kerana ia mampu untuk meningkatkan keputusan ramalan seandainya terdapat penambahbaikan dalam algoritma model yang dipelajari.

Gabungan teknik perlombongan data dan pembelajaran mesin dapat membantu penyedia platform penstriman dalam talian bagi menyediakan kandungan yang memenuhi kehendak penonton. Teknik perlombongan data dapat memberi manfaat dalam usaha mengekstrak, memproses dan memfokuskan faktor-faktor yang mempengaruhi rating sebuah rancangan TV atau filem manakala teknik pembelajaran mesin pula membantu untuk meramal sesuatu rancangan TV dan filem pilihan penonton berdasarkan data perlombongan.

Meskipun kebanyakan kajian ilmiah menggunakan teknik perlombongan data yang hampir sama, namun, kepelbagaian serta keluasan atribut menyebabkan model-model terdahulu perlu sentiasa disemak dan ditambah baik dari masa ke semasa. Keperluan pengguna yang semakin berubah membawa kepada penggunaan perkembangan teknik sains data oleh pencipta kandungan untuk menjana amalan baharu dan mencipta peluang untuk menghasilkan kandungan berkualiti lebih tinggi untuk memenuhi permintaan pengguna (Anmadwar et al. 2023). Oleh itu, kajian dijalankan dengan mengetengahkan kaedah kecerdasan buatan iaitu pembelajaran mesin dalam meramal dan mengkelaskan rating rancangan TV dan filem. Hal ini seterusnya dapat membantu penyedia platform untuk mendapatkan trend dan cita rasa terkini penonton bagi merancang strategi produk yang disajikan serta meningkatkan mutu servis.

Objektif kajian adalah seperti berikut:

1. Menganalisis trend dan statistik prestasi rancangan TV dan filem yang ditayangkan secara deskriptif.

2. Mengenal pasti prestasi rancangan TV dan filem yang ditayangkan menggunakan pembelajaran mesin.

3. Menentukan model pembelajaran mesin terbaik dalam membuat ramalan prestasi dan rancangan TV dan filem yang ditayangkan.

Kajian bertujuan memberikan cadangan model rating rancangan TV dan filem yang ditayangkan oleh platform penstriman dalam talian. Mengambil kira ketersediaan data yang luas serta popularity platform dalam kalangan penonton, Netflix dipilih sebagai platform dan rating IMDb merupakan metrik prestasi yang digunakan bagi kajian. Data dari portal IMDb dan platfom Netflix digabungkan menjadi satu set data untuk kajian . Data ini diperoleh daripada Eduardo Gonzalez yang membantu menggabungkan kedua-dua set data dan memuat naik di laman sumber terbuka Kaggle.com (Gonzalez 2022). Data mentah dari Kaggle.com melalui proses pemprosesan data dan hanya data yang memenuhi kriteria iaitu mempunyai sekurang-kurangnya 10000 undian dipilih untuk dijadikan set data. Kajian merangkumi lima fasa iaitu fasa tinjauan kajian, fasa pemprosesan data, fasa analisis deskriptif, fasa pemodelan dan yang terakhir fasa pengujian untuk mengenal pasti model terbaik untuk kajian.

2 KAJIAN LITERASI

a) Analisis Rating

Menurut Kamus Dewan Bahasa dan Pustaka Edisi Empat, analisis bermaksud penyelidikan atau penghuraian sesuatu (seperti keadaan, masalah, persoalan dan lain-lain) untuk mengetahui pelbagai aspek secara terperinci atau mendalam. Rating pula didefinisi sebagai berapa kadar kepopularan sesuatu rancangan yang diukur berdasarkan jumlah pendengar, penonton dan sebagainya. Justeru, analisis rating bermaksud penyelidikan atau penghuraian kadar kepopularan sesuatu rancangan yang diukur berdasarkan jumlah maklum balas penonton mahupun pendengar.

Kemajuan dan kepesatan teknologi membolehkan bidang perfileman dan sinematografi melakukan pelbagai penyelidikan dalam mencapai objektif memperoleh rating yang tinggi terhadap sesuatu karya. Pengaplikasian data analisis dalam industri menjadi satu kemestian untuk penggiat seni meramal kejayaan mutu sesuatu karya. Seiring dengan perkembangan teknologi, hasil kejayaan sesebuah rancangan TV atau filem bergantung pada regresi IMDb rating mahupun pengklasifikasian kejayaan atau kegagalan berdasarkan penggunaan teknik klasifikasi ramalan.

b) Faktor Mempengaruhi Prestasi Rating Filem dan Rancangan TV

Pelbagai faktor yang mempengaruhi kejayaan sesebuah rancangan TV atau filem. Faktor seperti genre, durasi, produksi, kategori rancangan dan sebagainya sering digunakan dalam kajian melibatkan kejayaan sesebuah rancangan TV dan filem. Kajian yang dijalankan oleh Dixit et al. (2020) menggariskan durasi, genre, dan bajet juga antara faktor yang dipercayai mempengaruhi penonton untuk menonton sesebuah filem. Selain itu, kajian oleh Gaenssle et al. (2018) menggariskan tiga faktor di sebalik kejayaan sesebuah filem iaitu bajet, jenama individu seperti pelakon dan pengarah dan faktor terakhir ulasan penonton. Hasil kajian mendapati faktor bajet memberi kesan positif terhadap kejayaan sesebuah filem. Bristi et al. (2019) menjalankan kajian membandingkan prestasi algoritma pembelajaran mesin dalam membuat ramalan. Faktor bajet, produksi, pengarah, genre, negara ditayang dan tahun dikeluar digunakan dalam menentukan kejayaan filem. Hasil kajian mendapati bajet mempunyai pengaruh kecil namun faktor pelakon tidak mempengaruhi kejayaan sesebuah filem. Kajian oleh Mhowwala et al. (2020) di mana kajian tersebut membuktikan kejayaan rating sesebuah filem sering dikaitkan dengan populariti pengarah dalam industri hiburan. Gupta et al. (2022) mencadangkan satu kajian untuk meramal kadar kejayaan atau kegagalan sesebuah filem bagi meningkatkan pertumbuhan industri perfileman. Enam atribut utama digunakan dalam kajian iaitu bulan keluaran, rating IMDb, tempoh, genre, bajet, hit atau flop. Lall dan Sivakumar (2021) membuat kajian dengan menggunakan atribut musim bagi mengukur sama ada penonton akan meneruskan menonton sesuatu rancangan TV atau meninggalkan rancangan TV tersebut pada musim seterusnya. Selain atribut musim, atribut durasi, tahun ditayangkan, genre dan kategori umur turut digunakan dalam kajian tersebut. Jadual 2.1 memaparkan rumusan faktor yang mempengaruhi prestasi filem atau rancangan TV dalam kajian lepas.

Jadual 2.1 Rumusan faktor yang mempengaruhi prestasi filem atau rancangan TV dalam kajian lepas

Nama Penulis	**Faktor Kajian**	**Hasil Kajian**
Dixit et al. (2020)	Undian, durasi, genre, dan bajet.	Bersetuju faktor kajian mempengaruhi prestasi rancangan TV atau filem.
Gaenssle et al. (2018)	Bajet, pelakon, pengarah dan ulasan penonton.	Bersetuju faktor kajian mempengaruhi prestasi rancangan TV atau filem.

Bristi et al. (2019)	Bajet, produksi, pengarah, genre, negara ditayangkan dan tahun dikeluar.	Bersetuju faktor kajian mempengaruhi prestasi rancangan TV atau filem.
Mhowwala et al. (2020)	Genre, durasi, kategori umur, rating, negara produksi, undian, pengarah, pelakon, penulis dan sosial media.	Bersetuju faktor kajian mempengaruhi prestasi rancangan TV atau filem.
Gupta et al. (2022)	Bulan keluaran, skor imdb, tempoh, genre, bajet, *hit* atau *flop.*	Bersetuju faktor kajian mempengaruhi prestasi rancangan TV atau filem.
Lall dan Sivakumar (2021)	Musim, durasi, tahun ditayangkan, genre dan kategori umur.	Bersetuju faktor kajian mempengaruhi sama ada penonton akan meneruskan tontonan sesuatu rancangan TV atau filem.

c) Pembelajaran Mesin

Pembelajaran mesin merupakan satu cabang dalam bidang sains data yang semakin berkembang pesat. Pembelajaran mesin digunakan untuk mengekstrak informasi daripada data di mana informasi tersebut boleh dinilai dalam bentuk yang boleh difahami dan digunakan untuk pelbagai tujuan (Baradwaj 2012). Teknik pembelajaran mesin yang sering digunakan dalam perlombongan data adalah teknik klasifikasi. Teknik klasifikasi adalah satu kaedah yang digunakan untuk meramal kelas sasaran untuk sesuatu set data. Teknik klasifikasi melibatkan fasa latihan dan fasa ujian. Di dalam fasa latihan, model dibina menggunakan set latihan di mana model tersebut mempelajari serta mengenal pasti atribut dan kelas label. Setelah model latihan dijana, model tersebut digunakan pada set data ujian untuk mengenal pasti kelas label bagi data ujian. Terdapat beberapa teknik pembelajaran mesin klasifikasi seperti *Decision Tree (DT), Naive Bayes (NB), K-nearest Neighbours (KNN), Random Forest (RF), Gradient Boosting (GB)* dan *Ada Boosting (AB).*

Pengelas *Naive Bayes (NB)* ialah pengelas berkebarangkalian berdasarkan teorem Bayes; di mana kebergantungan nilai kepada sesuatu kebarangkalian adalah berdasarkan peristiwa (A) dan bergantung dengan peristiwa yang lain (B). Pengelas NB menganggap ciri yang masuk ke dalam model adalah bebas antara satu sama lain. Iaitu mengubah nilai satu ciri, secara tidak langsung mempengaruhi atau mengubah nilai mana-mana ciri lain yang digunakan dalam algoritma (Gandhi 2018). Bristi et al. (2019) menjalankan kajian untuk meramal rating IMDb sesebuah filem menggunakan beberapa model termasuk *Naive Bayes.* Kajian bertujuan membandingkan prestasi algoritma pembelajaran mesin dalam membuat ramalan. Sebanyak 242 set data filem diekstrak daripada laman sesawang Wikipedia dan IMDb. Atribut rating diukur

dengan mengekstrak data daripada laman sesawang IMDb. Atribut diklasifikasikan kepada empat iaitu gagal, di bawah purata, purata, dan hit. Keputusan NB yang tinggi diperoleh apabila menggunakan algoritma klasifikasi dengan melakukan sampel semula set data dengan penggantian.

Pengelas *Decision Tree (DT)* menjadi salah satu teknik yang dapat membantu untuk membuat keputusan efektif. Teknik ini memiliki struktur sepohon pokok yang lengkap dengan akar, daun dan ranting. Nod pokok mewakili atribut manakala nod tepi/cabang mewakili nilai dari atribut, dan daun mewakili kelas. Kajian yang dilaksanakan oleh Sadashiv et al. (2021) membuktikan model DT memperoleh ketepatan yang tinggi iaitu 81% dibandingkan NB dengan hanya 72%. Kajian yang dijalankan oleh Bristi et al. (2019) juga membuktikan DT turut memperoleh ketepatan yang lebih tinggi apabila membuat perbandingan ketepatan tanpa melakukan sampel semula set data tanpa penggantian.

Pengelas *Random Forest (RF)* ialah teknik yang menggunakan pembelajaran gabungan, ia juga menggabungkan banyak pengelas lemah untuk menyediakan penyelesaian kepada masalah yang kompleks. RF terdiri daripada kombinasi beberapa DT, daripada ia bergantung pada satu pokok, ia mengambil ramalan dari setiap pokok dan berdasarkan undian majoriti ramalan, meramalkan output akhir (Saini 2022). Satu kajian dijalankan oleh Sivakumar dan Ekanayake (2021) untuk membuat ramalan kejayaan sesebuah filem dengan menggunakan ulasan treler di Youtube. Teknik *Natural Language Processing (NLP)* digunakan untuk mengekstrak kata kunci daripada ulasan pengguna. *Tokenizing*, *Stemming*, dan pengkategorian ulasan kepada positif atau negatif dilakukan berdasarkan analisis sentimental. Algoritma RF dipilih menggunakan ciri yang diekstrak daripada IMDb untuk meramalkan kejayaan sesebuah filem manakala NB menggunakan ulasan pengguna yang diekstrak daripada Youtube untuk meramal rating. Dua kesimpulan dicapai bahawa rating filem baru tidak boleh diramalkan terlebih dahulu melalui ulasan treler pada komen Youtube namun kejayaan filem baru boleh diramal lebih awal dengan menggunakan data atau ciri yang dikumpul daripada dalam talian. Dixit et al. (2020) turut mengaplikasikan model RF dalam pembangunan model klasifikasi untuk meramal prestasi filem dalam kajian beliau. Hasil kajian Sadashiv et al. (2021) turut menunjukkan model RF memperoleh ketepatan paling tinggi iaitu 85.2% dalam membuat ramalan awal kejayaan filem sebelum ditayangkan. Lall dan Sivakumar

(2021) turut menggunakan model NB dan RF dalam kajian untuk melihat sama ada penonton akan meneruskan tontonan atau meninggalkan tontonan bagi rancangan TV bermusim.

Pengelas *K-nearest Neighbours (KNN)* tergolong dalam domain pembelajaran model yang diselia. Model KNN mengklasifikasikan titik data baharu berdasarkan "jarak" kepada data yang serupa atau diketahui. Dalam kehidupan seharian, algoritma KNN sering digunakan dalam sistem pengesyoran atau dalam teknologi pengecaman (Schlee 2020). Satu kajian untuk meramal rating bagi filem belum ditayang dilaksanakan oleh Priyanganie (2021). Beliau memilih model KNN kerana model tersebut bersifat fleksibel dan teguh kepada noisy data. Di samping itu, model KNN juga menjadi pilihan Bristi et al. (2019) dalam kajian terhadap rating IMDb.

Boosting ialah satu teknik pembelajaran mesin untuk melatih koleksi algoritma pembelajaran mesin agar berfungsi lebih baik untuk meningkatkan ketepatan, mengurangkan berat sebelah dan mengurangkan varian. *Boosting* berfungsi dengan melatih model lemah secara berulang pada subset data latihan yang berbeza; model seterusnya direka bentuk untuk lebih baik daripada model sebelumnya yang kurang ketepatan klasifikasi (Lawton 2023). Namun antara yang terkenal dan sering digunakan ialah *Ada Boosting (AB)* dan *Gradient Boosting (GB).* AB ialah teknik pengukuhan adaptif di mana pemberat data dilaraskan berdasarkan kejayaan setiap algoritma (model lemah) dan diserahkan kepada model seterusnya untuk dibetulkan. GB pula merupakan satu teknik terkenal yang direka secara dinamik dan cepat sebagai tindak balas kepada pengesanan ralat dalam algoritma sebelumnya. Gupta et al. (2022) membuat satu kajian menggunakan model KNN, *Support Vector Machine (SVM)*, GB, *AB dan eXtreme Gradient Boosting (XB).* Hasil kajian ini juga tetap menunjukkan Boosting iaitu GB memperoleh ketepatan yang tinggi manakala KNN memperoleh ketepatan yang terendah. Jadual 2.2 merumuskan teknik pembelajaran mesin dalam kajian lepas.

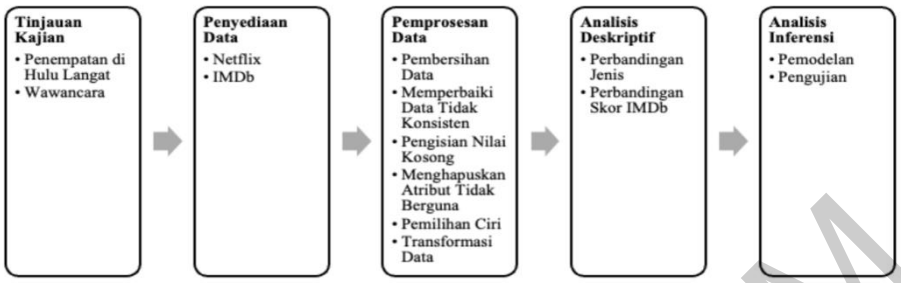
Jadual 2.2 Rumusan teknik pembelajaran mesin dalam kajian lepas

Nama Penulis	**Teknik**	**Hasil Kajian**
Briti et al. (2019)	NB, DT, KNN, *Bagging*, RF	RF model terbaik dengan ketepatan paling tinggi.
Sadashiv et al. (2021)	NB, DT, KNN, SVM, *Logistic Regression*(LR), RF	Model RF memperoleh ketepatan paling tinggi.

Mhowwala et al. (2020)	RF dan XB	Model XB model terbaik dengan ketepatan paling tinggi.
Dixit et al. (2020)	Regresi :*Simple Linear Regression*, SVM, RF, XB Klasifikasi : LR, SVM, GB, RF	Model Regresi : Model XB adalah model terbaik dengan ketepatan paling tinggi. Model Klasifikiasi : Model GB model terbaik dengan ketepatan paling tinggi.
Sivakumar dan Ekanayake (2021)	NLP, RF, NB	Rating filem baru tidak boleh diramalkan terlebih dahulu melalui ulasan treler pada komen *YouTube* namun kejayaan filem baru boleh diramal lebih awal dengan menggunakan data atau ciri yang dikumpul daripada dalam talian.
Gupta et al. (2022)	KNN, SVM, GB, AB, XB	Model GB model terbaik dengan ketepatan paling tinggi.
Priyanganie (2018)	LR, KNN, DT, RF, NB, *Bagging*, *Boosting*	Model DT dengan peningkatan *bagging* model terbaik dengan ketepatan paling tinggi.
Lall dan Sivakumar (2021)	LR, NB, SVM , RF	Model RF memperoleh ketepatan paling tinggi.

3 METODOLOGI KAJIAN

Metodologi kajian terbahagi kepada lima fasa. Fasa pertama ialah fasa tinjauan di mana fasa ini mendapatkan statistik penggunaan platform penstriman dalam talian di sebuah penempatan di Hulu Langat. Fasa kedua adalah fasa penyediaan set data di mana sumber data berada dikenal pasti, diperiksa dan dianalisis untuk mendapatkan gambaran keseluruhan butiran dan kuantiti data. Fasa ketiga adalah fasa pemprosesan data yang terdiri daripada pembersihan data, pemilihan data, dan transformasi data yang membantu mengubah data menjadi format yang boleh difahami untuk data pemodelan. Fasa keempat adalah analisis deskriptif atau ekplorasi data. Fasa terakhir adalah analisis inferensi teknik-teknik pembelajaran mesin untuk pemodelan rating IMDb. Dua aplikasi utama digunakan untuk kajian iaitu Microsoft Excel untuk analisis awal dan integrasi data dan Google Colab untuk pemprosesan data, ciri pemilihan serta kajian perbandingan untuk setiap pembelajaran mesin pengelasan.



Rajah 3.1 Metodologi kajian

a) Tinjauan Kajian

Pemilihan platform diperkukuhkan dengan hasil tinjauan secara rawak yang dibuat di sebuah penempatan di Hulu Langat. Hasil tinjauan mendapati 94.1% pengguna membuat langganan platform dan 100% pengguna melanggani *Over The Top (OTT)* platform. Pemilihan rating IMDb pula dibuat setelah menganalisa kredibiliti portal yang bertapak sejak 1990 dan juga hasil dari data tinjauan mendapati 78.5% pengguna mengetahui tentang rating IMDb. Pemilihan kajian juga diperkukuhkan lagi dengan hasil wawancara dalam talian bersama salah sorang pengarah pemasaran syarikat penstriman dalam talian milik Hong Kong di mana beliau percaya bahawa kajian memberi faedah yang besar kepada industri OTT.

b) Penyediaan Data

Data diperoleh daripada Eduardo Gonzalez yang membantu menggabungkan kedua-dua set data dan memuat naik di laman sumber terbuka Kaggle.com. Set data yang diguna untuk kajian mengandungi maklumat filem dan rancangan TV di platform Netflix sehingga Mei 2022. Eduardo Gonzalez memuat dua fail data mentah iaitu *raw_titles.csv* dan *raw_credits.csv* yang digabungkan untuk dijadikan set data baru untuk kajian.

c) Pra-Pemprosesan Data

Proses pembersihan data melibatkan aktiviti pengurangan data, mengisi nilai kosong, dan membuang atribut yang tidak diperlukan untuk fasa permodelan. Proses ini juga melibatkan aktiviti menyelenggara data yang tidak konsisten. Proses pembersihan yang dilakukan dalam kajian ialah pemilihan rekod data yang relevan dengan IMDb. IMDb menggariskan sesebuah filem atau rancangan TV perlu memperoleh sekurang-kurangnya 10000 undian untuk disenaraikan dalam laman sesawang IMDb. Hasil daripada analisis dan proses pembersihan pertama, sebanyak 302 rekod dan 14 atribut digunakan bagi kajian . Di samping proses pembersihan data, proses memperbaiki data tidak konsisten turut dititikberat dalam pemprosesan data. Atribut genre dan negara produksi mempunyai rekod yang tidak konsisten di mana setiap baris mempunyai bilangan data yang tidak sama. Bagi menghasilkan data yang lebih konsisten, fungsi *Text to Column* digunakan dalam aplikasi Microsoft Excel untuk membahagikan kumpulan rekod dalam setiap baris kepada individu rekod. Sebagai contoh, sekiranya terdapat dua rekod genre dalam satu jalur, ia akan dibahagi kepada dua jalur yang berbeza. Hanya genre pertama yang dipilih dan genre yang lain dibuang. Setelah mengenal pasti rekod data yang digunakan, kajian diteruskan dengan menganalisis data daripada setiap lajur. Tujuan analisis dibuat bagi memastikan bilangan rekod bagi setiap atribut dan sekiranya terdapat sebarang data yang hilang yang memerlukan proses pembersihan data. Pengurusan data hilang adalah penting kerana setiap data mempunyai kepentingan dalam bidang sains data. Pengisian nilai kosong dibuat dengan melengkapkan atribut yang mempunyai data hilang iaitu kategori umur, musim dan pengarah. Nilai kategori umur “R” diisi pada rekod data hilang kerana nilai ‘R’ merupakan mod bagi atribut kategori umur. Nilai “R” direkodkan sebanyak 128 kali. Bagi atribut musim pula, nilai “1” diisi bagi data yang hilang. Hal ini kerana setiap karya perlu mempunyai sekurang-kurangnya satu musim untuk filem atau rancangan TV. Atribut pengarah pula mempunyai 228 data yang hilang. Frasa “Tidak Ditakrif” digunakan bagi mengisi data yang hilang kerana informasi yang terbatas bagi setiap filem atau rancangan TV. Seterusnya ialah proses menghapuskan atribut tidak berguna. Proses ini dilakukan kepada atribut yang tidak menyumbang kepada algoritma dalam model ramalan. Menghapuskan atribut tidak berguna juga dapat membantu mempercepatkan masa pemprosesan komputer serta menjimatkan simpanan ketika pemodelan dilaksanakan. Dalam kajian, tiga atribut iaitu Indeks, ID dan ID IMDb dihapus kerana atribut tersebut tidak memberi sebarang sumbangan mahupun

implikasi kepada kajian. Jadual 3.1 memaparkan senarai atribut setelah proses pra-pemprosesan data.

Jadual 3.1 Set data selepas proses pra-pemprosesan data

Nama Atribut	Jenis Data	Deskripsi
Tajuk	Nominal	Tajuk filem atau rancangan TV
Jenis	Nominal	Jenis filem atau racangan TV
Tahun Ditayangkan	Integer	Tahun tayangan
Kategori Umur	Nominal	Kelayakan umur tontonan untuk filem atau rancangan TV
Durasi	Integer	Tempoh tayangan
Genre	Nominal	Genre filem atau rancangan TV
Negara Produksi	Nominal	Kod negara pengeluar filem atau rancangan TV
Musim	Integer	Bilangan musim
Skor IMDb	*Float*	Skor IMDb bagi filem atau rancangan TV
Undian IMDb	Integer	Bilangan undian IMDb bagi filem atau rancangan TV
Pengarah	Nominal	Nama pengarah

d) *Correlation Coefficient*

Correlation Coefficient ialah penilaian kolerasi antara atribut. Kolerasi membantu dalam proses meramal antara satu atribut kepada atribut yang lain. *Correlation Coefficient* yang sering digunakan ialah *Pearson Correlation* (Gupta 2023). Jadual 3.2 memaparkan kedudukan atribut hasil daripada pemilihan ciri.

Jadual 3.2 Kedudukan atribut berdasarkan teknik *Correlation Coefficient*

Nama Atribut	Skor
Skor IMDbJenis	1.000
Kategori Umur	0.590
Undian IMDb	0.510
Musim	0.382
Pengarah	0.368
Tahun ditayangkan	0.180
Tajuk	0.020
Negara Produksi	0.001
Genre	-0.080
Durasi	-0.111
	-0.322

Korelasi yang bernilai positif iaitu lebih besar daripada 0 menandakan atribut dan atribut kelas bergerak seiring ke arah yang sama. Sekiranya atribut meningkat atribut kelas juga meningkat. Korelasi negatif berlaku apabila korelasi kurang daripada 0 di mana atribut dan atribut kelas bergerak ke arah yang bertentangan. Sekiranya atribut meningkat atribut kelas akan berkurang.

e) Transformasi Data

Transformasi data adalah proses di mana data diubah atau disatukan supaya proses perlombongan data menjadi lebih cekap, dan corak hasil lebih mudah difahami. Strategi untuk transformasi data termasuk pelicinan, pembinaan atribut, pengagregatan, penormalan, pendiskretan dan penjanaan hierarki konsep untuk data nominal (Han et al. 2012). Penjanaan kelas label dilakukan dengan menggunakan data skor IMDb kepada rating. Tiga kelas label dijana berpandukan skor IMDb yang digariskan di portal IMDb. Jadual 3.3 menunjukkan pembinaan kelas label.

Jadual 3.3 Penjanaan kelas label

Kelas Label	**Skor IMDb (S)**
Rendah	1<= S < 5
Pertengahan	5 <= S < 8
Tinggi	8 <= S <=10

Kewujudan atribut kelas membawa kepada proses pembelajaran mesin berselia. Proses ini membenarkan model pembelajaran mesin mengklasifikasikan setiap data kepada kelas label tertentu. Proses transformasi berlaku ketika fasa pembangunan model.

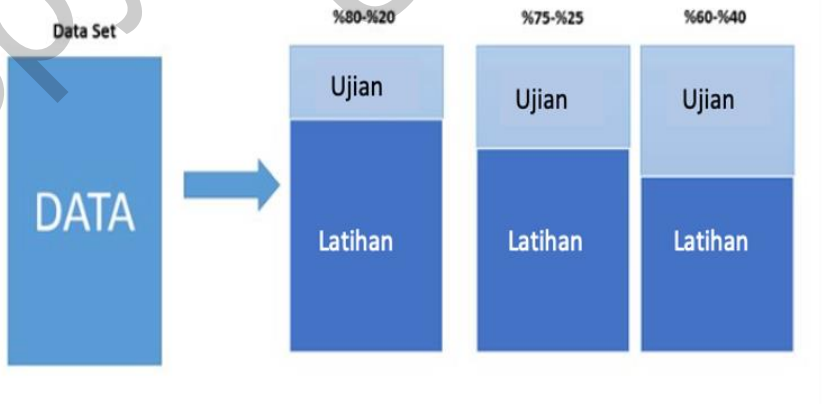
f) Analisis Deskriptif Data Netflix-Imdb

Set data yang dibersihkan mempunyai 302 rekod dan 11 atribut. Analisis deskriptif dilakukan untuk memahami ciri-ciri atribut serta mengenal pasti trend pada data. Tujuan menganalisis trend data adalah untuk mendapatkan maklumat penting dalam membantu kajian serta membuat keputusan. Atribut Jenis telah dipilih untuk dibuat perbandingan dengan atribut undian IMDb, kategori umur, genre dan negara produksi. Tujuan perbandingan dibuat untuk melihat trend dan taburan jenis bagi atribut-atribut ini. Di samping itu, perbandingan kelas atribut iaitu skor IMDb turut dibuat dengan semua atribut dalam set data. Namun perbandingan tidak dibuat dengan tiga atribut iaitu undian IMDb, kategori umur dan negara produksi. Berdasarkan semakan awal, dapat dilihat terdapat filem atau rancangan TV kategori umur pilihan ramai mempunyai bilangan undian

tinggi namun memperoleh skor IMDb yang rendah dan terdapat juga filem atau rancangan TV kategori umur kurang menjadi pilihan mempunyai bilangan undian sedikit tetapi memperoleh skor yang tinggi. Hal ini jelas kerana setiap individu mempunyai pandangan dan pendapat yang berbeza, oleh yang demikian perbandingan pola dan trend akan memberi keputusan yang *volatile* atau tidak menentu. Semakan negara produksi pula memperoleh keputusan skor IMDb yang lebih kurang antara setiap negara. Oleh yang demikian, perbandingan lanjut tidak perlu dilakukan.

g) Analisis Inferensi Data Netflix-Imdb

Model ujian pengelasan yang dipilih adalah model *Naive Bayes (NB), Decision Tree (DT), Random Forest (RF)* dan *K-nearest Neighbors (KNN).* Pemilihan model adalah kerana ianya paling banyak digunakan dalam kajian lepas dan paling relevan untuk digunakan dalam set data kajian. Dua model daripada teknik *boosting* iaitu *Gradient Boosting (GB)* dan *Ada Boosting (AB)* turut diguna dalam kajian. Hal ini kerana prestasi model *boosting* yang dapat meningkatkan ketepatan dalam kajian-kajian lepas dan kedua-dua model ini sesuai diguna dalam kajian klasifikasi. Bagi teknik pemecahan data pula, teknik yang digunakan ialah teknik *hold out* (Rajah 3.2) . *Hold out* adalah teknik yang membahagikan set data kepada dua iaitu set latihan dan ujian. Set latihan ialah model yang dilatih manakala set ujian pula digunakan untuk melihat prestasi model tersebut pada data yang tidak kelihatan. Pemisahan yang biasa digunakan ialah 80% data untuk latihan dan baki 20% data untuk ujian.



Rajah 3.2 Contoh pembahagian *hold out*

Hasil dapatan kajian akan dinilai berdasarkan ketepatan, kejituan dapatan semula dan pengiraan-F.

i. Ketepatan

Ketepatan ditakrif sebagai nisbah pemerhatian yang diramal dengan betul kepada jumlah pemerhatian.

Ketepatan =(TN + TP) / (TN + FP +TP + FN)

ii. Kejituan

Kejituan mengukur nisbah pemerhatian positif yang diramal dengan betul kepada jumlah pemerhatian positif yang diramal.

Kejituan =TP / (TP + FP)

iii. Dapatan Semula

Dapatan semula ialah nisbah pemerhatian positif yang diramal dengan betul kepada semua pemerhatian dalam kelas sebenar.

Dapatan Semula =TP / (TP + FN)

iv. Pengiraan F

Pengiraan F ialah purata kejituan dan mengambil kira kejituan dan dapatan semula.

Pengiraan F =2 x (Ketepatan x Dapatan Semula)/ (Ketepatan + Dapatan Semula)

4 DAPATAN KAJIAN

a) Hasil Analisis Deskriptif Data Netflix –Imdb

Kajian menggabungkan set data filem dan rancangan TV. Teknik *exploratory data analysis (EDA)* telah digunakan dengan membina graf perbandingan untuk memahami set data yang digunakan dalam kajian ini iaitu di mana karya jenis filem lebih banyak diundi berbanding karya jenis rancangan TV. Selain itu, hasil kajian juga mendapati kebanyakan penonton di platform Netflix

adalah berumur melebihi 17 tahun. Hal ini dibuktikan dengan mendapati genre kategori umur R mendominasi jenis filem sebanyak 44.04% di mana kategori umur membawa maksud filem tersebut sesuai ditonton oleh penonton 17 tahun ke atas manakala kategori umur TV-MA mendominasi jenis rancangan TV sebanyak 12.91% di mana TV-MA adalah rancangan yang sesuai ditonton oleh penonton matang. Di samping itu, hasil kajian turut menunjukkan genre pilihan penonton ialah genre drama dan rancangan TV manakala filem pilihan ramai adalah daripada negara US.

Hasil analisis graf diteruskan dengan melihat trend perbandingan setiap atribut dengan kelas atribut. Hasil daripada analisis graf memaparkan bahawa trend prestasi rating IMDb adalah dimalarkan. Ini bermaksud tiada perbezaan skor IMDb tinggi (Tinggi) yang jelas untuk dibincangkan. Kebanyakan filem memperoleh skor IMDb antara 7.0-7.9 dan rancangan TV pula 8.0-8.9 di mana rancangan TV memperoleh skor klasifikasi dalam kategori tinggi.

Perbandingan graf skor IMDb dan atribut genre pula menunjukkan genre dokumentasi memperoleh keputusan paling tinggi iaitu skor antara 9.0-10.0. Namun apabila dibuat penelitian lanjut, bilangan karya genre dokumentasi hanya mempunyai peratusan yang sangat kecil iaitu 0.33%. Genre yang menjadi pilihan penonton ialah drama dengan peratusan 9.27% mendapat skor 8.0-8.9 iaitu skor IMDb kategori tinggi.

Bagi trend mengikut tajuk dan pengarah, dapat dilihat atribut ini tidak dapat memberi kesan kepada skor IMDb kerana terdapat tajuk filem atau rancangan TV yang tidak mempunyai pengarah namun masih mendapat skor IMDb yang tinggi. Hal ini jelas menunjukkan penonton tidak menilai sesuatu karya dari segi tajuk mahupun pengarah karya tersebut. Di samping itu, trend mengikut durasi juga diperhatikan. Hasil kajian mendapati filem atau rancangan TV berdurasi 100 minit ke bawah sering memperoleh skor IMDb 8.0 ke atas iaitu klasifikasi tinggi. Hal ini kerana keterbatasan fokus manusia yang semakin mengurang mengikut masa. Dari segi tahun ditayangkan pula, filem atau rancangan 2005 ke atas sering mendapat skor IMDb tinggi. Hal ini kerana, seperti analisis graf perbandingan jenis dan kategori umur, kebanyakan penonton adalah berumur 17 tahun dan ke atas. Analisis trend terakhir ialah analisis bilangan musim, berdasarkan graf, filem atau rancangan TV yang mempunyai bilangan lebih daripada satu musim sering mendapat skor IMDb yang tinggi. Penelitian lanjut dibuat dan mendapati rancangan TV merupakan jenis karya yang mempunyai banyak musim dan kebanyakan karya bergenre drama

mendapat skor IMDb yang lebih tinggi. Oleh yang demikian dapat disimpulkan penonton sangat berminat pada rancangan TV bergenre drama yang bersiri atau bermusim.

b) Hasil Pembangunan Model Dan Prestasi Model

Proses pembangunan model diteruskan dengan hanya menggunakan atribut yang memiliki nilai *correlation coefficient* kurang daripada 0.9. Setiap model pembelajaran mesin telah melalui tiga kali pengujian dengan menggunakan teknik *hold out* dengan pemecahan data latihan dan data ujian yang berbeza. Jadual 4.1- Jadual 4.3, memaparkan hasil keputusan bagi semua algoritma pembelajaran mesin yang digunakan.

Jadual 4.1 Keputusan prestasi ramalan rating IMDb bagi teknik *hold out* 60% data latihan dan 40% data ujian

Model	**NB**	**DT**	**RF**	**KNN**	**GB**	**AB**
Kejituan	**0.86**	0.82	0.85	0.71	**0.87**	0.77
Dapatan Semula	**0.86**	0.82	0.85	0.74	**0.86**	0.78
Pengiraan F	**0.86**	0.81	0.84	0.71	**0.85**	0.74
Ketepatan(%)	**86**	82	85	74	**86**	78

Jadual 4.2 Keputusan prestasi ramalan rating IMDb bagi teknik hold out 75% data latihan dan 25% data ujian

Model	**NB**	**DT**	**RF**	**KNN**	**GB**	**AB**
Kejituan	0.80	0.82	0.82	0.75	**0.89**	0.82
Dapatan Semula	0.80	0.83	0.83	0.76	**0.87**	0.82
Pengiraan F	0.80	0.82	0.83	0.73	**0.85**	0.82
Ketepatan(%)	80	83	83	76	**87**	82

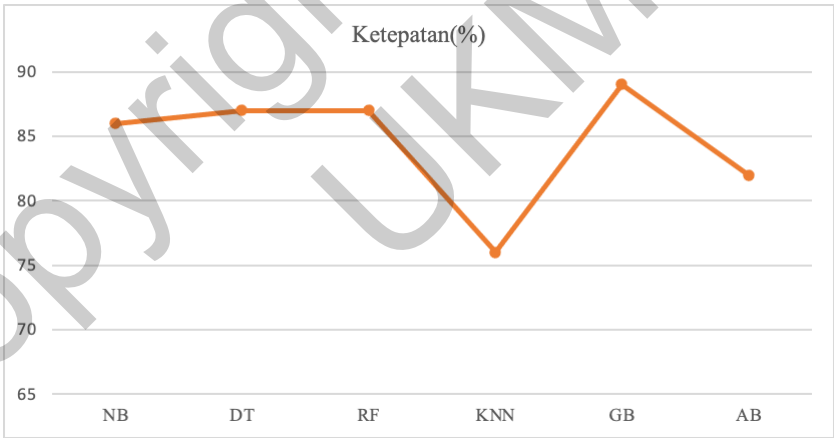
Jadual 4.3 Keputusan prestasi ramalan rating IMDb bagi teknik hold out 80% data latihan dan 20% data ujian

Model	**NB**	**DT**	**RF**	**KNN**	**GB**	**AB**
Kejituan	0.80	0.87	0.87	0.75	**0.89**	0.81
Dapatan Semula	0.80	0.87	0.87	0.75	**0.89**	0.80
Pengiraan F	0.80	0.86	0.87	0.71	**0.88**	0.81
Ketepatan(%)	80	87	87	75	**89**	80

Pengujian dan penilaian model dibina bagi mendapatkan model yang terbaik. Hasilnya model pengelasan prestasi rating IMDb bagi platfom Netflix berjaya dikenal pasti. Enam algoritma digunakan untuk pengelasan ini dan sebanyak 18 model pengujian berjaya dijalankan dengan menggunakan teknik *hold out*. Di antara 18 model yang dijana, proses pemilihan model terbaik dilaksana. Bagi teknik *hold out* pecahan 60% data latihan dan 40% data ujian, model yang terbaik ialah NB dan GB. Bagi teknik *hold out* pecahan 75% data latihan dan 25% data ujian serta teknik *hold out* pecahan 80% data latihan dan 20% data ujian, model terbaik ialah model GB. Model yang memperoleh ketepatan yang paling tinggi bagi setiap algoritma dipilih dan dibanding dengan model yang lain.

Jadual 4.4 Perbandingan keputusan prestasi ramalan rating IMDb bagi model-model kajian

Model	**NB**	**DT**	**RF**	**KNN**	**GB**	**AB**
Kejituan	0.86	0.87	0.87	0.75	**0.89**	0.82
Dapatan Semula	0.86	0.87	0.87	0.76	**0.89**	0.82
Pengiraan F	0.86	0.86	0.87	0.73	**0.88**	0.82
Ketepatan(%)	86	87	87	76	**89**	82



Rajah 4.1 Graf garisan ketepatan bagi setiap model

Berdasarkan Jadual 4.4 dan Rajah 4.1, dapat dilihat, model GB merupakan model yang terbaik dan memperoleh keputusan terbaik bagi ketepatan, kejituan, dapatan semula dan pengiraan F. Peratusan ketepatan yang tertinggi yang diperolah ialah 89%. Model KNN pula mendapat keputusan model yang paling rendah dan mendapat keputusan terendah bagi setiap parameter kejituan, dapatan semula, pengiraan F dan ketepatan. Walaupun model AB tergolong dalam

keluarga *boosting*, model ini mencatatkan keputusan yang rendah dengan 82% bagi ketepatan serta parameter yang lain. Hal ini kerana model AB tidak dapat membuat pemodelan efektif bagi set data yang tidak seimbang dan cabaran utama dalam kajian ialah data yang kecil dan tidak seimbang. Menurut Wang & Sun (2021), Algoritma *AdaBoost* ialah penyelesaian yang berkesan untuk pengelasan, tetapi ia masih memerlukan penambahbaikan dalam masalah data yang tidak seimbang. Model GB pula sangat efektif dalam mengendalikan data tidak seimbang. Hal ini dipersetujui oleh Cahyana et al. (2019) di mana kajian yang dilaksana menyimpulkan algoritma *Gradient Boosting* merupakan satu kaedah yang boleh mengklasifikasikan set data tidak seimbang dengan ketepatan dan prestasi yang tinggi.

Model DT dan RF mencatatkan peratusan ketepatan yang sama namun model RF mendapat nilai pengiraan F 0.1 lebih tinggi daripada DT. Justeru, hal ini melayakkan model RF dan DT menjadi model kedua terbaik selepas model GB. Walau bagaimanapun, keputusan ketepatan bagi model RF secara teorinya sepatutnya lebih baik daripada model DT kerana model RF dibina daripada beberapa bilangan DT. Namun keputusan ketepatan yang diperoleh bagi kajian adalah sama. Hal ini kerana bilangan data yang kecil menyebabkan ciri-ciri paling penting yang dipilih untuk pemodelan adalah sama bagi kedua-dua model. Saiz set data yang relatif mudah juga tidak memerlukan kompleksiti tambahan yang terdapat dalam model RF turut menyumbang kepada perkara ini. Oleh yang demikian, penggunaan model DT sudah mencukupi untuk kajian. Model ketiga terbaik ialah model NB yang memperoleh ketepatan 86% dan diikuti model AB yang mendapat 82% ketepatan dan 0.82 bagi kejituan, dapatan semula dan pengiraan F. Oleh yang demikian, dapat disimpulkan teknik *boosting* dapat meningkatkan ketepatan model. Model GB bersifat adaptif dan fleksibel dapat menghasilkan ketepatan yang paling tinggi bahkan pada set data yang kecil dan tidak seimbang. Meskipun model AB mungkin tidak menghasilkan keputusan yang sama, ini tidak bermakna model tersebut tidak dapat memberikan hasil yang baik. Jika diguna pada set data yang seimbang, berkemungkinan model AB juga akan menghasilkan keputusan yang sama.

5 KESIMPULAN

Kajian berjaya mengenal pasti teknik dan algoritma terbaik untuk pengelasan rating IMDb bagi filem dan rancangan TV di platform Netflix dengan mencadangkan model GB sebagai model

terbaik. Model ini memperoleh ketepatan yang paling tinggi dan membuktikan *boosting* berjaya meningkatkan ketepatan model. Kemampuan model GB yang bersifat adaptif dan sangat fleksibel dapat memberi keputusan ketepatan yang terbaik pada set data yang kecil dan tidak seimbang. Walaupun keputusan yang sama tidak diperoleh bagi model AB, namun tidak bermakna model ini tidak dapat memberi keputusan yang baik. Sekiranya model ini diguna pada data yang seimbang, berkemungkinan model ini turut menghasilkan ketepatan yang tinggi. Uji kaji yang dijalankan dapat memberi panduan asas kepada kajian baru dalam bidang klasifikasi hiburan dan dapat memberi panduan tentang penggunaan GB dalam meningkatkan ketepatan model.

Hasil analisis deskriptif dan trend yang diperoleh sedikit sebanyak dapat memberi maklumat kepada penggiat industri hiburan serta penyedia platform penstriman dalam talian bagi membuat keputusan strategi bisnes serta pemasaran dalam meningkatkan mutu servis kepada pelanggan. Justeru, dengan penambahbaikan mutu servis secara tidak langsung memberi kebaikan kepada pengguna.

Objektif kajian adalah untuk mengaplikasi teknik pembelajaran mesin terhadap prestasi rating IMDb dan faktor atribut yang mempengaruhi prestasi tersebut. Namun model yang dibina adalah terhad kepada prestasi rating sahaja. Data sebenar adalah bersifat luas dan tidak tersusun. Bilangan data yang kecil serta kelas label yang tidak seimbang berkemungkinan menyebabkan kebarangkalian tinggi untuk model salah mengklasifikasikan kelas label serta mengurangkan ketepatan model tersebut.

Oleh yang demikian, dicadangkan kajian pada masa depan menggunakan data yang besar dan seimbang bagi mendapatkan keputusan ketepatan yang lebih baik. Atribut lain boleh ditambah dalam perlombongan data agar dapat memberi gambaran jelas hubung kait dengan prestasi rating IMDb dan secara tidak langsung dapat membantu meningkatkan ketepatan model.

Hasil kajian dijangka dapat menyumbang kepada dua aspek utama iaitu kajian dalam bidang pembelajaran mesin serta memberi penanda aras bagi penyedia platform penstriman dalam talian bagi meningkatkan mutu servis mereka. Dengan terhasilnya model ini, diharapkan penyedia platform penstriman dalam talian lebih cakna dalam merancang strategi bisnes dan pemasaran serta dapat membekalkan karya-karya yang bermutu dan memenuhi cita rasa penonton.

RUJUKAN

Anmadwar, R., Kulkarni, P. & Pasomba, J. 2023. Role of big data analytics in media and entertainment industry. AIP Conference Proceedings, 2736(1): 6. https://doi.org/10.1109/CIACT.2018.8480403

Azahari, M. 2022. Pembelajaran Mesin: Satu Keperluan Masa Kini. Universiti Tun Hussein Onn Malaysia. https://news.uthm.edu.my/ms/2022/10/pembelajaran-mesin-satu-keperluan-masa-kini [20 Oktober 2023].

Baradwaj, B.K. & Pal, S. 2012. Mining educational data to analyze students' performance. *IJACSA* (2, 6): 63-69. https://doi.org/10.48550/arXiv.1201.3417

Bristi, W.R., Zaman, Z. & Sultana, N. 2019. Predicting imdb rating of movies by machine learning techniques. *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT):* 1-5. https://doi.org/10.1109/ICCCNT45670.2019.8944604

Cahyana, N., Khomsah, S., & Aribowo, A., S. 2019. Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting, *5th International Conference on Science in Information Technology (ICSITech*: 217-222. https://doi.org/ 10.1109/ICSITech46713.2019.8987499.

Dixit, P., Hussain, S. & Singh, G. 2020. Predicting the IMDB rating by using EDA and machine learning Algorithms. *International Journal of Scientific Research in Computer Science, Engineering dan Information Technology* (IJSCSEIT) : 441-446.

Gonzalez, E. 2022. Netflix Top Rated Movies and TV Shows. https://www.kaggle.com/datasets/thedevastator/netflix-top-rated-movies-and-tv-shows-2020-2022 [20 Ogos 2023].

Gaenssle, S., Budzinski, O. and Astakhova, D. 2018. Conquering the box office: Factors influencing success of international movies in Russia. *Review of Network Economics* 17(4) : 245-266.

Gandhi, R. 2018. Naive Bayes Classifier. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [20 Oktober 2023].

Gupta, A. 2023. Feature Selection Techniques in Machine Learning. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/ [20 November 2023].

Gupta, V., Jain, N., Garg, H., Jhunthra, S., Mohan, S., Omar, A.H. & Ahmadian, A. 2022. Predicting attributes based movie success through ensemble machine learning. *Multimedia Tools and Applications :*1-30.

Han, J., Kamber, M. & Pei, J. 2012. *Data Mining Concepts and Technique 3rd edition* : (111-113).

Lall, S. & Sivakumar, R. 2021. A Real-World Dataset of Netflix Videos and User Watch-Behavior: Analysis and Insights. *ICC 2021-IEEE International Conference on Communications* : 1-7. https://doi.org/10.1109/ICC42927.2021.9500669

Lawton, G. 2023. What is boosting in machine learning? https://www.techtarget.com/searchenterpriseai/feature/What-is-boosting-in-machine-learning [15 Disember 2023].

Maddodi, S. & Prasad, K. 2019. Netflix Bigdata Analytics-The Emergence of Data Driven Recommendation. *International Journal of Case Studies in Business, IT, and Education (IJCSBE)* 3(2) : 41-51. https://doi.org/10.5281/zenodo.3510316

Mhowwala, Z., Sulthana, A.R. & Shetty, S.D. 2020. Movie Rating Prediction using Ensemble Learning Algorithms. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(8). https://doi.org/10.14569/IJACSA.2020.0110849

Priyanganie, A.G.D.L.C. 2021. *Classifier To Predict The Ratings Of Upcoming Movies,* Universiti of Colombo.

Sadashiv, S., Sween, S. & Sankruth, S. 2021. Movie success prediction using machine learning. *Int. Res. J. Mod. Eng. Technol. Sci*, 3: 1-4 .

Saini, A. 2022. An Introduction to Random Forest Algorithm for Beginners. https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/ [25 Oktober 2023].

Schlee, A. 2020. K-Nearest-Neighbor-IMDB-Project. https://www.linkedin.com/pulse/k-nearest-neighbor-imdb-project-alexander-schlee [27 Oktober 2023].

Sharma, G.A.,Tayade, A.A., Bodke, P. & Ajmire, P. 2019. An Analytical Study of Data Mining for Entertainment. *2nd National Conference on green Technology and Science for Sustainable Development.*

Sivakumar, P. & Ekanayake, J., 2021. Predicting ratings of Youtube videos based on the user comments. *FARS 2021 The 2nd Faculty Annual Research Session*: 71-73.

Wang, W. & Sun, D. 2021. The improved AdaBoost algorithms for imbalanced data classification., *Information Sciences,* 563: 358–374. https://doi.org/10.1016/j.ins.2021.03.042.