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

NUR IZYANI BINTI AHMAD

UNIVERSITI KEBANGSAAN MALAYSIA

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

NUR IZYANI BINTI AHMAD

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI SAINS DAN TEKNOLOGI MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

01 Februari 2024

NUR IZYANI BINTI AHMAD
P113707

PENGHARGAAN

Syukur ke hadrat Illahi yang tidak terhingga kerana dengan limpah kurnia dan izin-Nya, dapat saya menyiapkan kajian ilmiah tahun akhir ini seperti yang telah diharapkan.

Penghargaan yang tidak terhingga dan jutaan terima kasih khas kepada penyelia yang amat dihormati, Puan Siti Aishah Hanawi di atas tunjuk ajar, pandangan dan bimbingan beliau yang telah diberikan kepada saya bagi menyempurnakan kajian tahun akhir ini. Tidak dilupakan, ucapan terima kasih kepada pihak kakitangan Fakulti Teknologi dan Sains Maklumat di atas bantuan dan kerjasama yang diberikan sepanjang tempoh pembelajaran.

Ucapan terima kasih ini juga ditujukan khas buat kedua ibubapa saya yang banyak membantu saya, juga buat suami tercinta yang menyokong serta ahli keluarga saya yang sentiasa menjadi tulang belakang kejayaan saya selama ini. Saya dedikasikan penyelidikan kajian saya ini kepada suami Mohammad Azwan Sahrudin, dan anak-anak tersayang Iman Medina, Hud Anaqi, Ameena Sofia dan Dayyan Umar.

Akhir kata, semoga dengan segala usaha yang telah dilakukan ini akan mendapat keberkatan dan keredhaan daripada Allah S.W.T dan menjadi harapan saya agar kajian ilmiah saya ini dapat membantu organisasi dalam mempertingkatkan produktiviti dengan lebih efisyen serta efektif pada masa hadapan. Ribuan terima kasih diucapkan sekali lagi. Sekian.

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.

RATING CLASSIFICATION OF MOVIES AND TELEVISION SHOWS ON THE NETFLIX STREAMING PLATFORM BASED ON MACHINE LEARNING MODELS

ABSTRACT

In today's era, the entertainment industry plays a crucial role in economic development. The rapid advancement of the industry has resulted in a revolution in the entertainment sector, where online streaming platforms such as Netflix, Amazon TV, Iflix, and others have become the primary choice for viewers. The proliferation of streaming platforms, also known as Over The Top (OTT), has intensified competition among these platforms to ensure that the content they produce meets user preferences. In the entertainment context, the success of a film or television show relies on the ratings given by viewers. Several literacy studies outline various factors such as genre, season, director, duration, and others that influence IMDb ratings. The diversity and breadth of attributes necessitate constant review of past studies, highlighting the use of artificial intelligence methods such as machine learning in predicting and classifying TV and film ratings. Therefore, the study aims to propose a model for rating TV shows and films streamed by online platforms using machine learning algorithms. Several machine learning models such as Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), K-nearest Neighbors (KNN), Gradient Boosting (GB), and Ada Boosting (AB) are employed in the study. The research results indicate that the machine learning model achieving high accuracy in predicting IMDb score ratings is the GB model, with an accuracy of 89%.

KANDUNGAN

Halaman

PENGAKUAN		**ii**
PENGHARGAAN		**iii**
ABSTRAK		**iv**
ABSTRACT		**v**
KANDUNGAN		**vi**
SENARAI JADUAL		**ix**
SENARAI ILUSTRASI		**xi**
SENARAI SINGKATAN		**xii**
BAB I	**PENGENALAN**	
1.1	Pendahuluan	1
1.2	Penyataan Masalah	1
1.3	Objektif Kajian	2
1.4	Skop Kajian	3
1.5	Organisasi Tesis	3
1.6	Kesimpulan	4
BAB II	**KAJIAN LITERASI**	
2.1	Pengenalan	5
2.2	Definisi Analisis Rating	5
2.3	Analisis Rating	6
	2.3.1 Faktor mempengaruhi prestasi rancangan TV atau Filem	6
2.4	Pengenalan Kepada Pembelajaran Mesin	9
	2.4.1 *Naive Bayes*	9
	2.4.2 *Decision Tree*	11
	2.4.3 *Random Forest*	12
	2.4.4 *K-nearest Neighbours*	14
	2.4.5 *Boosting*	15
2.5	Kesimpulan	19

BAB III KAEDAH

3.1	Pengenalan		20
3.2	Tinjauan Kajian		21
3.3	Penyediaan Data		23
	3.3.1	Sumber Data	23
	3.3.2	Deskriptif Data	23
	3.3.3	Hasil Integrasi Data	25
3.4	Pemprosesan Data		25
	3.4.1	Proses Pembersihan Data	26
	3.4.2	Memperbaiki Data Tidak Konsisten	27
	3.4.3	Proses Pengisian Nilai Kosong	27
	3.4.4	Menghapuskan Atribut Tidak Berguna	29
	3.4.5	Pemilihan Ciri	30
	3.4.6	Transformasi Data	31
3.5	Analisis Deskriptif Data Netflix-Imdb		32
3.6	Analisis Inferensi Data Netflix-Imdb		33
	3.6.1	Pembangunan Model Pembelajaran Mesin Data Rating IMDb	33
	3.6.2	Penilaian Prestasi Model Klasifikasi Rating IMDb	34
3.7	Kesimpulan		35

BAB IV DAPATAN KAJIAN

4.1	Pengenalan		37
4.2	Hasil Analisis Deskriptif Data Netflix-Imdb		37
	4.2.1	Atribut Jenis	37
	4.2.2	Kelas Label Skor IMDb	44
4.3	Keputusan Uji Kaji Model Pembelajaran Mesin		51
4.4	Hasil Pembangunan Model Dan Prestasi Model		51
4.5	Kesimpulan		59

BAB V RUMUSAN DAN PENUTUP

5.1	Pengenalan		60
5.2	Rumusan Dan Penemuan Penyelidikan		60
5.3	Sumbangan Kajian		64

5.4	Cadangan Perluasan Kajian	65
5.5	Penutup	65
RUJUKAN		**67**
LAMPIRAN		
Lampiran A	Keputusan Kajian Soal Selidik	70
Lampiran B	Sampel Pemprosesan Data Hilang dan Pengisian Data	77
Lampiran C	Deskripsi Kategori Umur	79
Lampiran D	Kod Negara	80

SENARAI JADUAL

No. Jadual		Halaman
Jadual 2.1	Rumusan faktor yang mempengaruhi prestasi filem atau rancangan TV dalam kajian lepas	8
Jadual 2.2	Rumusan teknik pembelajaran mesin dalam kajian lepas	17
Jadual 3.1	Keputusan tinjauan rating IMDb bagi filem atau rancangan TV di platform OTT	21
Jadual 3.2	Senarai atribut *raw_titles*.csv	24
Jadual 3.3	Senarai atribut *raw_credit*s.csv	24
Jadual 3.4	Senarai atribut setelah proses integrasi	25
Jadual 3.5	Senarai atribut set data	26
Jadual 3.6	Jumlah rekod bagi setiap atribut	27
Jadual 3.7	Jumlah rekod bagi kategori umur	29
Jadual 3.8	Atribut yang dihapuskan secara manual	30
Jadual 3.9	Kedudukan atribut berdasarkan teknik *Correlation Coefficient*	31
Jadual 3.10	Penjanaan kelas label	32
Jadual 3.11	Perbandingan atribut data Netflix-IMDb	33
Jadual 4.1	Bilangan dan peratusan setiap kategori umur mengikut jenis filem dan rancangan TV.	39
Jadual 4.2	Bilangan dan peratusan setiap genre mengikut jenis filem dan rancangan TV.	41
Jadual 4.3	Bilangan dan peratusan setiap negara produksi mengikut jenis filem mahupun rancangan TV.	43
Jadual 4.4	Bilangan dan peratusan skor IMDb mengikut jenis filem mahupun rancangan TV.	45
Jadual 4.5	Bilangan dan peratusan skor IMDb mengikut genre.	46
Jadual 4.6	Atribut dikekalkan dalam pemodelan	52
Jadual 4.7	Matriks kekeliruan NB	53
Jadual 4.8	Keputusan prestasi ramalan rating IMDb bagi model NB	53

Jadual 4.9	Matriks kekeliruan DT	54
Jadual 4.10	Keputusan prestasi ramalan rating IMDb bagi model DT	54
Jadual 4.11	Matriks kekeliruan RF	54
Jadual 4.12	Keputusan prestasi ramalan rating IMDb bagi model RF	55
Jadual 4.13	Matriks kekeliruan KNN	55
Jadual 4.14	Keputusan prestasi ramalan rating IMDb bagi model KNN	55
Jadual 4.15	Matriks kekeliruan GB	56
Jadual 4.16	Keputusan prestasi ramalan rating IMDb bagi model GB	56
Jadual 4.17	Matriks kekeliruan AB	57
Jadual 4.18	Keputusan prestasi ramalan rating IMDb bagi model AB	57
Jadual 4.19	Perbandingan keputusan prestasi ramalan rating IMDb bagi model-model kajian	57
Jadual 5.1	18 model pengujian	62
Jadual 5.2	Keputusan prestasi ramalan rating IMDb bagi teknik *hold out* 60% data latihan dan 40% data ujian	63
Jadual 5.3	Keputusan prestasi ramalan rating IMDb bagi teknik *hold out* 75% data latihan dan 25% data ujian	63
Jadual 5.4	Keputusan prestasi ramalan rating IMDb bagi teknik *hold out* 80% data latihan dan 20% data ujian	63
Jadual 5.5	Perbandingan ketepatan prestasi ramalan rating IMDb bagi model-model kajian	64

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 2.1	Contoh ramalan DT berdasarkan atribut	11
Rajah 2.2	Model ramalan RF	13
Rajah 2.3	Model ramalan KNN	14
Rajah 2.4	Model *Boosting*	16
Rajah 3.1	Metodologi kajian	21
Rajah 3.2	Gambaran data hilang	21
Rajah 3.3	Contoh pembahagian *hold out*	34
Rajah 3.4	Matriks kekeliruan	35
Rajah 4.1	Peratusan filem dan rancangan TV	38
Rajah 4.2	Bilangan undian filem dan rancangan TV	38
Rajah 4.3	Graf bagi bilangan setiap kategori umur mengikut jenis filem dan rancangan TV.	40
Rajah 4.4	Graf perbandingan genre mengikut jenis filem dan rancangan TV	42
Rajah 4.5	Graf perbandingan negara produksi mengikut jenis filem dan rancangan TV	44
Rajah 4.6	Graf skor IMDb mengikut jenis filem mahupun rancangan TV	45
Rajah 4.7	Graf skor IMDb mengikut genre	47
Rajah 4.8	Graf skor IMDb mengikut tajuk	48
Rajah 4.9	Graf skor IMDb mengikut pengarah	49
Rajah 4.10	Perbandingan graf garisan skor IMDb mengikut durasi	49
Rajah 4.11	Perbandingan graf garisan skor IMDb mengikut tahun ditayangkan	50
Rajah 4.12	Perbandingan graf garisan skor IMDb mengikut bilangan musim	51
Rajah 4.13	Graf garisan ketepatan bagi setiap model	58

SENARAI SINGKATAN

NB	*Naive Bayes*
DT	*Decision Tree*
RF	*Random Forest*
KNN	*K-Nearest Neighbours*
SVM	*Support Vector Machine*
LR	*Logistic Regression*
GB	*Gradient Boosting*
AB	*Ada Boosting*
XB	*XGBoost*
EDA	*Exploratory data analysis*
IMDb	Pangkalan Data Filem Internet
TV	Televisyen
TP	Positif Benar / *True Positive*
TN	Negatif Benar / *True Negative*
FP	Positif Palsu / *False Positive*
FN	Negatif Palsu / *False Negative*

BAB I

PENGENALAN

1.1 PENDAHULUAN

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. Bagi para saintis, kebolehan untuk mengenal pasti maklumat platform penstriman adalah sesuatu yang penting. Hal ini kerana, para saintis dan penyedia platform penstriman dalam talian dapat memahami dan membuat ramalan cita rasa penonton bagi memastikan platform relevan serta berdaya saing dalam menyampaikan kandungan yang berkesan kepada 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).

Teknik perlombongan data hiburan merupakan satu teknik yang mencabar kerana jumlah data yang diperoleh dari pangkalan data bersifat sentiasa berubah dan terlalu banyak. Oleh yang demikian, hasil ramalan rating sesebuah filem atau rancangan TV adalah penting dalam menentukan perancangan strategi rancangan-rancangan yang akan disajikan serta menambah baik kualiti servis yang diberikan kepada penonton (Maddodi & Prasad 2020).

1.2 PENYATAAN MASALAH

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.

Klasifikasi merupakan proses membahagikan data ke dalam kumpulan, manakala ramalan data adalah proses membuat ramalan nilai pemboleh ubah berterusan (Baniodeh 2021). Ramalan serta analisis rating sesebuah filem atau rancangan TV adalah penting untuk kemajuan industri hiburan khususnya platform dalam talian. Namun demikian, proses ini sangat mencabar kerana terdapat pelbagai faktor yang

mempengaruhi rating tersebut. Hal ini membawa kepada perkembangan pesat kepada teknik pembelajaran mesin.

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.

Pelbagai kajian menggunakan teknik perlombongan data yang hampir sama telah dilakukan, namun, kepelbagaian serta keluasan atribut menyebabkan model-model terdahulu perlu sentiasa disemak dan ditambah baik dari masa ke semasa. 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.

1.3 OBJEKTIF KAJIAN

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.

1.4 SKOP KAJIAN

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.

1.5 ORGANISASI TESIS

Tesis mempunyai lima bab iaitu:

1. Bab 1: Bab ini menerangkan latar belakang kajian, penyataan masalah, objektif kajian dan skop kajian.
2. Bab 2: Bab ini memberi penerangan berkaitan kajian kesusasteraan yang dijalankan selari dengan kajian pengelasan rating IMDb. Teknik perlombongan data dan pembelajaran mesin hasil kajian sebelum ini turut dibincangkan dalam bab 2.
3. Bab 3: Bab 3 menerangkan fasa tinjauan, fasa pemprosesan, data serta fasa analisis data secara statistik dan deskriptif. Fasa tinjauan melibatkan fasa membuat tinjauan awal bagi melihat penggunaan platform dalam talian dalam kalangan pengguna. Fasa kedua iaitu fasa pemprosesan data melibatkan proses

pembersihan dan integrasi data. Fasa ketiga iaitu fasa analisis deskriptif memaparkan statistik data melalui graf dan carta.

4. Bab 4: Bab ini menerangkan fasa rating dan pemodelan. Fasa pemodelan ialah fasa membangunkan model prestasi rating menggunakan teknik pembelajaran mesin *Naive Bayes* (NB), *Decision Tree* (DT), *Random Forest* (RF), *K-nearest Neighbours* (KNN), *Gradient Boosting* (GB) dan *Ada Boosting* (AB). Prestasi model dibandingkan antara satu sama lain bagi mengenal pasti model yang terbaik. Hasil pembangunan model seterusnya diuji dengan parameter yang bersesuaian dan model yang memperoleh ketepatan yang paling tinggi merupakan model terbaik bagi kajian.
5. Bab 5: Bab terakhir yang merumuskan kesimpulan kajian berserta sumbangan dan kajian pada masa depan.

1.6 KESIMPULAN

Kajian memberi sumbangan penyelidikan dalam bidang perlombongan data dan pembelajaran mesin dalam industri hiburan. Hasil kajian boleh digunakan untuk membantu meningkatkan ketepatan model ramalan rating IMDb, dan ia juga dapat digunakan oleh penyedia platform penstriman dalam talian seperti Netflix untuk meramal pilihan rancangan TV dan filem pilihan penonton. Seterusnya ia dapat memberi nilai tambah kepada penonton dalam mendapatkan pilihan rancangan TV dan filem yang lebih bermutu dalam platform penstriman dalam talian langganan mereka.

BAB II

KAJIAN LITERASI

2.1 PENGENALAN

Sektor hiburan semakin berkembang pesat dan menjadi sektor kritikal yang menyumbang kepada pembangunan sesebuah negara. Hal ini dibuktikan dengan kejayaan laporan *Theme* daripada *The Motion Picture Associations* pada tahun 2019 yang merekodkan hasil jualan dan penstriman teater filem mencecah $100 bilion buat kali pertama dalam sejarah. Ini didorong oleh rekod global $42.2 bilion dalam *box office* dan $58.8 juta dalam jualan hiburan rumah (Loria 2020). Bab ini membincangkan kajian literasi yang meliputi pengumpulan sumber berkaitan dengan topik kajian untuk dijadikan sebagai rujukan. Bab ini juga membincangkan definisi analisis rating serta kajian-kajian terdahulu melibatkan teknik perlombongan data dan pembelajaran mesin yang digunakan.

2.2 DEFINISI 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.

Seperti perbincangan di dalam Bab I, rating merupakan satu kaedah penting bagi penyedia platform penstriman dalam talian untuk mengetahui serta menilai kualiti dan kejayaan kandungan yang mereka sajikan. Hal ini penting untuk kelangsungan bisnes dan kemajuan platform dalam mengatur strategi pemasaran agar dapat menyediakan kandungan yang bermutu tinggi dan memenuhi cita rasa pengguna.

2.3 ANALISIS RATING

Kejayaan sesuatu filem bergantung kepada perspektif bagaimana rancangan TV atau filem itu dilihat atau dianalisis. Pada peringkat awal bidang hiburan, kejayaan sesebuah karya ditentukan dengan hasil jualan kasar *box office*. Nilai tahunan yang direkodkan menjadi penanda ukur kejayaan karya tersebut.

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.

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).

2.3.1 Faktor mempengaruhi prestasi rancangan TV atau Filem

Terdapat beberapa 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) menunjukkan perkembangan pasaran filem antarabangsa dan domestik di Rusia, di mana pasaran filem antarabangsa mendahului dari tahun 2002 hingga 2014. Kajian ilmiah juga 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 untuk mencari model mampan untuk meramal rating IMDb sesebuah filem. Kajian bertujuan 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.

Di samping itu, terdapat juga kajian yang membuktikan bahawa faktor individu mempengaruhi kejayaan menarik minat penonton terhadap sesebuah karya. Hal ini dibuktikan dengan kajian oleh Mhowwala et al. (2020) di mana kajian tersebut membuktikan kejayaan rating sesebuah filem sering dikaitkan dengan popularitipengarah dalam industri hiburan.

Gupta et al. (2022) mencadangkan satu kajian untuk meramal kadar kejayaan atau kegagalan sesebuah filem bagi meningkatkan pertumbuhan industri perfileman. Objektif kajian adalah untuk membantu industri filem dalam pengurusan kos dan penyasaran penonton dengan meramal rating IMDb. 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-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.

2.4 PENGENALAN KEPADA PEMBELAJARAN MESIN

Pembelajaran mesin merupakan satu cabang dalam bidang sains data yang semakin berkembang pesat. Pembelajaran mesin melibatkan tiga peringkat iaitu peringkat pertama penyediaan dan pembersihan data, seterusnya peringkat integrasi serta analisis data dan yang terakhir peringkat pembangunan model. Perkembangan teknologi dan penggunaan komunikasi digital menjadi penyumbang utama kepada perkembangan data secara digital. Penggunaan media sosial seperti Facebook dan Instagram serta perkembangan e-borang menjadi faktor utama kepada kewujudan pangkalan data yang besar. Namun begitu data-data yang disimpan adalah bersifat mentah dan perlu diolah untuk mendapatkan data yang lebih bermakna.

Keperluan untuk mendapatkan data yang lebih bermakna membawa kepada perkembangan perlombongan data dengan menggunakan pembelajaran mesin. 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).

2.4.1 *Naive Bayes*

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). Berikut merupakan formula teorem Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P = Simbol kebarangkalian.

$P(A|B)$ = Kebarangkalian peristiwa A berlaku sekiranya peristiwa B telah berlaku.

$P(B|A)$ = Kebarangkalian peristiwa B berlaku sekiranya peristiwa A telah berlaku.

$P(A)$ = Kebarangkalian peristiwa B berlaku.

$P(B)$ = Kebarangkalian peristiwa A berlaku.

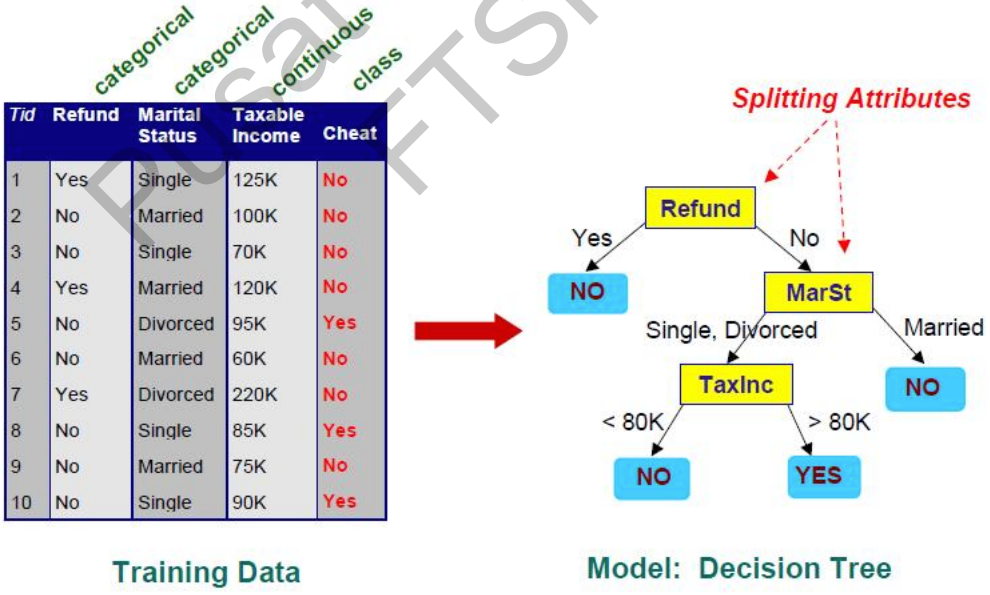
Bristi et al. (2019) menjalankan kajian untuk meramal rating IMDb sesebuah filem menggunakan beberapa model termasuk *Naive Bayes*. Kajian bertujuan membandingkan prestasi algoritma pembelajaran mesin dalam membuat ramalan. Sebanyak 242 set data filem diekstrak daripada laman sesawang Wikipedia dan IMDb. 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.

Kelebihan NB ialah algoritma boleh dilaksanakan dengan mudah dan ramalan dibuat dengan cepat. Disebabkan kelebihan ini, ia mudah dikembangkan dan secara tradisinya merupakan algoritma pilihan untuk aplikasi dunia sebenar yang memerlukan tindak balas segera kepada permintaan pengguna. Namun begitu, terdapat juga kekurangan pada algoritma NB di mana seandainya pemboleh ubah tidak didapati dalam set latihan, algoritma akan memberikan nilai *null* pada atribut tersebut. Hal ini akan menjejaskan markah ketepatan bagi model tersebut dan *tuning* perlu dilakukan untuk mendapatkan keputusan yang lebih baik.

2.4.2 *Decision Tree*

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. Pengelas DT dibina dari atas ke bawah (Rajah 2.1) dan perlu melalui langkah-langkah berikut:

1. Langkah 1 : Memilih kelas atribut.
2. Langkah 2 : Mengira tahap kepentingan atribut melalui formula entropi. Atribut dengan nilai tertinggi seterusnya dipilih menjadi nod akar.
3. Langkah 3 : Atribut terbaik seterusnya dijadikan ranting.
4. Langkah 4 : Pengiraan entropi diteruskan, sekiranya nod akar bernilai sifar, cabang pokok berhenti seolah-olah daun di hujung ranting. Sekiranya lebih dari sifar, ranting tersebut berpecah kepada dua cabang sehingga nilai entropi menjadi sifar.



Rajah 2.1 Contoh ramalan DT berdasarkan atribut

Sumber: mines.humanoriented.com

Terdapat beberapa kajian menggunakan *Decision Tree* dalam meramal rating IMDb . Kajian yang dilaksanakan oleh Sadashiv et al. (2021) membuktikan model DT

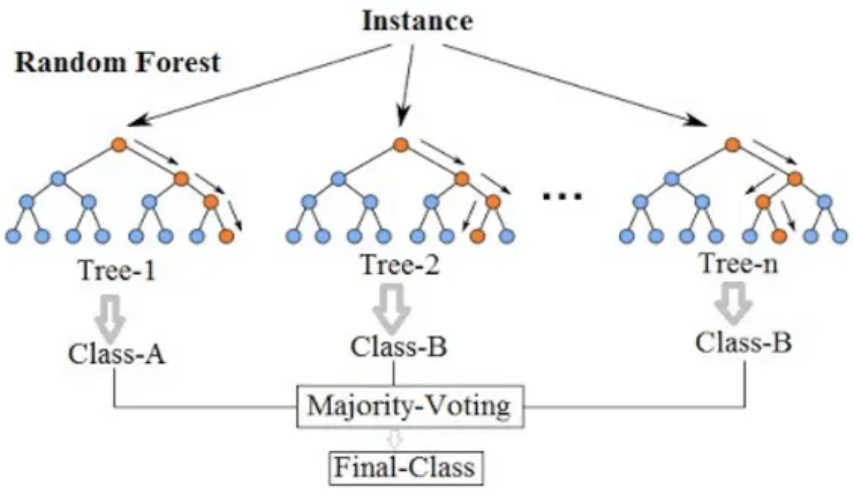
memperoleh ketepatan yang tinggi iaitu 81% dibandingkan NB dengan hanya 72%. Kajian yang dijalankan oleh Bristi et al. (2019) juga membuktikan DT turut memperoleh ketepatan yang lebih tinggi apabila membuat perbandingan ketepatan tanpa melakukan sampel semula set data tanpa penggantian.

Teknik pembelajaran DT adalah mudah untuk difahami dan diaplikasikan. Teknik ini juga tidak memerlukan masa yang lama untuk membuat keputusan. Walau bagaimanapun, bagi mendapatkan hasil yang tepat, teknik ini memerlukan pemprosesan data dan pemilihan atribut yang tepat. Hal ini bertujuan mengelakkan keputusan yang tidak tepat atau berat sebelah kerana bilangan atribut yang terlalu banyak atau tidak normal.

2.4.3 *Random Forest*

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). Berikut merupakan langkah-langkah untuk pelaksanaan pengelas RF (Rajah 2.2):

1. Langkah 1 : Membina subset data daripada data asal melalui kaedah pensampelan baris dan pensampelan ciri.
2. Langkah 2 : Model DT dijana daripada setiap subset data.
3. Langkah 3 : Output daripada setiap DT dinilai. Output akhir dianggap berdasarkan Pengundian Majoriti bagi masalah klasifikasi.



Rajah 2.2 Model ramalan RF

Sumber: medium.com

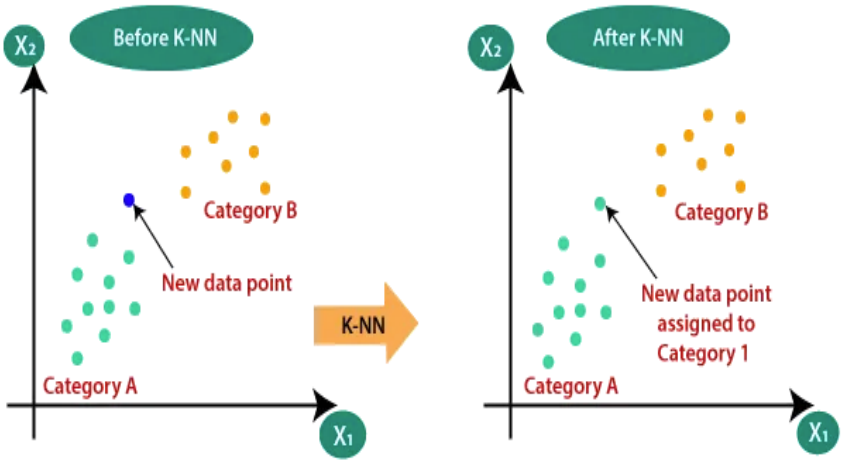
Satu kajian dijalankan oleh Sivakumar dan Ekanayake (2021) untuk membuat ramalan kejayaan sesebuah filem dengan menggunakan ulasan treler di Youtube. Teknik *Natural Language Processing* (NLP) digunakan untuk mengekstrak kata kunci daripada ulasan pengguna. *Tokenizing, Stemming*, dan pengkategorian 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. Model RF menjadi model terbaik bagi kajian ini.

Secara ringkasnya dapat disimpulkan bahawa RF merupakan model canggih yang dapat memberikan ketepatan yang lebih tinggi daripada DT kerana RF memfokuskan kepada mengenal pasti kepentingan ciri. Namun proses tersebut menyebabkan kos penyimpanan dan perkomputeran yang tinggi bagi model pembelajaran RF.

2.4.4 *K-nearest Neighbours*

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). Berikut merupakan langkah-langkah untuk pelaksanaan pengelas KNN (Rajah 2.3):

1. Langkah-1: Pilih nombor K jiran.
2. Langkah-2: Kira jarak Euclidean bagi nombor K jiran.
3. Langkah-3: Ambil K jiran terdekat mengikut jarak Euclidean yang dikira.
4. Langkah-4: Di antara K jiran ini, hitung bilangan titik data dalam setiap kategori.
5. Langkah-5: Tetapkan titik data baharu kepada kategori yang bilangan jirannya adalah maksimum.



Rajah 2.3 Model ramalan KNN

Sumber: medium.com

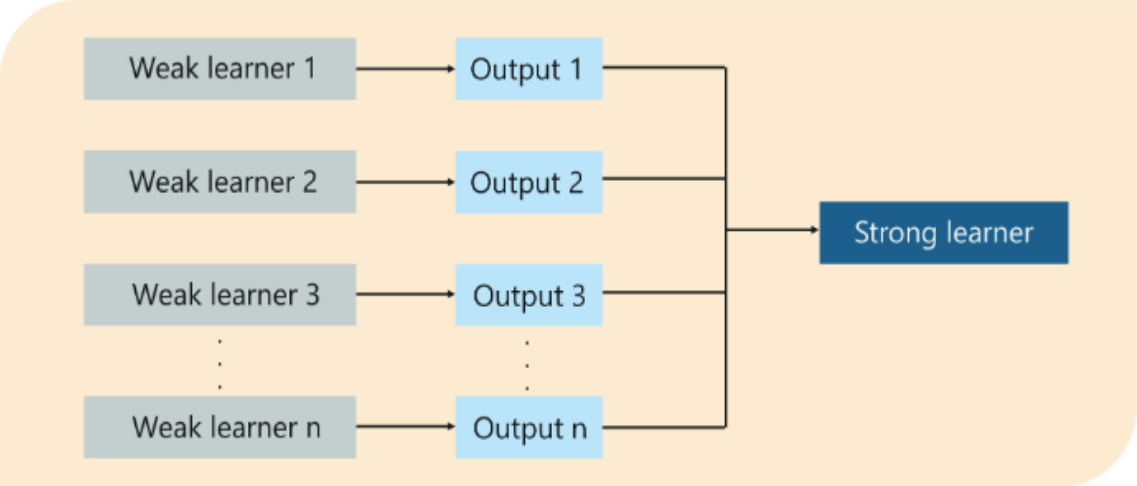
Satu kajian untuk meramal rating bagi filem belum ditayang dilaksanakan oleh Priyanganie (2021). Beliau memilih model KNN kerana model tersebut bersifat fleksibel dan teguh kepada *noisy* data. Di samping itu, model KNN juga menjadi pilihan Bristi et al. (2019) dalam kajian terhadap rating IMDb.

KNN adalah satu model yang mudah ditafsir dan difahami. Ia juga cepat kerana tidak memerlukan fasa latihan untuk dilaksanakan. Walau bagaimanapun kos dan masa perkomputeran KNN pada fasa pengujian adalah tinggi kerana setiap proses dimulakan pada fasa pengujian kerana ketiadaan fasa latihan.

2.4.5 *Boosting*

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). Berikut merupakan langkah-langkah untuk model *boosting* (Rajah 2.4) :

1. Langkah 1: Menetapkan berat awal untuk semua model.
2. Langkah 2: Melatih model lemah.
3. Langkah 3: Membuat pengiraan dan analisis pada ralat bagi model yang tidak dapat mengklasifikasi dengan baik atau model lemah.
4. Langkah 4: Menambahbaik berat bagi model yang lemah.
5. Langkah 5: Menggabungkan model-model lemah bagi semua kitaran untuk menghasilkan satu model yang mempunyai ketepatan yang tinggi.



Rajah 2.4 Model *Boosting*

Sumber: edureka.co

Terdapat pelbagai jenis algoritma *boosting* dalam pembelajaran mesin. Namun antara yang terkenal dan sering digunakan ialah *Ada Boosting* (AB), *Gradient boosting* (GB) dan *XGBoost* (XB). 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. XB pula melatih himpunan algoritma serentak dan selari, dan kemudian pemberat diselaraskan dan disalurkan semula kepada kesemuanya secara kolektif untuk meningkatkan ketepatan keseluruhannya. Setiap algoritma dilatih secara berasingan merentas berbilang CPU atau GPU, yang mengurangkan masa latihan dan meningkatkan prestasi (Lawton 2023).

Boosting merupakan model pembelajaran mesin yang semakin menjadi pilihan kerana ketepatannya. Antara kajian ilmiah yang menggunakan GB ialah kajian ilmiah oleh Omtunde et al. (2022) yang membuat ramalan kadar kejayaan filem melibatkan 26 atribut. Hasil kajian beliau mendapati GB memperoleh ketepatan yang paling tinggi. Dixit et al. (2020) membuat kajian bagi meramal rating IMDb menggunakan model regresi dan klasifikasi. Hasil kajian ilmiah mereka menunjukkan bahawa bilangan pengundi, tempoh, belanjawan dan genre utama mempengaruhi rating IMDb. XB adalah model terbaik untuk model regresi dan GB pula merupakan model terbaik untuk model klasifikasi.

Satu kajian yang dijalankan oleh Mhowwala et al. (2020) untuk menilai kejayaan atau kegagalan sesebuah filem sebelum ditayang dan membandingkannya dengan prestasinya. Beliau menggabungkan data dari IMDb, komen di Youtube dan maklumat di Wikipedia. Hasil kajian mendapati XB menunjukan ketepatan yang terbaik dalam meramal rating IMDb. Gupta et al. (2022) membuat satu kajian menggunakan model KNN, *Support Vector Machine* (SVM) , GB, AB dan XB. Hasil kajian ini juga tetap menunjukkan *Boosting* iaitu GB memperoleh ketepatan yang tinggi manakala KNN memperoleh ketepatan yang terendah. Jadual 2.2 merumuskan hasil teknik perlombongan yang digunakan dalam kajian-kajian sebelum ini.

Jadual 2.2 Rumusan teknik pembelajaran mesin dalam kajian lepas

Nama Penulis	**Objektif**	**Teknik**	**Hasil Kajian**
Briti et al. (2019)	Membuat ramalan rating IMDb.	NB, DT, KNN, *Bagging*, RF	RF model terbaik dengan ketepatan paling tinggi.
Sadashiv et al. (2021)	Membuat ramalan awal kejayaan filem sebelum ditayangkan.	NB, DT, KNN, SVM, *Logistic Regression* (LR), RF	Model RF memperoleh ketepatan paling tinggi.
Mhowwala et al. (2020)	Membuat ramalan rating sesebuah filem bagi menentukan kejayaan atau kegagalan sebelum tayangan.	RF dan XB	Model XB model terbaik dengan ketepatan paling tinggi.
Dixit et al. (2020)	Membuat ramalan rating IMDb menggunakan model regresi dan klasifikasi.	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.

bersambung…

…sambungan

Sivakumar dan Ekanayake (2021)	Membuat ramalan awal kejayaan filem.	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)	Membuat ramalan kadar kejayaan atau kegagalan sesebuah filem untuk membantu industri filem dalam pengurusan kos dan pemilihan penonton dengan meramal rating IMDb.	KNN, SVM, GB, AB, XB	Model GB model terbaik dengan ketepatan paling tinggi.
Omotunde et al. (2022)	Membuat ramalan kadar kejayaan filem dengan menggunakan 26 atribut.	GB, SVM	GB model terbaik dengan ketepatan paling tinggi.
Priyanganie (2018)	Membuat ramalan rating sesebuah filem sebelum ditayangkan.	LR, KNN, DT, RF, NB, *Bagging, Boosting*	Model DT dengan peningkatan *bagging* model terbaik dengan ketepatan paling tinggi.
Lall dan Sivakumar (2021)	Membuat ramalan sama ada penonton akan meneruskan atau meninggalkan rancangan TV bermusim	LR, NB, SVM , RF	Model RF memperoleh ketepatan paling tinggi.

2.5 KESIMPULAN

Kajian literasi sebelum ini dapat memberi maklumat tentang ketersediaan data, penggunaan metodologi kajian lepas dan membantu mengenal pasti model-model pembelajaran mesin yang digunakan dalam kajian sebelum ini. Daripada Jadual 2.1 kajian-kajian ilmiah sebelum ini menyokong terdapat pelbagai faktor yang mempengaruhi prestasi sesebuah filem atau rancangan TV. Di samping itu, daripada Jadual 2.2, dapat disimpulkan bahawa pelbagai teknik terkini diguna dalam peramalan rating yang mana teknik *boosting* adalah teknik terkini yang menjadi pilihan kerana mempunyai fungsi untuk meningkatkan ketepatan. Oleh itu, kajian ini dilakukan untuk meramal dan mengkelaskan rating dengan menggunakan teknik pembelajaran mesin NB, DT, RF, KNN serta dua teknik *boosting* iaitu GB dan AB .

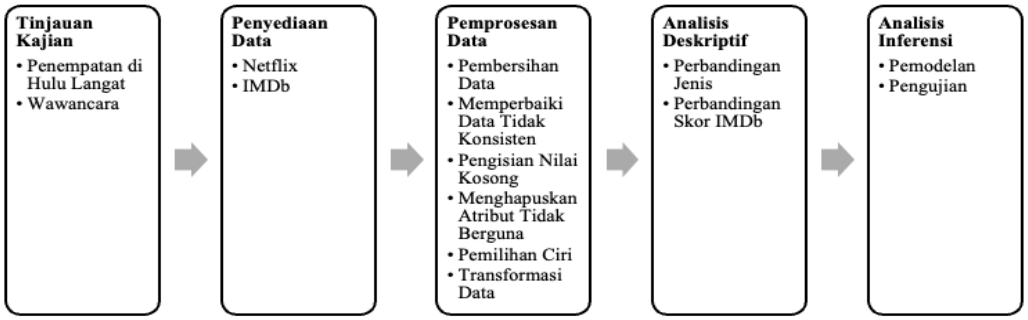
BAB III

KAEDAH

3.1 PENGENALAN

Kajian bertujuan memberikan cadangan model rating rancangan TV dan filem yang ditayangkan di Netflix. Pemilihan Netflix sebagai penyedia platfom penstriman dalam kajian adalah kerana data yang luas serta popularit platfom dalam kalangan penonton.

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

3.2 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 (Lampiran A). Jadual 3.1 memaparkan keputusan tinjauan rating IMDb bagi filem atau rancangan TV di platform OTT.

Jadual 3.1 Keputusan tinjauan rating IMDb bagi filem atau rancangan TV di platform OTT

No	Soalan	Jawapan	Peratusan(%)
1.	Umur	20-30	29.4
		31-40	52.9
		41-50	17.6
		51-60	0
2.	Jantina	Perempuan	52.9
		Lelaki	47.1
3.	Adakah anda kini melanggan sebarang platform	Ya	100
	OTT (contohnya Netflix, Amazon Prime, Disney+)	Tidak	0
4.	Seberapa kerap anda menggunakan platform OTT	Setiap hari	52.9
	untuk menonton filem atau rancangan TV?	Hujung Minggu	23.5
		Jarang-jarang	23.5

bersambung…

…sambungan

5.	Adakah anda berpendapat kandungan di platform OTT lebih pelbagai dan berbeza berbanding saluran TV tradisional?	Ya	100
		Tidak	0
6.	Adakah anda percaya bahawa platform OTT menawarkan nilai yang lebih baik berbanding langganan TV kabel atau satelit tradisional?	Ya	100
		Tidak	0
7.	Adakah anda berpendapat bahawa popularitи platfom OTT akan terus berkembang dan akhirnya melampaui penonton TV tradisional?	Ya	100
		Tidak	0
8.	Platform mana yang anda langgani?	Netflix	94.1
		Iflix	11.8
		Amazon TV	5.9
		Viu	23.5
		Tonton	0
		Disney Plus	29.4
		Apple TV	5.9
9.	Adakah tuan/puan pernah mendengar berkaitan rating IMDb?	Ya	78.5
		Tidak	23.5
10.	Adakah rating IMDb mempengaruhi keputusan anda untuk menonton sesuatu filem atau rancangan TV?	Ya	88.2
		Tidak	11.8
11.	Adakah anda mempercayai ulasan dan penilaian pengguna di IMDb ketika membuat keputusan sama ada menonton filem atau rancangan TV?	Ya	82.4
		Tidak	17.6
12.	Adakah anda percaya bahawa penilaian yang tinggi di IMDb secara umum menunjukkan filem berkualiti ?	Ya	76.5
		Tidak	23.5
13.	Adakah anda berpendapat bahawa penilaian IMDb mempengaruhi popularitи dan kejayaan sebuah filem atau rancangan TV?	Ya	82.4
		Tidak	17.6
14.	Adakah anda akan mengesyorkan filem atau rancangan TV kepada seseorang berdasarkan rating yang tinggi?	Ya	58.8
		Tidak	41.2
15.	Selain rating IMDb, ulasan berkaitan filem atau rancangan TV di media sosial juga mempengaruhi keputusan pemilihan tontonan. Adakah anda bersetuju dengan perkara ini?	Ya	94.1
		Tidak	5.9

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.

3.3 PENYEDIAAN DATA

3.3.1 Sumber Data

Penyediaan sumber data merupakan proses penting dalam perlombongan data. Dalam kajian, data daripada portal IMDb dan platform Netflix digabungkan menjadi satu set data. Data diperoleh daripada Eduardo Gonzalez yang membantu menggabungkan kedua-dua set data dan memuat naik di laman sumber terbuka Kaggle.com. Kaggle.com ialah komuniti dalam talian untuk saintis data dan pengamal pembelajaran mesin. Kaggle membolehkan pengguna mencari dan menerbitkan set data, meneroka dan membina model dalam persekitaran sains data berasaskan web serta bekerjasama dengan para saintis data mahupun jurutera pembelajaran mesin untuk sesebuah projek.

3.3.2 Deskriptif Data

Set data yang diguna untuk kajian mengandungi maklumat filem dan rancangan TV di platform Netflix sehingga Mei 2022. Eduardo Gonzalez memuat naik dua fail data mentah dan empat fail data yang diolah untuk tujuan kajian. Namun begitu, dua fail data mentah digabungkan untuk dijadikan set data baru untuk kajian.

Set data pertama yang digunakan ialah *raw_titles.csv*. Set data ini mempunyai 5805 rekod dan 13 atribut. Jadual 3.2 adalah jadual senarai atribut bagi *raw_titles.csv*.

Jadual 3.2 Senarai atribut *raw_titles*.csv

Nama Atribut	**Jenis Data**	**Deskriptif**
Indeks	Nominal	Indeks
ID	Nominal	Id filem atau rancangan TV
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
ID IMDb	Nominal	Id IMDb
Skor IMDb	*Float*	Skor IMDb bagi filem atau rancangan TV
Undian IMDb	Integer	Bilangan undian IMDb bagi filem atau rancangan TV

Set data kedua ialah *raw_credits.csv*. Set data ini mempunyai 4523 rekod dan lima atribut. Jadual 3.3 memaparkan senarai atribut bagi *raw_credits.csv*.

Jadual 3.3 Senarai atribut *raw_credit*s.csv

Nama Atribut	**Jenis Data**	**Deskriptif**
Indeks	Nominal	Indeks
ID	Nominal	Id filem atau rancangan TV
Person ID	Integer	Id pelakon atau pengarah
Nama	Nominal	Nama pelakon atau pengarah
Watak	Nominal	Nama karakter dalam filem atau rancangan TV
Peranan	Nominal	Peranan pengarah atau pelakon dalam filem atau rancangan TV

Kedua-dua sumber data diintegrasi menjadi satu set data yang baru. Proses integrasi dan pembersihan dilakukan untuk mendapatkan satu set data baru yang bersih dan berkualiti.

3.3.3 Hasil Integrasi Data

Data *raw_titles.csv* dan *raw_credits.csv* digabungkan menjadi satu set data baru. Jadual 3.4 menunjukkan senarai atribut setelah proses integrasi dilakukan.

Jadual 3.4 Senarai atribut setelah proses integrasi

Bil	Nama Atribut
1	Indeks
2	ID
3	Tajuk
4	Jenis
5	Tahun Ditayangkan
6	Kategori Umur
7	Durasi
8	Genre
9	Negara Produksi
10	Musim
11	ID IMDb
12	Skor IMDb
13	Undian IMDb
14	Pengarah

3.4 PEMPROSESAN DATA

Majoriti set data dunia sebenar sangat terdedah kepada data yang hilang, tidak konsisten dan bising disebabkan oleh sifat asal data tersebut. Ia diagregatkan daripada sumber yang pelbagai menggunakan teknik perlombongan data dan pergudangan (Baheti 2021). Penggunaan algoritma data pada data mentah mungkin menjana hasil yang tidak berkualiti kerana algoritma tidak dapat mengenal pasti corak dengan berkesan. Oleh itu, pemprosesan data yang terdiri daripada pembersihan data, pemilihan ciri dan transformasi data dapat membantu mengubah data menjadi format yang difahami dan menghasilkan pemodelan data yang berkualiti.

3.4.1 Proses Pembersihan 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 pertama 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 . Jadual 3.5 menunjukkan senarai atribut akhir serta jenis data bagi setiap atribut.

Jadual 3.5 Senarai atribut set data

Nama Atribut	**Jenis Data**	**Deskriptif**
Indeks	Nominal	Indeks
ID	Nominal	Id filem atau rancangan TV
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
ID IMDb	Nominal	Id IMDb
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

3.4.2 Memperbaiki Data Tidak Konsisten

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 (Lampiran B).

3.4.3 Proses Pengisian Nilai Kosong

Setelah mengenal pasti rekod data yang digunakan, kajian diteruskan dengan menganalisis data daripada setiap lajur (Lampiran B). 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. Jadual 3.6 merumuskan bilangan rekod bagi setiap atribut.

Jadual 3.6 Jumlah rekod bagi setiap atribut

Nama Atribut	Bilangan Data	Bilangan Data Hilang
Indeks	302	0
ID	302	0
Tajuk	302	0
Jenis	302	0
Tahun Ditayangkan	302	0
Kategori Umur	297	5
Durasi	302	0
Genre	302	0
Negara Produksi	302	0

bersambung…

…sambungan

Musim	74	228
ID IMDb	302	0
Skor IMDb	302	0
Undian IMDb	302	0
Pengarah	228	74

Berikut merupakan visual bilangan rekod dalam setiap atribut menggunakan pengaturcaraan python dalam aplikasi Google Colab. Rajah 3.2 memaparkan hasil visualisasi tersebut.

Rajah 3.2 Gambaran data hilang

Pengisian nilai kosong dibuat dengan melengkapkan atribut 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. Jadual 3.7 memaparkan hasil pivot menggunakan Microsoft Excel bagi atribut kategori umur. 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.

Jadual 3.7 Jumlah rekod bagi kategori umur

Kategori Umur	**Bilangan Data**
PG	23
PG-13	72
R	128
TV-14	30
TV-MA	39
TV-PG	3
TV-Y7	2
Data Hilang	5

3.4.4 Menghapuskan Atribut Tidak Berguna

Proses menghapuskan atribut tidak berguna sering 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 dihapus kerana atribut tersebut tidak memberi sebarang sumbangan mahupun implikasi kepada kajian. Jadual 3.8 menyenaraikan atribut yang dihapus secara manual.

Jadual 3.8 Atribut yang dihapuskan secara manual

Nama Atribut	Jenis Data	Deskriptif
Indeks	Nominal	Indeks
ID	Nominal	Id filem atau rancangan TV
ID IMDb	Nominal	Id IMDb

3.4.5 Pemilihan Ciri

Pemilihan ciri ialah proses mengurangkan bilangan atribut apabila membangunkan model pembelajaran mesin. Pengurangan bilangan atribut dapat membantu mengurangkan kos pengiraan pemodelan dan dalam beberapa situasi dapat meningkatkan prestasi model (Browlee 2020). Objektif proses pemilihan ciri adalah untuk menilai dan mengenal pasti atribut mana yang memberi impak kepada model data. Terdapat pelbagai kaedah pemilihan ciri, antaranya *Chi Square Test, Fisher's Score, Information Gain* dan *Mutual Information, Correlation Coefficient* dan sebagainya. Dalam kajian, *Correlation Coefficient* dipilih sebagai kaedah pemilihan ciri.

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 3.9 memaparkan kedudukan atribut hasil daripada pemilihan ciri.

Jadual 3.9 Kedudukan atribut berdasarkan teknik *Correlation Coefficient*

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

3.4.6 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 3.10 menunjukkan pembinaan kelas label.

Jadual 3.10 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.

3.5 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. Berikut merupakan atribut yang digunakan dalam analisis deskriptif data Netflix-IMDb (Jadual 3.11). 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. Hasil analisis deskriptif dibincangkan dalam Bab IV.

Jadual 3.11 Perbandingan atribut data Netflix-IMDb

Atribut	Perbandingan
Jenis	Undian IMDb
	Kategori Umur
	Genre
	Negara Produksi
Skor IMDb	Jenis
	Genre
	Tajuk
	Pengarah
	Durasi
	Tahun ditayangkan
	Musim

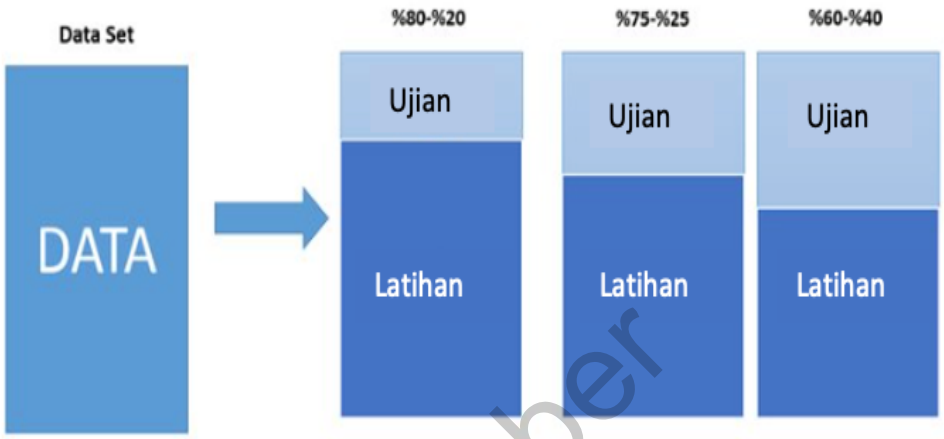
3.6 ANALISIS INFERENSI DATA NETFLIX-IMDB

3.6.1 Pembangunan Model Pembelajaran Mesin Data Rating IMDb

Di dalam kajian, pemilihan model algoritma pengelasan sedia ada dipilih berdasarkan penerangan di Bab II. 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 3.3) . *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. Dalam kajian , beberapa set pemisahan digunakan dan hasil dapatan kajian dibincangkan dalam Bab IV.



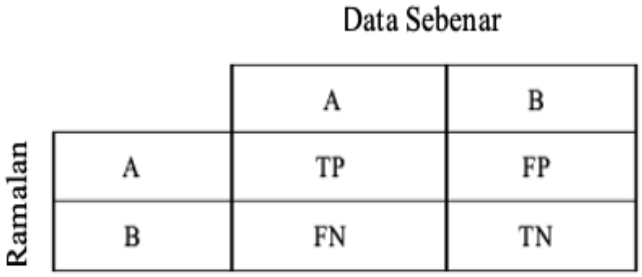
Rajah 3.3 Contoh pembahagian *hold out*

Sumber: researchgate.net

3.6.2 Penilaian Prestasi Model Klasifikasi Rating IMDb

Hasil dapatan analisis model seperti ketepatan, kejituan dapatan semula dan pengiraan-F digunakan untuk perbandingan model dalam Bab IV. Selain itu, matriks kekeliruan juga diguna untuk melihat kekerapan model dalam mengklasifikasikan label data dengan tepat dan betul.

Matriks kekeliruan digunakan untuk membuat perbandingan ramalan dan kelas sebenar bagi dua kelas label negatif dan positif. Positif benar (*True Positive*, TP) dan negatif benar (*True Negative*, TN) menunjukkan sama ada data itu dikelas dengan betul manakala positif palsu (*False Positive*, FP) dan negatif palsu (*False Negative*, FN) menunjukkan data dilabel secara salah oleh algoritma (Rajah 3.4). Ketepatan, kejituan dan dapatan semula juga boleh dikira daripada matriks kekeliruan.



Rajah 3.4 Matriks kekeliruan

a. Ketepatan

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

$$\text{Ketepatan} = (TN + TP) / (TN + FP + TP + FN)$$

b. Kejituan

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

$$\text{Kejituan} = TP / (TP + FP)$$

c. Dapatan semula

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

$$\text{Dapatan Semula} = TP / (TP + FN)$$

d. Pengiraan F

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

$$\text{Pengiraan F} = 2 \times (\text{Ketepatan} \times \text{Dapatan Semula}) / (\text{Ketepatan} + \text{Dapatan Semula})$$

3.7 KESIMPULAN

Bab III membincangkan tentang fasa pertama dan kedua dalam skop kajian. Metodologi menerangkan kaedah yang digunakan dalam kajian. Analisis deskriptif adalah untuk mengekplorasi data serta melihat trend dan statistik atribut. Hasil pelaksanaan fasa pembangunan model diterangkan dengan lebih lanjut dalam Bab IV. Hasil dapatan

kajian kesemua model dinilai dari aspek ketepatan, kejituan, dapatan semula dan pengiraan F.

BAB IV

DAPATAN KAJIAN

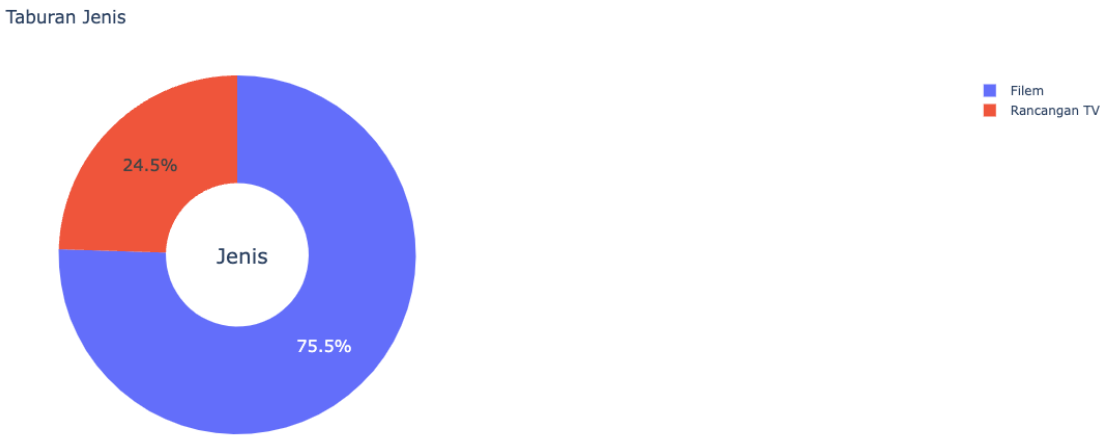
4.1 PENGENALAN

Bab ini membincangkan hasil analisis deskriptif explorasi data dan keputusan yang diperoleh daripada pemodelan algoritma pembelajaran mesin dalam proses perlombongan data. Dalam bahagian ini, keputusan algoritma NB, DT, RF, KNN, GB dan AB dibincangkan dengan lebih terperinci. Maklumat yang diperoleh dalam analisis deskriptif akan dikaji dengan lebih mendalam bagi melihat sebarang hubungkait atribut dan pemodelan. Bab ini juga menerangkan hasil ukuran keputusan menggunakan teknik-teknik penilaian prestasi yang diterangkan dalam Bab III.

4.2 HASIL ANALISIS DESKRIPTIF DATA NETFLIX-IMDB

4.2.1 Atribut Jenis

Platform Netflix membekalkan pelbagai jenis karya untuk tontonan penonton. Terdapat dua jenis taburan yang digunakan untuk set data ini iaitu filem dan rancangan TV. Rajah 4.1 menunjukkan carta pai di mana 75.5% adalah filem manakala 24.5% adalah rancangan TV .

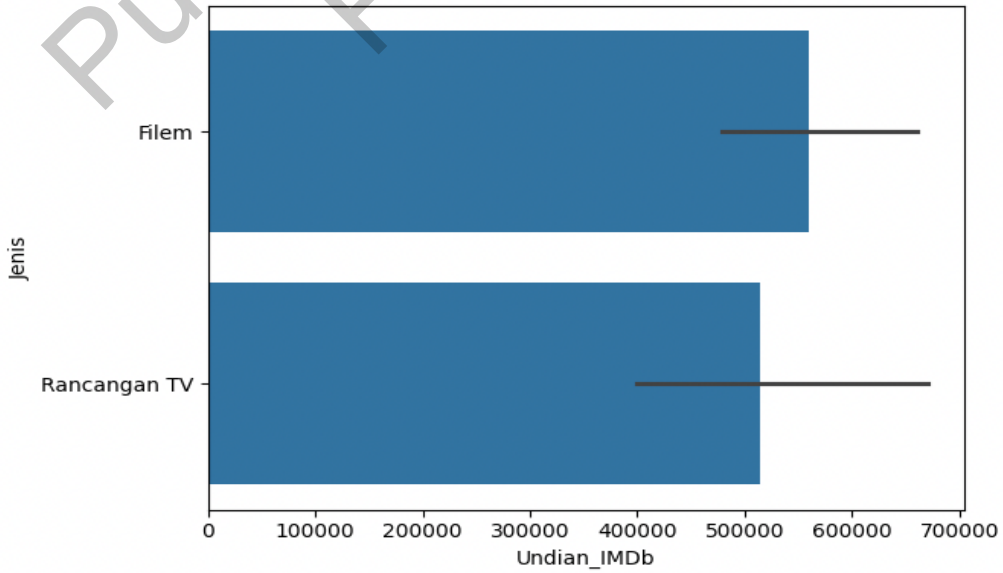


Rajah 4.1 Peratusan filem dan rancangan TV

Berdasarkan carta pai di atas, dapat dirumuskan bahawa penonton lebih memilih karya filem daripada rancangan TV di Netflix.

a. Perbandingan Jenis Mengikut Undian IMDb

Perbandingan undian IMDb mengikut jenis telah dianalisis untuk melihat undian IMDb bagi filem dan rancangan TV. Rajah 4.2 menunjukkan jenis filem mendapat undian lebih banyak berbanding jenis rancangan TV .



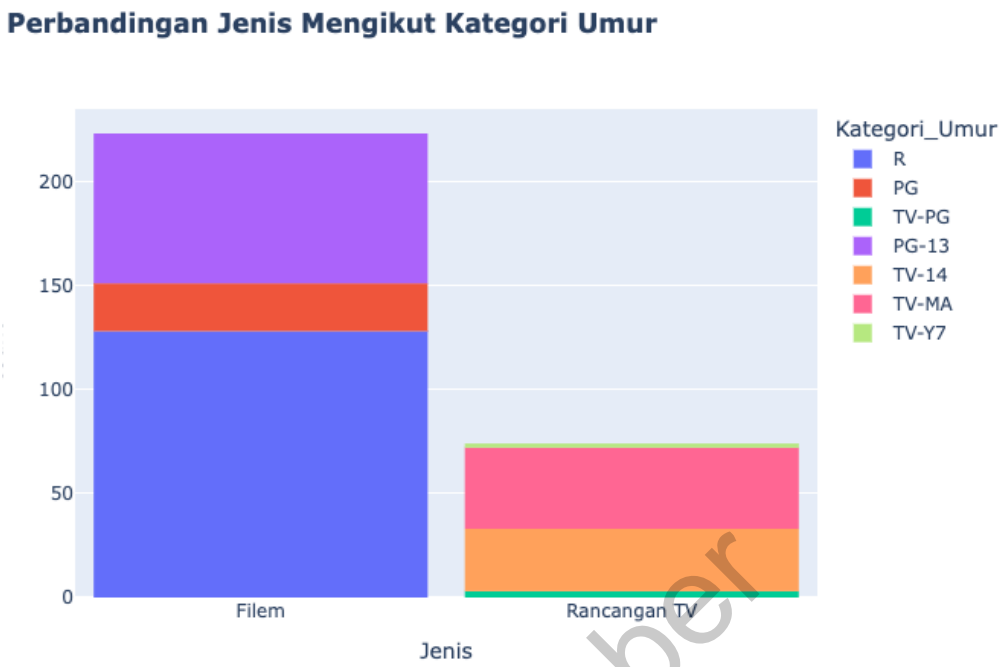
Rajah 4.2 Bilangan undian filem dan rancangan TV

b. Perbandingan Jenis Mengikut Kategori Umur

Perbandingan kategori umur mengikut jenis telah dianalisis untuk melihat kategori umur yang selalu ditonton dan diundi dalam setiap kategori. Jadual 4.1 memaparkan bilangan dan peratusan setiap kategori umur mengikut jenis filem dan rancangan TV. Rajah 4.3 memaparkan graf bagi bilangan setiap kategori umur mengikut jenis filem dan rancangan TV.

Jadual 4.1 Bilangan dan peratusan setiap kategori umur mengikut jenis filem dan rancangan TV.

Jenis	**Kategori umur**	**Bilangan**	**Peratusan(%)**
Filem	PG	23	7.62
	PG-13	72	23.84
	R	133	44.04
Rancangan TV	TV-14	30	9.93
	TV-MA	39	12.91
	TV-PG	3	0.99
	TV-Y7	2	0.66



Rajah 4.3 Graf bagi bilangan setiap kategori umur mengikut jenis filem dan rancangan TV.

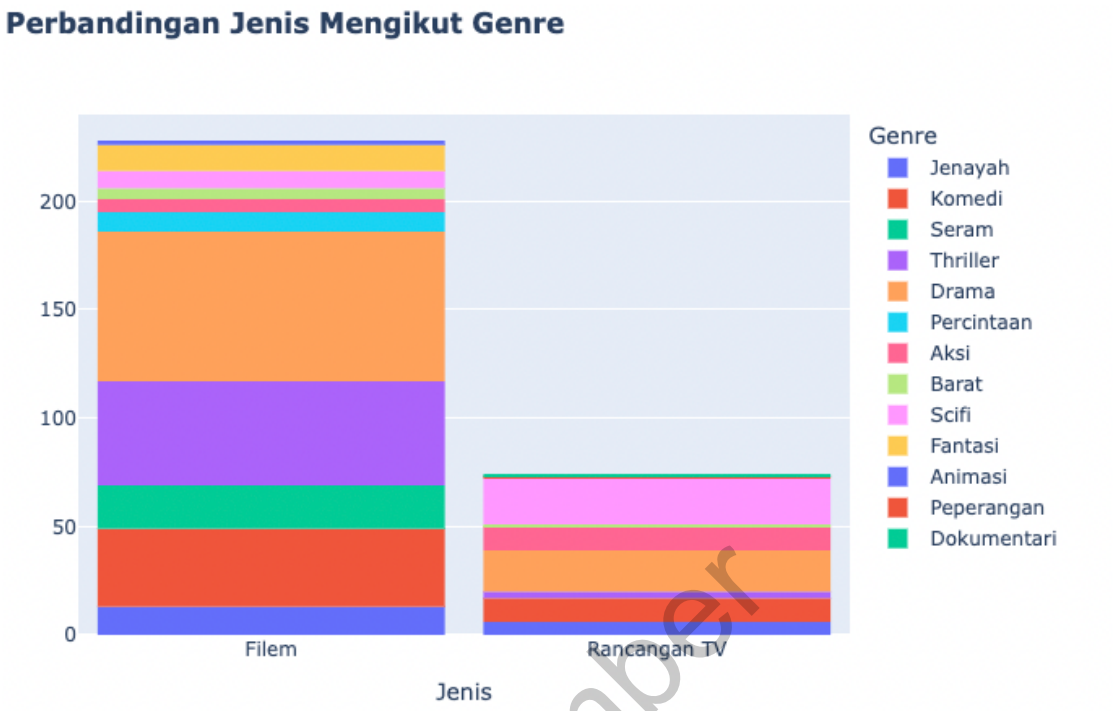
Berdasarkan Jadual 4.1 dan Rajah 4.3, kategori umur R mendominasi jenis filem sebanyak 44.04% di mana kategori umur membawa maksud filem tersebut sesuai ditonton oleh penonton 17 tahun ke atas manakala kategori umur TV-MA mendominasi jenis rancangan TV sebanyak 12.91% di mana TV-MA adalah rancangan yang sesuai ditonton oleh penonton matang. Penerangan bagi kategori umur yang lain dilampirkan dalam Lampiran C. Berdasarkan analisis di atas dapat disimpulkan penonton Netflix kebanyakan daripada penonton berumur melebihi 17 tahun.

c. Perbandingan Jenis Mengikut Genre

Perbandingan genre dianalisis untuk melihat genre yang mendominasi dalam setiap kategori. Jadual 4.2 memaparkan bilangan dan peratusan setiap genre mengikut jenis filem mahupun rancangan TV. Rajah 4.4 memaparkan bilangan setiap genre mengikut jenis filem dan rancangan TV.

Jadual 4.2 Bilangan dan peratusan setiap genre mengikut jenis filem dan rancangan TV.

Jenis	**Genre**	**Bilangan**	**Peratusan(%)**
Filem	Aksi	6	1.99
	Animasi	2	0.66
	Komedi	36	11.92
	Jenayah	13	4.30
	Drama	69	22.85
	Fantasi	12	3.97
	Seram	20	6.62
	Percintaan	9	2.98
	Scifi	8	2.65
	Thriller	48	15.89
	Barat	5	1.66
Rancangan TV	Aksi	11	3.64
	Komedi	11	3.64
	Jenayah	6	1.99
	Dokumentasi	1	0.33
	Drama	19	6.29
	Scifi	21	6.95
	Thriller	3	0.99
	Peperangan	1	0.33
	Barat	1	0.33



Rajah 4.4 Graf perbandingan genre mengikut jenis filem dan rancangan TV

Berdasarkan Jadual 4.2 dan Rajah 4.4, genre Drama mendominasi jenis filem sebanyak 22.85% manakala genre Scifi mendominasi jenis rancangan TV sebanyak 6.95%. Walau bagaimanapun genre drama bagi jenis rancangan TV turut mencatatkan peratusan kedua tinggi sebanyak 6.29% dengan perbezaan yang kecil sebanyak 0.66%. Oleh itu, dapat disimpulkan penonton lebih berminat menonton genre drama bagi kedua-dua jenis taburan filem dan rancangan TV.

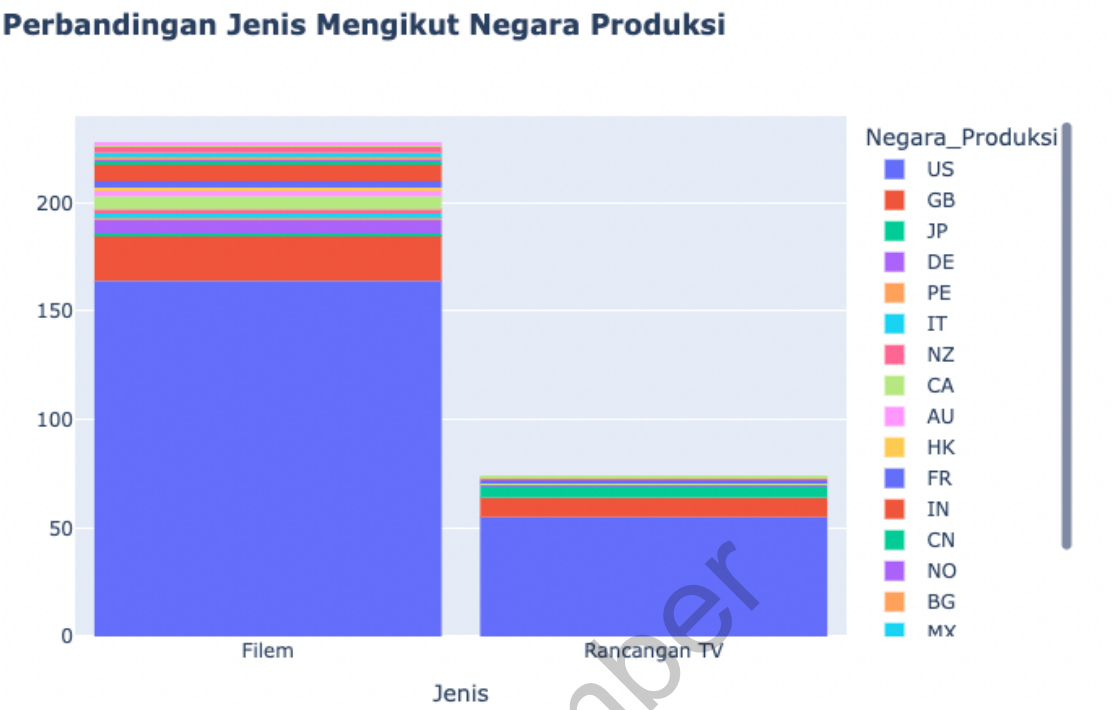
d. Perbandingan Jenis Mengikut Negara Produksi

Perbandingan negara produksi ialah perbandingan terakhir untuk atribut jenis. Jadual 4.3 memaparkan bilangan dan peratusan setiap negara produksi mengikut jenis filem mahupun rancangan TV.

Jadual 4.3 Bilangan dan peratusan setiap negara produksi mengikut jenis filem mahupun rancangan TV.

Jenis	**Negara Produksi**	**Bilangan**	**Peratusan(%)**
Filem	AU	3	0.99
	BE	1	0.33
	BG	1	0.33
	CA	6	1.99
	CN	1	0.33
	DE	6	1.99
	ES	3	0.99
	FR	3	0.99
	GB	21	6.95
	HK	1	0.33
	IN	8	2.65
	IT	2	0.66
	JP	1	0.33
	KR	1	0.33
	MX	2	0.66
	NO	1	0.33
	NZ	2	0.66
	PE	1	0.33
	US	164	54.30
Rancangan TV	CA	1	0.33
	DE	1	0.33
	ES	1	0.33
	FR	1	0.33
	GB	9	2.98
	JP	5	1.66
	KR	1	0.33
	US	55	18.21

Rajah 4.5 menunjukkan bilangan setiap negara produksi mengikut jenis filem dan rancangan TV.



Rajah 4.5 Graf perbandingan negara produksi mengikut jenis filem dan rancangan TV

Berdasarkan Jadual 4.3 dan Rajah 4.5, produksi negara Amerika Syarikat (US) mendominasi sebanyak 54.30% bagi filem manakala 18.21% bagi rancangan TV. Oleh yang demikian, walaupun filem dan rancangan TV di Netflix adalah daripada pelbagai negara, filem dan rancangan TV dari negara produksi US adalah yang paling banyak ditayangkan dan ditonton di platform tersebut. Penerangan bagi negara produksi yang lain dilampirkan dalam Lampiran D.

4.2.2 Kelas Label Skor IMDb

Skor IMDb telah dipilih sebagai kelas label untuk kajian ini. *Exploratory data analysis* (EDA) telah dilakukan sebelum proses transfomasi data skor IMDb daripada jenis data *float* ke nominal.

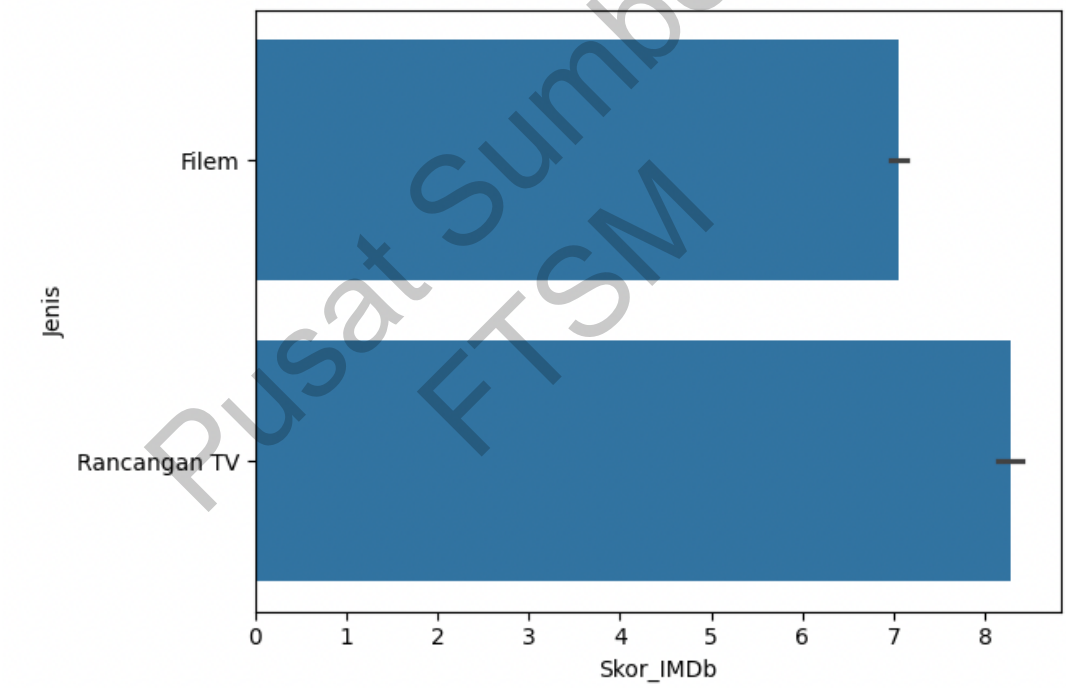
a. EDA Atribut Skor IMDb Mengikut Jenis

EDA atribut jenis adalah untuk melihat skor IMDb yang paling banyak diperoleh oleh kedua-dua filem dan rancangan TV. Jadual 4.4 memaparkan bilangan dan peratusan skor IMDb bagi filem dan rancangan TV.

Jadual 4.4 Bilangan dan peratusan skor IMDb mengikut jenis filem mahupun rancangan TV.

Jenis	Skor IMDb	Bilangan	Peratusan (%)
Filem	4.0-4.9	1	0.3
	5.0-5.9	19	6.29
	6.0-6.9	78	25.83
	7.0-7.9	98	32.45
	8.0-8.9	32	10.60
Rancangan TV	6.0-6.9	3	0.99
	7.0-7.9	16	5.30
	8.0-8.9	49	16.23
	9.0-10.0	6	1.99

Rajah 4.6 memaparkan graf skor IMDb mengikut jenis filem dan rancangan TV.



Rajah 4.6 Graf skor IMDb mengikut jenis filem mahupun rancangan TV

Berdasarkan Jadual 4.4 dan Rajah 4.6, kebanyakan filem memperoleh skor IMDb antara 7.0-7.9 dan rancangan TV pula 8.0-8.9. Peratusan bagi filem ialah 32.45% dan rancangan TV pula 16.23%. Walaupun rancangan TV mendapat rating yang tinggi, peratusannya rendah kerana bilangan rancangan TV dalam set data ini lebih rendah daripada filem.

b. EDA Atribut Skor IMDb Mengikut Genre

EDA atribut genre adalah untuk melihat skor IMDb yang paling banyak diperoleh oleh kedua-dua filem dan rancangan TV. Jadual 4.5 memaparkan bilangan dan peratusan skor IMDb bagi genre.

Jadual 4.5 Bilangan dan peratusan skor IMDb mengikut genre.

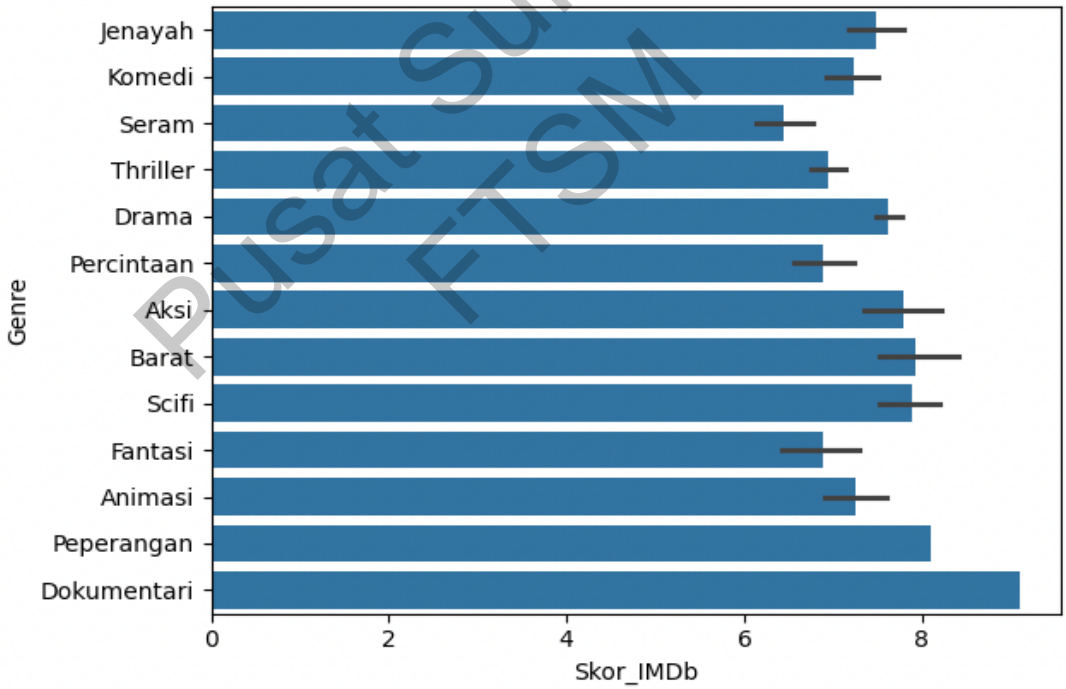
Genre	Skor IMDb	Bilangan	Peratusan(%)
Aksi	8.0-8.9	8	2.65
	7.0-7.9	4	1.32
	6.0-6.9	3	0.99
	5.0-5.9	1	0.33
	9.0-10.0	1	0.33
Animasi	7.0-7.9	1	0.33
	6.0-6.9	1	0.33
Komedi	8.0-8.9	16	5.30
	7.0-7.9	12	3.97
	6.0-6.9	15	4.97
	5.0-5.9	4	1.32
Jenayah	8.0-8.9	5	1.66
	7.0-7.9	9	2.98
	6.0-6.9	5	1.66
Dokumentasi	9.0-10.0	1	0.33
Drama	8.0-8.9	28	9.27
	7.0-7.9	44	14.57
	6.0-6.9	15	4.97
	9.0-10.0	1	0.33
Fantasi	8.0-8.9	1	0.33
	7.0-7.9	4	1.32
	6.0-6.9	6	1.99
	5.0-5.9	1	0.33
Seram	8.0-8.9	1	0.33
	7.0-7.9	5	1.66
	6.0-6.9	7	2.32
	5.0-5.9	7	2.32
Percintaan	7.0-7.9	4	1.32
	6.0-6.9	4	1.32
	5.0-5.9	1	0.33

bersambung…

…sambungan

Scifi	8.0-8.9	13	4.30
	7.0-7.9	8	2.65
	6.0-6.9	3	0.99
	5.0-5.9	2	0.66
	9.0-10.0	3	0.99
Thriller	8.0-8.9	6	1.99
	7.0-7.9	19	6.29
	6.0-6.9	22	7.28
	4.0-4.9	1	0.33
	5.0-5.9	3	0.99
Peperangan	8.0-8.9	1	0.33
Barat	8.0-8.9	2	0.66
	7.0-7.9	4	1.32

Rajah 4.7 memaparkan graf bagi skor IMDb yang paling banyak diperoleh bagi atribut genre.

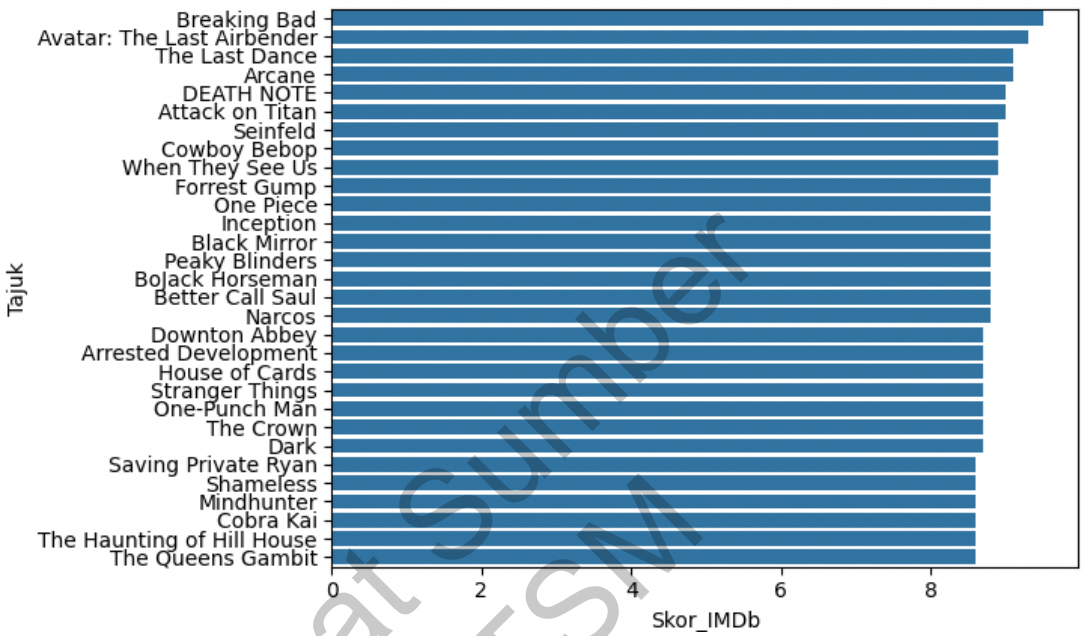


Rajah 4.7 Graf skor IMDb mengikut genre

Berdasarkan Jadual 4.5 dan Rajah 4.7, genre yang memperoleh skor tertinggi iaitu 9.0-10.0 ialah dokumentasi dengan peratusan sebanyak 0.33% manakala genre yang memperoleh skor terendah ialah *thriller* iaitu 4.0-4.9 dengan peratusan sebanyak 0.33%.

c. EDA Atribut Skor IMDb Mengikut Tajuk

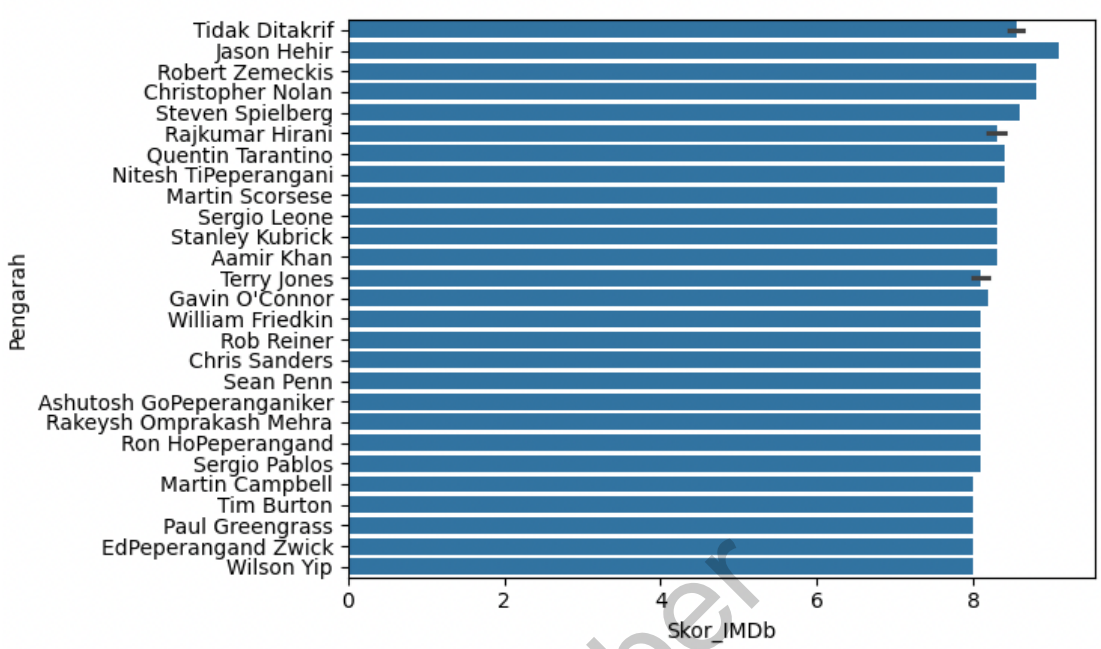
Rajah 4.8 memaparkan graf filem atau rancangan TV yang memperoleh skor IMDb melebihi 8.0. Berdasarkan graf tersebut, rancangan TV *Breaking Bad* memperoleh skor IMDb yang tertinggi dengan memperoleh skor melebihi 9.5.



Rajah 4.8 Graf skor IMDb mengikut tajuk

d. EDA Atribut Skor IMDb Mengikut Pengarah

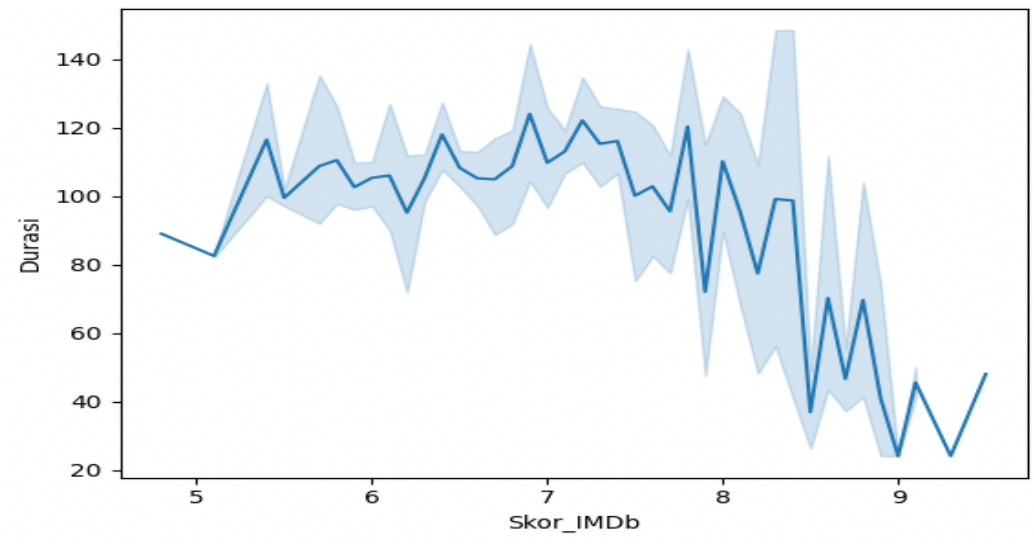
Rajah 4.9 memaparkan graf pengarah filem atau rancangan TV yang memperoleh skor IMDb melebihi 8.0. Berdasarkan graf tersebut, pengarah Jason Hehir, pengarah daripada rancangan *The Last Dance* telah memperoleh skor IMDb yang tertinggi dengan memperoleh skor 9.1. Walaupun terdapat filem atau rancangan lain yang memperoleh skor IMDb yang lebih tinggi, namun maklumat nama pengarah tidak terdapat dalam set data.



Rajah 4.9 Graf skor IMDb mengikut pengarah

e. EDA Atribut Skor IMDb Mengikut Durasi

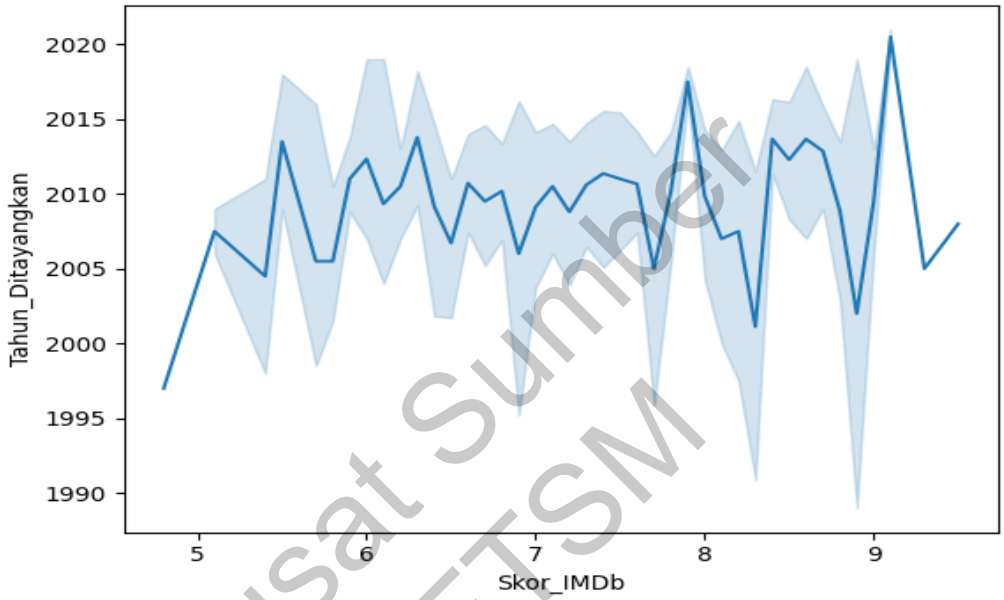
Rajah 4.10 memaparkan perbandingan graf garisan durasi filem atau rancangan TV dengan skor IMDb. Berdasarkan graf tersebut, rata-rata filem atau rancangan TV memperoleh skor yang pelbagai, namun filem atau rancangan TV yang berdurasi kurang 100 minit sering memperoleh keputusan IMDb skor melebihi 8.0.



Rajah 4.10 Perbandingan graf garisan skor IMDb mengikut durasi

f. EDA Atribut Skor IMDb Mengikut Tahun Ditayangkan

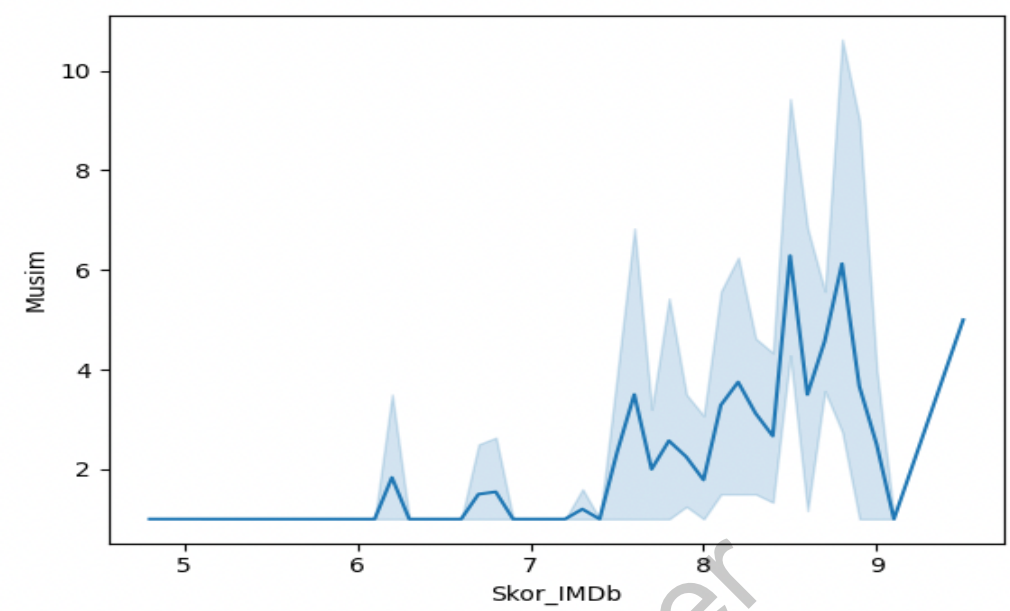
Rajah 4.11 memaparkan perbandingan graf garisan tahun filem atau rancangan TV yang ditayangkan dengan skor IMDb. Berdasarkan graf tersebut, filem atau rancangan TV yang ditayangkan pada tahun 2005 ke atas memperoleh skor IMDb melebihi 5.0. Filem atau rancangan TV yang ditayangkan pada tahun 2000 dan ke bawah pula memperoleh skor kurang daripada 5.0.



Rajah 4.11 Perbandingan graf garisan skor IMDb mengikut tahun ditayangkan

g. EDA Atribut Skor IMDb Mengikut Musim

Rajah 4.12 memaparkan perbandingan graf garisan bilangan musim filem atau rancangan TV ditayangkan dengan skor IMDb. Berdasarkan graf tersebut, filem atau rancangan TV yang melebihi empat musim mempunyai skor melebihi 7.0 berbanding filem atau rancangan TV yang kurang dari empat musim.



Rajah 4.12 Perbandingan graf garisan skor IMDb mengikut bilangan musim

4.3 KEPUTUSAN UJI KAJI MODEL PEMBELAJARAN MESIN

Analisis deskriptif dan EDA merupakan langkah asas sebelum perlombongan model. Tujuan analisis deksriptif dan EDA adalah untuk mengenal pasti sifat sesuatu atribut serta hubungkait antara satu atribut dengan atribut yang lain mahupun hubungkait antara sesuatu atribut dengan kelas atribut dalam kajian. Konklusi awal boleh dibuat melalui analisis deskriptif mahupun EDA. Namun untuk memperkukuhkan lagi kajian, pemodelan data menggunakan algoritma pembelajaran mesin perlu dilaksana. Sub topik seterusnya membincangkan keputusan hasil pembangunan model serta prestasi setiap model yang dibangun.

4.4 HASIL PEMBANGUNAN MODEL DAN PRESTASI MODEL

Bab III menerangkan proses pemilihan ciri yang dibuat sebelum pembangunan model. Melalui Jadual 3.9, atribut Skor IMDb mempunyai *correlation coefficient* yang paling tinggi dengan atribut kelas iaitu dengan nilai 1.0. Oleh yang demikian, atribut ini dibuang sebelum proses pemodelan bagi mengelakkan sebarang keputusan yang meragukan. Hanya atribut yang memiliki *correlation coefficient* kurang daripada 0.9 dikekalkan untuk tujuan pemodelan. Jadual 4.6 menunjukkan atribut yang dikekalkan untuk pembangunan model.