SELEKSI FITUR UNTUK PREDIKSI RATING FILM HOLLYWOOD MENGGUNAKAN MODEL K-NEAREST NEIGHBOR

Andi Bode
Fakultas Ilmu Komputer Universitas Ichsan Gorontalo
Jl. Achmad Nadjamuddin No. 17 Telp. 0435-829975 Fax. 0435-829976 Kota Gorontalo
andibode22@gmail.com

ABSTRAK

Film kali pertamanya dipertontonkan untuk masyarakat dengan membayar di *grand cafe boulevard de capucines, paris,*perancis 28 Desember 1895. Saat ini perkembangan dunia perfilman meningkat dengan pesat, dengan banyaknya film-film yang silih berganti di tayangkan dengan melihatjumlah film terhitung mencapai ribuan, sehingga menimbulkan kesulitan bagi masyarakat penikmat film, terkadang masyarakat hanya melihat poster atau trailer pada film sehingga ketika menonton film fullnya tidak sesuai ekspetasi. Keadaan ini membuat rating dari film tersebut turun dan hasilnya membuat produser mengalami kerugian biaya dalam memproduksi film. Perlu dilakukan prediksi pada *rating* film, dengan melihat *rating* film bisa menentukan apakah suatu film menerik untuk di tonton atau tidak. Pada penelitian ini medel yang di gunakan untuk memprediksi adalah K-*Nearest Neighbor* model ini dapat menghasilkan data yang efektif dan memiliki konsistensi yang kuat. Untuk meningkatkan hasil *accuracy* model K-NN perlu ditambahkan seleksi fitur*Forward Selection* dan *Backward Elimination*. Dengan seleksi fitur tingkat *accuracy* yang dihasilkan meningkat secara signifikan yaitu 98,92%, *recaal* 98,00% dan *precision* 100,00%.

Kata Kunci:***Film, K-Nearest Neighbor, Forward Selection, Backward Elimination***

Abstract

The first time the film was shown to the public by paying at the grand cafe boulevard de capucines, Paris, France December 28, 1895. At present the development of the world of cinema is increasing rapidly, with the number of films that are alternately aired by thousands of films so causing difficulties for the people who enjoy the film, sometimes people only see posters or trailers on the film so that when watching the full film does not match expectations. This situation made the rating of the film go down and the result made the producer experience a cost loss in producing films. It is necessary to make predictions on the rating of the film, by looking at the rating of the film can determine whether a movie is to be watched or not. In this study the medel used to predict is this K-Nearest Neighbor model can produce data that is effective and has strong consistency. To improve the accuracy of the K-NN model, a selection of Forward Selection and Backward Elimination features should be added. With feature selection, the level of accuracy produced increases significantly, namely 98.92%, recalal 98.00% and precision 100.00%.

Keyword: Films, K-Nearest Neighbor Forward Selection, Backward Eliminatio

I. PENDAHULUAN

Beberapa bentu seni meliputi seni sastra, seni rupa, seni musik, seni tari, ataupun seni peran memiliki penggemar masing-masing. Karya seni mengalami perkembangan tahun ke tahun sehingga tercipta perpaduan antara seni musik, seni peran, seni sastra yang di kemas dalam bentuk film. Film merupakan sarana yang digunakan untuk menyebarkan sebuah hiburan kepada masyarakat luas [1]. Film kali pertamanya dipertontonkan ke masyarakat dengan membayar di *grand cafe boulevard de capucines, paris,* perancis 28 Desember 1895. Kejadian tersebut sekaligus disebut lahirnya film dan bioskop didunia. Saat ini perkembangan dunia film sangat melonjak, dengan melihat semakin banyak film-film silih berganti untuk di tayangkan. Jumlah film tidak diragukan lagi terhitung mencapai ribuan,sehingga menimbulkan kesulitan bagi masyarakat penikmat film, sering kalinya masyarakat penikmat film minat dengan film yang ingin mereka lihat hanya dengan melihat poster serta trailer film tersebut,hasilnya ketika menonton filmnya ekspetasinya tidak sesuaibahkan film tersebut tidak menyajikan alur cerita yang menarik. Keadaan ini menimbulkan rating dari film yang ada menjadi menurun dan sehingga berdampak pada produser yang mengalami kerugian atas biaya dalam proses produksi film.

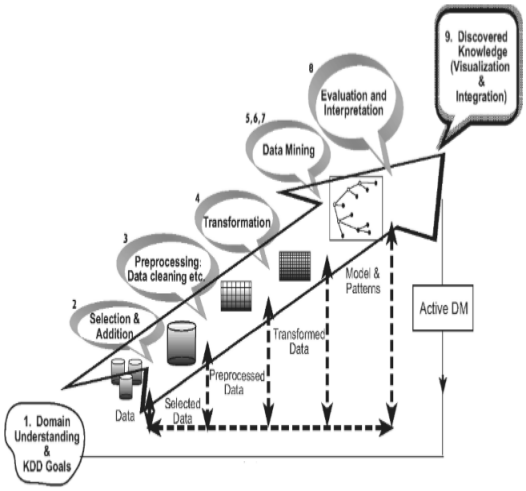
Prediksi pada rating film sangatlah diperlukan, rating dapat menentukan apakah suatu film menerik atau tidak untuk ditonton. Rating menentukan nilai tingkat peminat film tersebut bagi masyarakat. Kemudian produser dapat melihat rating pada produksi film yg dibuat jika tinggi, maka produser dapat membuat sequel pada film, alur cerita dibuat lebih menarik, meningkatkan kualitas gambar untuk meningkatkan laba pendatapatan dari film yang dibuat. Metode yang sering digunakan dalam prediksi adalah model K-*Nearest Neighbor* (KNN). Prediksi merupakan perkiraan nilai masa yang akan datang. K-NN merupakan salah satu model yang digunakan untuk analisis klasifikasi terhadap objek yang mempunyai jarak terdekat (ecluidean), tetapi sudah beberapa dekade terakhir model K-NN dapat digunakan untuk prediksi [2]. Model K-NN juga merupakan metodel non-parametrik yang digunakan untuk data klasifikasi dan regresi. Model K-NN menghitung antara jarak data dari semua sampel. Jarak yang sering diguanakan adalah jarak ecluident. Model K-NN merupakan model penerapan dari algoritma *supervised,* yang mana algoritma tersebut dibagi menjadi dua bagian yaitu *supervised leraning* dan *unsupervised learning*, model ini memiliki konsistensi yang kuat dan memiliki hasil data yang efektif [3].Menurut penelitian terdahulu yang dilakukan oleh Riszki W. P dan Yusuf S. N dengan judul Prediksi *Rating* Film Menggunakan Metode *Naïve Bayes* menghasilkan nilai *accuracy* 55,80%, *precision* 32,41%, serta *recall* 46,70% [4]. Seleksi fitur diterapkan untuk meninggkatkan hasil *accuracy* dari performa model K-NN. Fitur seleksi digunakan untuk memastikan bahwa setiap variable relevan.

Pada penelitian ini data yang digunakan adalah kaggle dataset Movie Metadata berasal dengan jumlah record 200 dan 6 variabel yang terdiri dari atribut *color, director name, num critic for review, duration, director facebook like, actor 3 facebook like.*Penerapan seleksi fitur dan modelK-*Nearest Neighbour* untuk prediksi Rating Film *Hollywood* dengan tujuan untuk dapat mengetahui minat penikmat film dari cerita, kualitas gambar, aktor, atau durasi film tersebut.

II. METODE PENELITIAN

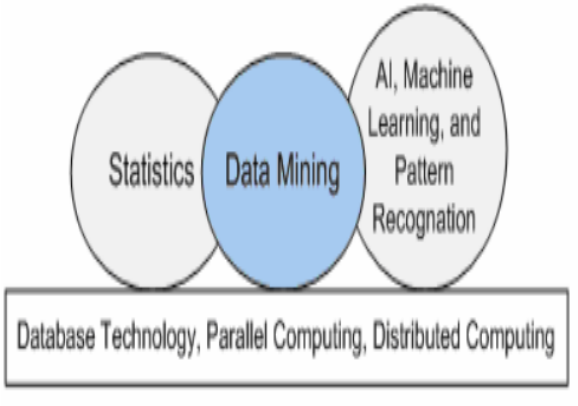
1. Data Mining

Data mining adalah suatu proses menambang (mining) pengetahuan sekumpulan data yang sangat besar. Data mining merupakan suatu langkah dalam *knowledge discovery database* (KDD).KDD) adalah seluruh proses *non-trivial* untuk mencari dan mengidentifikasi pola *(pattern)* didalam data, kemudian dapat bermanfaat dan dimengerti[5] (M. Rudyanto Ariel, Kusrini, Ricky ImanuelNdaumanu, 2014 :1).



Gambar 1. Proses KDD

Data mining dimanfaatkan untuk menelusuri data untuk membangun sebuah model, kemudian model tersebut digunakan untuk mengenali pola data yang lain yang tidak berada dalam basis data. Teknik *data mining* merupakan teknik klasifikasi, dimana klasifikasi adalah teknik pembelajaran dalam prediksi suatu nilai dari target variabel kategori.Data mining dapat diklasifikasikan kedalam dua kategori yaitu deskriptif dan prediktif [6] (Witten, I. H. dkk, 2011).Hasil dari *data mining* secara umum diintegrasikan dengan *decision support system (DSS).* Integrasi memerlukan langkah postprocessing sebagai penjamin bahwa hanya hasil yang *valid* digabungkan dengan DSS. Ukuran-ukuran statistik dan metode pengujian hipotesis digunakan saat postprocessing untuk mengeliminasi hasil *data mining* yang palsu.



Gambar 2. Data mining sebagai pertemuan dari banyak disiplin ilmu

2. K-Nearest Neighbor

Model K-*Nearest Neighbor* (K-NN) model yang menghitung jarak data uji terhadap setiap data latih kemudian mengambil ***k*** tetangga terdekat awal data. Kemudian, algoritma ini akan total jumlah data yang mengikuti kelas yang ada dari ***k*** tetangga. Dimana kelas yang dinyatakan dengan data terbanyak pengikutnynya menjadi kelas pemenang yang dijadikan sebagai kelas pada data uji ***y'***.Persamaan matematika K-NN adalah sebagai berikut:

$$y' = \underset{v}{argmax} \sum_{i=1}^{n} x_i y_i \in D_z I(v = y_i)$$

v adalah data yang masuk dalam kelas y_i.

Yang menjadi permasalah K-NN adalah pemilihan nilai ***k*** yang tepat. Cara pemilihan mayoritas dari ***k*** tetangga untuk nilai ***k*** yang besar dapat mengakibatkan distorsi data yang besar, karena setiap tetangga mempunyai bobot sama terhadap data uji, sedangkan ***k*** yang terlalu kecil dapat

menyebabkan *niose*.Cara mengatasi masalah tersebut, penambahan terhadapbobot untuk menghitung calon kelas yang sebaiknya diambil oleh data uji ***k***tetangga terdekat. Persamaan perhitungan bobot dari setiap tetangga terdekat dapat dilihat dibawah ini:

$$W_i = \frac{1}{d(x', x_i)^2}$$

Contoh:

$\mathbf{z} = (x', y')$adalah data uji dengan vektor x' dan class y' yang belum diketahui.Hitung jarak $\mathbf{d} = (x', \mathbf{x})$, jarak antara data uji **z**ke setiap vektor data latih.

Prediksi:

$$\mathbf{y}' = \mathbf{argmax} \sum_{i=1}^{n} (x_i y_i) \in d_z \, w_i I(v = y_i)$$

Dekat jauhnya tetangga biasanya dihitung berdasarkan:*Euclidean Distance* = $\sum (X_j - Y_j)^2$; atau *Manhattan Distance (cityblock atau Sum of Absolut Distance)* = $\sum ABS(X_j - Y_j)$*Cosine;* atau*Correlation;* atau*Hamming*[7] (Prasetyo, 2012)

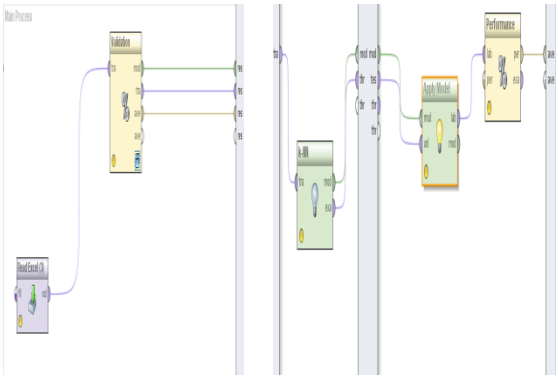
3. Seleksi Fitur

Seleksi Fitur merupakan permasalahan yang berterkaitan erat dengan pengurangan variabel. Tujuan dari seleksi fitur yaitu untuk mengidentifikasi fitur atau variabeldidalam kumpulan data yang sama pengaruhnya, kemudianmengeliminasi semua fitur atau informasi yang diangap tidak relevan dan kurang efektif. Tujuan utama dari seleksi fitur adalah mengurangi dimensi dari data, sehingga menghasilkan performa yang lebih efektif serta algoritma data mining yang lebih cepat (yaitu data algoritma Mining lebih cepat cepat dioperasikan dan lebih efektif dengan penerapan seleksi fitur)Seleksi fitur dilakukan untuk dapat mengidentifikasi dan mencari variable yang diangap relevan. Kemudian seleksi fiturberfungsi untuk pengurangan fitur, untuk mengeliminasi dari dataset dari variabel yang idak relevan.*Backward Elimination* konvensional menurun adalah proses mengeliminasi sebuah model regressor *one by one* berdasarkan penurunan didalam sebuah model yang baik. Ada beberapa alur model *backward elimination*: Mulailah dengan semua prediktor dalam model, Hapus prediktor dengan p-*value* tertinggi atau lebih besar dari target, Perbaiki model dan ulangi langkah ke – 2 dan Hentikan ketika semua jika p-value kurang dari target [8].Metode Sequential *Forward Selection* atau dikatakan metode seleksi maju merupakan algoritma pencarian sangat sederhana. *Forward Selection*berdasar pada model *regresi linear*. *Forward Selection* adalah salah satu cara yang digunakan untuk mereduksi datasetberdimensi tinggi dengan mengeluarkan atribut-atribut yang dianggap tidak relevan atau redudan. Metode *Forward Selection* adalah model diawali dari nol peubah (*empty model*), selanjutnya satu persatu peubah dimasukan sampai dengan kriteria tertentu dipenuhi [9].

III. HASIL DAN PEMBAHAN

Pada penelitian saat ini penerapan seleksi fitur *Forward Selection* dan *Backward Elimination* pada model K-*Nearest Neighbor* untuk prediksi Rating Film *Hollywood* dengan tujuan mendapatkan model *accuracy* tertinggi. Pada penelitian ini mengimplementasikan *software RapidMiner*, fungsi dari*software* ini adalah untuk mengimpor sebuah informasi dari berbagai macam sumber *database* kemudian diperiksa dan dianalisa didalam sebuah aplikasi. *RapidMiner* juga bisa disebut sebagai solusi untuk prediksi dan analisa komputasi statistic. Bisa dilihat pada gambar-gambar dibawah ini merupakan hasil dari penelitian yang dilakukan menggunakan *software*

RapidMiner.



Gambar 3. Klasifikasi K-*Nearest Neighbor*

accuracy: 98.92% +/- 2.16% (mikro: 98.90%)

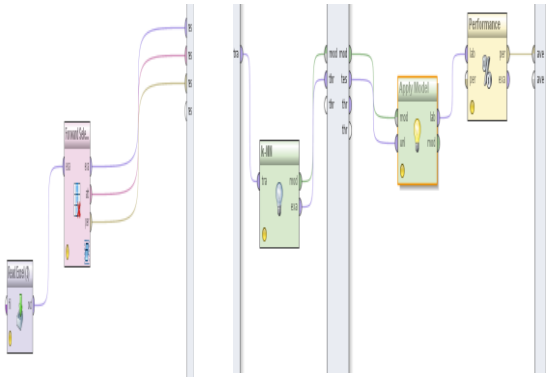
	true Bagus	true Cukup	class precision
pred. Bagus	84	2	97.67%
pred. Cukup	0	95	100.00%
class recall	100.00%	97.94%	

Gambar 4. Tingkat *Accuracy* pada K-*Nearest Neighbor*

Tabel.1 Nilai *Accuracy* Pengujian Menggunakan Model K-*Nearest Neighbor*

Nilai *K*	1	3	5	7	9
Accuracy	62,50 %	65,81 %	65,21 %	64,59 %	65,74 %
Recall	60,11 %	64,44 %	66,56 %	69,59 %	71,44 %
Precision	66,76 %	71,44 %	70,63 %	69,54 %	68,59 %

Gambar 4 menjelaskan model K-NN nilai *class* "CUKUP" 69,66%, sedangkan *class* "BAGUS" 61,96%. Pada table 1 menjelaskan ringkasan dari hasil eksperimen Rating Film *Hollywood* model K-NN dengan nilai ***k*** 1, 3, 5, 7, 9 yang dilihat pada nilai *accuracy* tertinggi. Nilai *accuracy* tertinggi adalah 65,81%, *recall* 64,44% dan *precision* 71,44% pada nilai ***k*** 3.



Gambar 5. Klasifikasi K-NN dengan *Forward Selection*

precision: 71.44% +/- 15.42% (mikro: 69.66%) (positive class: Cukup)

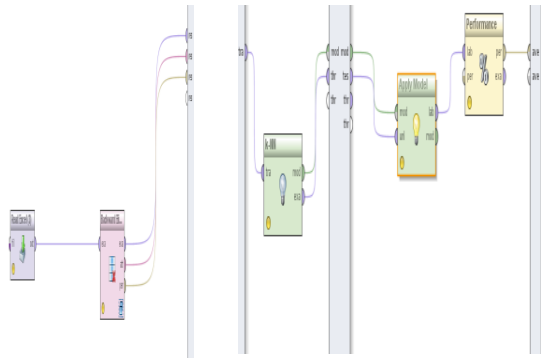
	true Bagus	true Cukup	class precision
pred. Bagus	57	35	61.96%
pred. Cukup	27	62	69.66%
class recall	67.86%	63.92%	

Forward Selection

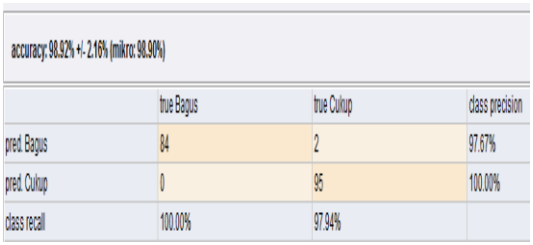
Tabel.2 Nilai *Accuracy* Pengujian Menggunakan Model K-NN dengan *Forward Selection*

Nilai *K*	1	3	5	7	9
Accuracy	97,78 %	98,89 %	98,89 %	98,89 %	98,92 %
Recall	97,89 %	98,00 %	98,00 %	98,00 %	98,00 %
Precision	98,18 %	100,00 %	100,00 %	100,00 %	100,00 %

Gambar 6 menjelaskan model K-NN dengan *Forward Selection*nilai *class* "CUKUP" 100,00%, sedangkan *class* "BAGUS" 97,67%. Pada tabel 2 menjelaskan ringkasan dari hasil eksperimen Rating Film *Hollywood* model K-NN dengan*Forward Selection* pada nilai ***k*** 1, 3, 5, 7, 9 yang dilihat pada nilai *accuracy* tertinggi. Nilai *accuracy* tertinggi adalah 98,92%, *recall* 98,00% dan *precision* 100,00% pada nilai ***k*** 9.



Gambar 7. Klasifikasi K-NN dengan *Backward Elimination*



accuracy: 98.92% +/- 2.16% (mikro: 98.90%)

	true Bagus	true Cukup	class precision
pred. Bagus	84	2	97.67%
pred. Cukup	0	95	100.00%
class recall	100.00%	97.94%	

Gambar 8. Tingkat *Accuracy* K-NN dengan *Backward Elimination*

Tabel.3 Nilai *Accuracy* Pengujian Menggunakan Model K-NN dengan *Backward Elimination*

Nilai *K*	**1**	**3**	**5**	**7**	**9**
Accuracy	97,84 %	69,73 %	72,44 %	98,92 %	98,92 %
Recall	97,00 %	66,78 %	72,33 %	98,00 %	98,00 %
Precision	99,09 %	74,61 %	75,76 %	100,00 %	100,00 %

Gambar 8 menjelaskan model K-NN denga *Backward Elimination* nilai *class* "CUKUP" 100,00%, sedangkan *class* "BAGUS" 97,67%. Pada table 3 menjelaskan ringkasan dari hasil eksperimen *Rating Film Hollywood* model K-NN dengan *Backward Elimination* pada nilai *k* 1, 3, 5, 7, 9 yang dilihat pada nilai *accuracy* tertinggi. Nilai *accuracy* tertinggi terjadi pada nilai *k* 7 dan nilai *k* 9, dengan hasil nilai *accuracy* sama yaitu 98,92%, *recaal* 98,00% dan *precision* 100,00%.

IV. KESIMPULAN DAN SARAN

Dari hasil eksperimen pada penelitian ini penerapan model K-*Nearest Neighbor* (K-NN) dengan fitur seleksi dalam prediksi *Rating Film Hollywood* yang dilakukan terdiri dari***k*** 1, 3, 5, 7, 9, bahwa penambahan fitur seleksi pada model K-NN menghasilkan nilai accuracy lebih tinggi dibandingkan dengan penerapan model K-NN tanpa fitur seleksi. Berdasarkan eksperimen menunjukkan perbedaan yang begitu signifikan, hasil K-NN nilai *accuracy* yaitu 65,81%, *recall* 64,44% dan *precision*71,44% dengan ***k*** 3, K-NN dan *Forward Selection* nilai *accuracy* 98,92%, *recall* 98,00% dan *precision* 100,00% dengan ***k*** 9, dan K-NN dan *Backward Elimination*nilai *accuracy* tertinggi terjadi pada nilai ***k*** 7 dan nilai ***k*** 9, dengan hasil nilai *accuracy* sama yaitu 98,92%, *recall* 98,00% dan *precision* 100,00%. Sehingga dapat ditarik kesimpulan penerapan model K-*Nearest Neighbor* dengan fitur seleksi dapat meninggkatkan akurasi dalam memprediksi Rating Film *Hollywood.*

Disarankan bagi peneliti selanjutnya untuk dapat menggunakan model algoritma lainnya, untuk dapat membandingkan tingkat keakurasian hasil prediksi Rating Film *Hollywood*.

DAFTAR PUSTAKA

[1]. Mudjiono Y. 2011, Kajian Semiotika Dalam Film, Jurnal Ilmu Komunikasi, Volume 1 Nomor 1, April, ISSN 2088-981X, 125-138.

[2]. Bode Andi. 2017. *K-Nearest Neighbor* Dengan *Feature Selection* Menggunakan *Backward Elimination* Untuk Prediksi Harga Komoditi Kopi Arabika, Universitas Ichsan Gorontalo, *ILKOM Jurnal Ilmiah*, Vol. 9, No. 2, Agustus, ISSN: 2087-1716.

[3]. Drajana Ivo C. R, 2018, Prediksi Jumlah Produksi *Coconut Oil* Menggunakan k-*Nearest Neighbor* dan *Backward Elimination*, *Jurnal Tecnoscienza*, Volume 3 Nomor 1, Oktober.

[4].PratiwiRiszki W. dan Nugroho Yusuf S.2016, Prediksi *Rating* Film Menggunakan Metode *Naïve Bayes, Jurnal Teknik Elektro,* Vol. 8. N0. 2, Desember, ISSN 1411-0059

[5]. Ndaumanu Ricky I, Kusrini, Arief M. R, 2014, Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode *K-Nearest Neighbor,* JATISI*jurnal Teknik Informatika dan Sistem Informasi,* Vol. 1. No. 1, September, ISSN 2407-4322

[6]. Witten, I. H., Frank, E., Hall, M. A. 2011, *Data Mining Practical Machine Learning Tools and Techniquers* (3rd ed). USA: Elsevier

[7]. Prasetyo, Eko. 2012, *Data Mining* Konsep dan Aplikasi Menggunakan Matlab. Penerbit Andi Yogyakarta.

[8]. Yunita, 2017, Seleksi Fitur Menggunakan *Backward Elimination* Pada Prediksi Cuaca Dengan *Neural Network*, IJCIT (Indonesian Journal on Computer and Information Technology),