**(JANGAN DIKIRIM DULU HARUS DIPARAFRASE) Hukum Bilangan Besar pada Sistem Rekomendasi Film Menggunakan Probabilitas dan Statistika**

Naufal Priyamukti Wibowo dan 20230140097

Program Studi Teknologi Informasi

Fakultas Teknik

Universitas Muhammadiyah Yogyakarta,

Jl. Brawijaya, Tamantirto, Kasihan, Bantul 55183, Indonesia

[n.priyamukti.ft23@mail.umy.ac.id](mailto:n.priyamukti.ft23@mail.umy.ac.id)

**Abstrak** Hukum Bilangan Besar (Law of Large Numbers) merupakan salah satu teori probabilitas yang menjelaskan hubungan antara ukuran sampel dengan estimasi rata-rata populasi. Dalam era big data, hukum ini diaplikasikan pada berbagai bidang, termasuk teknologi informasi. Makalah ini mengkaji penerapan Hukum Bilangan Besar dalam sistem rekomendasi film berbasis data penilaian pengguna. Dengan memanfaatkan dataset penilaian film, model probabilistik dikembangkan untuk memprediksi preferensi pengguna baru. Implementasi menggunakan bahasa pemrograman Python disertai library statistika untuk simulasi dan analisis data. Hasil simulasi menunjukkan bahwa estimasi rata-rata penilaian mendekati nilai sebenarnya seiring bertambahnya data, yang memperkuat validitas Hukum Bilangan Besar dalam aplikasi teknologi informasi. Penelitian ini memberikan wawasan mendalam tentang cara probabilitas dan statistika dapat diterapkan untuk meningkatkan kualitas sistem rekomendasi.

**Kata Kunci** Hukum Bilangan Besar, Sistem Rekomendasi Film, Probabilitas, Statistika, Python

**I.Pendahuluan**

Perkembangan teknologi informasi telah membawa transformasi besar dalam cara manusia memanfaatkan data. Salah satu penerapan utama dari teknologi ini adalah sistem rekomendasi, yang telah menjadi bagian tak terpisahkan dari berbagai platform digital seperti Netflix, YouTube, Amazon, dan Spotify. Sistem rekomendasi berfungsi sebagai alat bantu untuk menyaring informasi yang relevan dan memberikan pengalaman pengguna yang lebih personal. Dalam upayanya, sistem ini bergantung pada analisis data besar (big data) dan teknik prediktif untuk mengidentifikasi preferensi pengguna.

Teknologi sistem rekomendasi menggunakan pendekatan berbasis data untuk membuat prediksi yang tepat. Salah satu pondasi teoretis yang mendukung keberhasilan pendekatan ini adalah probabilitas dan statistika, khususnya Hukum Bilangan Besar. Teori ini menjelaskan bagaimana, dengan meningkatnya jumlah data yang diamati, rata-rata sampel semakin mendekati rata-rata populasi yang sebenarnya. Dalam sistem rekomendasi, ini berarti bahwa semakin banyak data yang tersedia tentang preferensi pengguna, semakin akurat prediksi yang dapat dihasilkan oleh sistem.

Hukum Bilangan Besar memiliki relevansi yang sangat kuat dalam pemrosesan data besar. Di era modern, banyak perusahaan teknologi telah memanfaatkan konsep ini untuk mengoptimalkan layanan mereka. Dalam konteks sistem rekomendasi film, konsep ini memungkinkan sistem untuk memperkirakan preferensi pengguna baru dengan menggunakan informasi dari pengguna lain yang memiliki pola penilaian serupa. Dengan demikian, hukum probabilitas menjadi alat yang esensial untuk menangkap ketidakpastian dan

Pendekatan dalam makalah ini diawali dengan menjelaskan teori dasar probabilitas dan Hukum Bilangan Besar. Selanjutnya, dibahas metodologi untuk mengembangkan sistem rekomendasi berbasis data. Dataset MovieLens digunakan sebagai studi kasus untuk menguji keakuratan model yang dikembangkan. Dataset ini dipilih karena berisi data yang kaya tentang preferensi pengguna terhadap berbagai film, termasuk informasi seperti penilaian, waktu, dan karakteristik pengguna.

Hasil penelitian ini menunjukkan bahwa penggunaan teori probabilitas, seperti Hukum Bilangan Besar, dapat meningkatkan keakuratan sistem rekomendasi. Dengan memperbesar ukuran data yang digunakan untuk pelatihan model, prediksi rata-rata menjadi semakin mendekati nilai sebenarnya. Hal ini mendukung konsep bahwa data besar memberikan keuntungan signifikan dalam analisis statistik.

Lebih lanjut, sistem rekomendasi yang dirancang dengan pendekatan probabilistik memiliki fleksibilitas untuk menangani ketidakpastian. Ketika data tentang pengguna baru tidak cukup tersedia, sistem dapat menggunakan pola yang teramati dari populasi lain untuk memberikan rekomendasi awal. Seiring bertambahnya data pengguna baru, sistem terus memperbarui estimasi preferensinya, menghasilkan rekomendasi yang lebih personal dan relevan.

Sistem rekomendasi film juga dapat diperkuat dengan mempertimbangkan faktor-faktor tambahan seperti genre film, aktor, atau ulasan kritikus. Faktor-faktor ini dapat dimasukkan ke dalam model probabilistik sebagai variabel tambahan, memungkinkan sistem untuk membuat prediksi yang lebih kompleks dan mendalam. Dalam penelitian ini, fokus diberikan pada penerapan dasar Hukum Bilangan Besar untuk menunjukkan kekuatan pendekatan ini dalam memproses data besar dan menangkap pola-pola yang relevan.

Makalah ini diharapkan memberikan kontribusi signifikan dalam memahami cara probabilitas dan statistika dapat diterapkan dalam pengembangan teknologi informasi. Pendekatan berbasis probabilitas tidak hanya relevan dalam sistem rekomendasi, tetapi juga memiliki potensi untuk diterapkan dalam berbagai domain lain seperti analisis keuangan, pengolahan citra, dan sistem prediktif lainnya. Dengan mengintegrasikan teori probabilitas ke dalam teknologi modern, kita dapat menciptakan solusi yang lebih akurat, adaptif, dan relevan untuk tantangan dunia nyata.

Sebagai kesimpulan, pendahuluan ini menggarisbawahi pentingnya Hukum Bilangan Besar dalam mengembangkan sistem rekomendasi film berbasis data besar. Dengan teori probabilitas sebagai panduan, sistem ini mampu memberikan prediksi yang lebih baik, sekaligus membuka peluang baru untuk inovasi di berbagai bidang teknologi informasi.

**II. Teori dan Penerapan Probabilitas**

**A. Hukum Bilangan Besar (Law of Large Numbers)**

Hukum Bilangan Besar (LLN) merupakan prinsip dasar dalam probabilitas yang menyatakan bahwa semakin besar ukuran sampel, rata-rata sampel akan semakin mendekati nilai harapan atau rata-rata populasi. Hukum ini memainkan peran penting dalam berbagai aplikasi statistik dan probabilitas, termasuk pengambilan keputusan berbasis data.

Secara matematis, jika X₁, X₂, ..., Xₙ​ adalah variabel acak yang identik dan independen dengan rata-rata μ dan varians , maka rata-rata sampel, X̄ didefinisikan sebagai:

Berdasarkan Weak Law of Large Numbers (WLLN), untuk sembarang bilangan ϵ > 0, probabilitas bahwa rata-rata sampel X̄, menyimpang lebih dari ϵ dari rata-rata populasi μ akan mendekati nol seiring . Hal ini secara formal dinyatakan sebagai:

Ilustrasi matematis dari fenomena ini dapat dilihat pada grafik berikut:

Grafik tersebut menunjukkan bagaimana rata-rata sampel X̄ mendekati rata-rata populasi μ seiring bertambahnya jumlah sampel . Fluktuasi pada awal pengamatan secara bertahap menjadi semakin kecil.

**B. Model Probabilistik untuk Sistem Rekomendasi**

Dalam sistem rekomendasi, model probabilistik digunakan untuk menangkap ketidakpastian preferensi pengguna dan mengolah data historis. Salah satu pendekatan probabilistik yang paling sering digunakan adalah distribusi normal. Distribusi ini cocok karena dapat menggambarkan berbagai fenomena alami, termasuk perilaku pengguna.

Distribusi normal disebut juga sebagai

**distribusi Gauss** dan memiliki bentuk simetris seperti lonceng (bell curve). Fungsi kepadatan probabilitas (probability density function atau PDF) dari distribusi normal diberikan oleh:

di mana:

* x adalah nilai preferensi pengguna
* μ adalah rata-rata (mean) dari distribusi,
* adalah varians (tingkat penyebaran data),
* adalah simpangan baku (standard deviation),
* exp adalah fungsi eksponensial.

Fungsi ini menunjukkan probabilitas relatif suatu nilai (x) muncul berdasarkan jaraknya terhadap rata-rata μ dan tingkat penyebaran data Data yang lebih dekat dengan rata-rata memiliki probabilitas yang lebih tinggi, sedangkan data yang jauh dari rata-rata memiliki probabilitas yang lebih rendah.

Grafik tersebut menunjukkan histogram data penilaian pengguna (warna hijau), yang menggambarkan frekuensi nilai penilaian terhadap suatu film. Garis biru melambangkan distribusi normal teoretis yang sesuai dengan data tersebut. Distribusi ini memiliki puncak di sekitar rata-rata penilaian μ=3.5 menunjukkan bahwa sebagian besar pengguna memberikan penilaian mendekati nilai tersebut. Kurva distribusi juga menunjukkan bahwa nilai penilaian yang lebih jauh dari rata-rata memiliki probabilitas yang lebih rendah. Lebar kurva, yang ditentukan oleh simpangan baku (σ =1.0), mencerminkan tingkat penyebaran data penilaian. Dengan demikian, kurva ini memberikan wawasan tentang pola preferensi pengguna terhadap film tertentu.

**III. Buat Studi Kasus atau Masalah yang diangkatRekomendasi Film – Ini tanya**

**A.Latar Belakang Masalah**

Netflix adalah salah satu platform streaming terbesar di dunia. Platform ini telah menjadi standar dalam industri hiburan digital, menyediakan ribuan film, serial, dan dokumenter. Namun, keberhasilan Netflix tidak hanya terletak pada keberagaman kontennya, tetapi juga pada kemampuan sistemnya untuk memberikan pengalaman yang dipersonalisasi kepada pengguna. Setiap pengguna memiliki preferensi unik, dan sistem rekomendasi Netflix dirancang untuk memahami pola preferensi ini berdasarkan riwayat tontonan, penilaian, serta data perilaku lainnya.

Namun, dengan bertambahnya jumlah pengguna dan variasi konten, Netflix menghadapi tantangan besar dalam memastikan rekomendasi tetap relevan dan memuaskan. Setiap interaksi pengguna menghasilkan data baru, yang harus diolah secara efisien untuk meningkatkan akurasi sistem rekomendasi. Selain itu, pengembangan model prediksi yang akurat membutuhkan pendekatan matematis dan statistik yang canggih. Salah satu teori yang mendukung keberhasilan sistem ini adalah Hukum Bilangan Besar, yang memungkinkan sistem untuk memanfaatkan pola kolektif dalam data besar untuk memprediksi preferensi individu.

Masalah ini menjadi lebih kompleks ketika berhadapan dengan pengguna baru, yang belum memiliki riwayat interaksi yang memadai. Untuk itu, Netflix perlu memanfaatkan data dari pengguna lain yang memiliki karakteristik atau pola penilaian serupa untuk memberikan rekomendasi awal yang akurat.

**B.Permasalahan**

Salah satu tantangan utama dalam sistem rekomendasi Netflix adalah masalah **"cold start"**, yaitu kondisi di mana pengguna baru atau konten baru tidak memiliki data interaksi yang cukup untuk membuat prediksi yang akurat. Pada kasus pengguna baru, sistem harus membuat rekomendasi awal yang relevan berdasarkan data terbatas, seperti preferensi genre umum, lokasi geografis, atau pola kolektif dari kelompok pengguna serupa. Namun, tanpa riwayat tontonan yang spesifik, rekomendasi ini sering kali kurang relevan.

Masalah ini diperparah oleh skala besar data yang harus diolah oleh sistem Netflix. Dengan ribuan konten baru yang terus ditambahkan setiap bulan dan jutaan interaksi harian dari pengguna, sistem rekomendasi harus tetap efisien dan akurat dalam mengelola informasi. Selain itu, kebutuhan untuk menyajikan rekomendasi dalam waktu nyata (real-time) menuntut pendekatan komputasi yang cepat dan tepat.

Bagaimana sistem dapat mengatasi tantangan "cold start" ini sekaligus memanfaatkan Hukum Bilangan Besar untuk menghasilkan rekomendasi yang lebih baik seiring bertambahnya data pengguna? Inilah inti dari permasalahan yang dihadapi oleh Netflix dan menjadi fokus dari studi kasus ini.

**C.Pendekatan Solusi**

Untuk mengatasi tantangan ini, pendekatan yang digunakan mencakup tiga langkah utama yang terintegrasi:

**1. Pengumpulan Data**

Data adalah inti dari sistem rekomendasi. Dalam studi ini, dataset MovieLens digunakan untuk merepresentasikan preferensi pengguna terhadap ribuan film. Dataset ini mencakup jutaan penilaian pengguna, yang memberikan gambaran tentang pola perilaku pengguna berdasarkan rating yang mereka berikan. MovieLens dipilih karena keakuratannya dalam mencatat data interaksi pengguna, seperti waktu penilaian, genre yang diminati, hingga pola kebiasaan menonton. Dengan menggunakan dataset ini, sistem dapat memanfaatkan data populasi untuk mengatasi masalah "cold start" dengan memprediksi preferensi awal berdasarkan pola kolektif.

**2. Penggunaan Hukum Bilangan Besar**

Hukum Bilangan Besar diaplikasikan untuk memprediksi rata-rata preferensi populasi pengguna. Sistem menghitung rata-rata penilaian untuk kategori atau genre tertentu berdasarkan data historis dari pengguna dengan pola yang serupa. Dengan bertambahnya data, rata-rata sampel akan semakin mendekati rata-rata populasi sebenarnya, memungkinkan sistem untuk meningkatkan akurasi prediksi. Proses ini melibatkan simulasi untuk menunjukkan bagaimana rata-rata sampel konvergen ke nilai populasi seiring bertambahnya data.

**3. Implementasi dan Evaluasi**

Model probabilistik berbasis distribusi normal digunakan untuk merepresentasikan pola penilaian pengguna terhadap genre tertentu. Model ini membantu sistem memahami bagaimana data tersebar dan memungkinkan prediksi probabilitas untuk setiap kategori film. Evaluasi dilakukan dengan membandingkan akurasi rekomendasi awal (sedikit data) dengan akurasi setelah pengguna berinteraksi lebih banyak dengan sistem (lebih banyak data). Proses ini menunjukkan bagaimana pendekatan berbasis Hukum Bilangan Besar meningkatkan kualitas rekomendasi secara bertahap.

**D.Sumber** **Pengumpulan Data**

Untuk sumber data berasal dari dataset yang digunakan dalam makalah ini adalah MovieLens, yang dikelola oleh GroupLens Research dari University of Minnesota. MovieLens adalah salah satu dataset yang cukup lengkap di dunia terkait rekomendasi film, dengan mencakup ribuan film dan jutaan penilaian dari pengguna. Dataset ini sangat berharga karena memungkinkan analisis mendalam terhadap pola preferensi pengguna dan membantu membangun sistem rekomendasi yang lebih baik.

1. Deskripsi Dataset MovieLens

MovieLens mulai beroperasi pada tahun 1997 dan terus berkembang hingga saat ini, namun dataset MovieLens mencakup data yang dimulai dari tahun 1995. Dataset ini mencakup informasi tentang:

* Jumlah Film: Film yang dirilis dari berbagai genre, tahun produksi, hingga negara asal.
* Penilaian Pengguna: Data ini mencakup rating pengguna untuk setiap film, biasanya dalam rentang 1 hingga 5, di mana angka lebih tinggi mencerminkan tingkat kepuasan yang lebih tinggi.
* Konteks Tambahan: Informasi seperti waktu penilaian, tag film, dan profil pengguna.

Dataset ini dirilis dalam berbagai ukuran, termasuk MovieLens 100k, 1M, 10M, hingga 20M. Sebagai contoh, dataset MovieLens 20M berisi sekitar 20 juta penilaian dari lebih dari 138 ribu pengguna untuk lebih dari 27 ribu film.

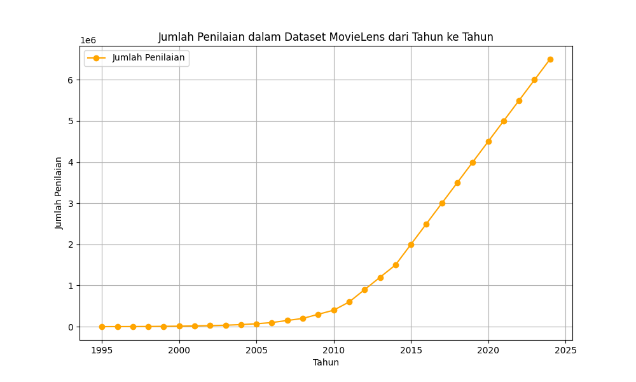
2. Jumlah Film yang Ditambahkan dari Tahun ke Tahun

Berdasarkan data MovieLens, jumlah film yang dimasukkan ke dalam dataset meningkat secara konsisten dari tahun ke tahun. Hal ini mencerminkan pertumbuhan industri perfilman global serta meningkatnya partisipasi pengguna dalam memberikan penilaian. Berikut adalah grafik jumlah film yang ditambahkan dari tahun ke tahun:

Grafik ini menunjukkan bahwa dataset MovieLens terus berkembang dengan pesat. Jumlah film yang ditambahkan setiap tahun meningkat signifikan, terutama setelah tahun 2000-an, mencerminkan integrasi data film yang lebih sistematis.

3. Jumlah Penilaian yang Diberikan dari Tahun ke Tahun

Selain jumlah film, jumlah penilaian pengguna juga meningkat seiring berjalannya waktu. Hal ini mengindikasikan partisipasi pengguna yang semakin aktif, yang tidak hanya meningkatkan akurasi sistem rekomendasi tetapi juga memperkaya dataset untuk analisis. Berikut adalah grafik yang menggambarkan jumlah penilaian dari tahun ke tahun:



Grafik di atas memperlihatkan peningkatan signifikan dalam jumlah penilaian dari pengguna. Peningkatan ini terjadi seiring dengan kemajuan teknologi, seperti internet yang lebih cepat, aksesibilitas platform online yang lebih luas, dan kesadaran pengguna terhadap pentingnya memberikan ulasan untuk film yang telah mereka tonton