**Laporan Tugas Kelompok**

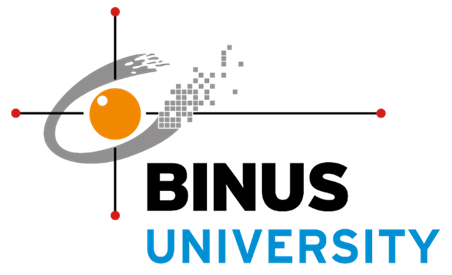
**Mata Kuliah *Database Analysis***

**Oleh:**

**Budi Nugroho - 2602102025**

**Djordy - 2602138315**

**Julian - 2602124801**



**Computer Science Program**

**Computer Science Study Program**

**School of Computer Science**

**Universitas Bina Nusantara**

**Jakarta**

**2024**

**Bab I**

**Pendahuluan**

Dalam menganalisis film-film teratas dengan rating IMDB, penting bagi kita untuk memahami konteks dan faktor-faktor yang berperan dalam menentukan peringkat sebuah film. Kualitas produksi, jalan cerita, dan kehadiran aktor dapat mempengaruhi peringkat, sementara analisis genre, rentang waktu, dan aspek-aspek sosial dan budaya juga harus dipertimbangkan. Selain itu, kita perlu menggunakan sumber daya dan metode analisis yang sesuai agar hasilnya akurat. Jangan lupakan pula mempertimbangkan kondisi ekonomi saat film tersebut dirilis. Dengan memahami semua ini, analisis kita dapat memberikan pandangan yang lebih komprehensif tentang tren dan faktor-faktor yang mempengaruhi kesuksesan film, bermanfaat untuk pengambilan keputusan industri film dan sebagai dasar penelitian lebih lanjut. Dalam Dataset yang diambil adalah rating Movie IMDB sampai Desember 2023 dari website Kaggle[1].

Melihat daftar film dengan rating IMDb bisa jadi kaya sumber inspirasi dan wawasan tentang selera penonton. Dari sana, kita bisa menangkap film-film favorit yang banyak dicintai orang, memahami genre yang sedang ngehits, dan merasakan sentuhan khusus dari sutradara atau aktor favorit. Dengan melihat perubahan peringkat dari waktu ke waktu, kita bisa menangkap perubahan selera penonton, serta melihat sejauh mana hubungan antara anggaran produksi dengan keberhasilan finansial dan peringkat IMDb. Tentu saja, melihat feedback dan review penonton membantu kita memahami apa yang benar-benar disukai dan diapresiasi oleh mereka. Dengan melibatkan data ini, industri film bisa membuat keputusan yang lebih baik dalam memberikan konten yang sesuai dengan harapan penonton, membuat film lebih bermakna, dan tentunya, sukses di pasar.

[1] <https://www.kaggle.com/datasets/kianindeed/imdb-movie-dataset-dec-2023/data>

**Bab II**

**Studi Literatur**

**2.1 Klasifikasi**

Klasifikasi adalah *supervised machine learning* dimana model mencoba memprediksi label yang benar dari data yang telah dimasukkan. Dalam klasifikasi, model dilatih sepenuhnya menggunakan data training, kemudian dilakukan evaluasi pada data test sebelum digunakan untuk melakukan prediksi pada data baru yang diberikan.

**2.2 Logistic Regression**

Logistic Regression adalah *supervised machine learning* yang menyelesaikan tugas klasifikasi biner dengan memprediksi probabilitas suatu hasil, peristiwa, atau observasi. Model ini memberikan hasil biner yang terbatas pada dua kemungkinan hasil. Contohnya adalah ya atau tidak, 0 atau 1, dan benar atau salah.

**2.3 Random Forest**

Random Forest adalah *supervised learning algorithm*. “Hutan” yang dibuat adalah kumpulan *decision tree*, hal ini dilatih oleh *bagging method*. Gagasan umum dari *bagging method* adalah kombinasi learning model yang meningkatkan hasil secara keseluruhan. Random Forest menggabungkan hasil dari banyak *decision tree* untuk mendapatkan hasil yang lebih akurat dan prediksi yang stabil.

**2.4 Python**

Python adalah bahasa pemrograman yang sering digunakan. Dalam implementasi algoritma Logistic Regression dan Random Forest akan menggunakan beberapa *library* Python. Melalui *library* Python menyediakan alat untuk melakukan proses data mining dengan mudah[7]. Beberapa *library* yang digunakan adalah numpy, sklearn, seaborn, pandas dan matplotlib.

**Bab III**

**Metode**

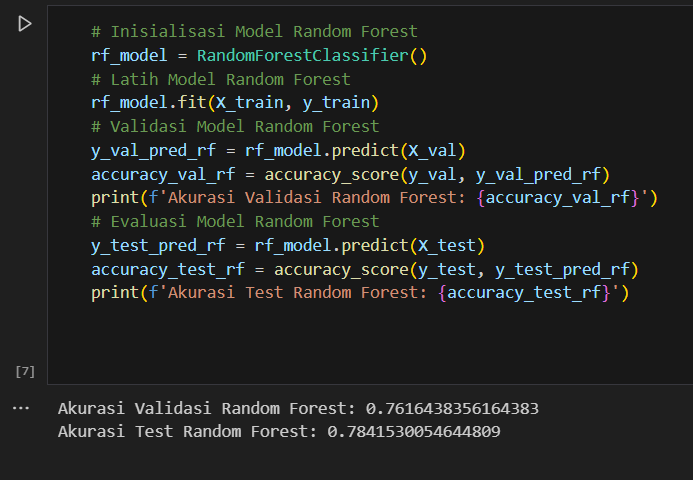
Dalam penelitian ini, kami akan membandingkan dua metode yang digunakan untuk melakukan klasifikasi film dalam analisis data IMDb. Pertama, metode yang kami digunakan disebut Random Forest, yang seperti arsitektur pohon yang dibangun untuk membantu memahami struktur data secara hierarkis. Ini membantu kita dalam mengelompokkan film-film ke dalam kategori-kategori yang lebih besar dan kemudian memecahnya menjadi kelompok-kelompok yang lebih kecil, yang membuatnya cocok untuk data besar seperti data IMDb. Metode kedua yang kami gunakan adalah Logistic Regression, yang lebih sederhana tetapi efektif. Keduanya akan mencoba mengklasifikasikan film-film ke dalam dua kelompok yaitu direkomendasikan (1) dan tidak direkomendasikan (0). Kedua metode ini akan digunakan pada dataset IMDb untuk melihat bagaimana mereka mengklasifikasikan film-film, serta untuk membandingkan kinerja mereka. Atribut yang akan digunakan adalah ‘Rating’, ‘Meta Score’, dan ‘Votes’. ‘Meta Score’ dan ‘Votes’ akan digunakan untuk menentukan apakah film direkomendasikan atau tidak. Dengan atribut ‘Rating’ dibuat Recommendation, yang akan merekomendasikan film yang memiliki score lebih atau sama dengan 7,

Pertama, dataset didownload dibaca dengan *python*. Kemudian, datasetnya di bersihkan dengan menghilangkan atribut yang tidak dipakai dan menghilangkan data yang tidak valid. Setelah itu dilakukan pembagian data, menjadi data latihan, pengujian dan validasi. Akhirnya dibuat model dan dilakukan evaluasi akurasi model dengan akurasi dan *confusion matrix*.

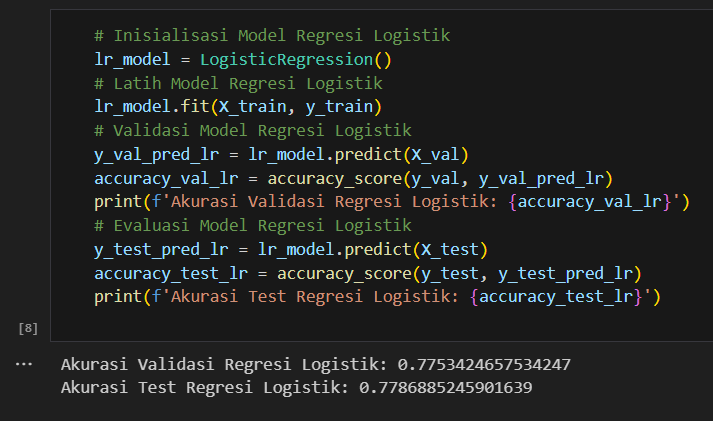
**Bab IV**

**Hasil**

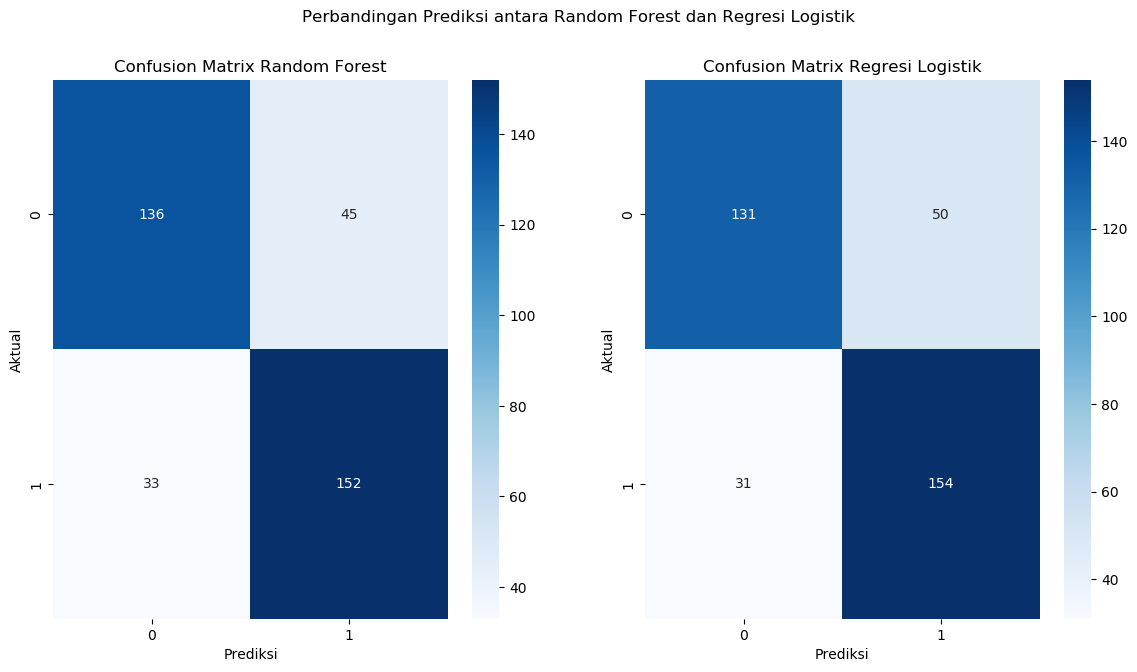
**4.1 Random Forest**

****

**4.2 Logistic Regression**

****

**4.3 Confusion Matrix**

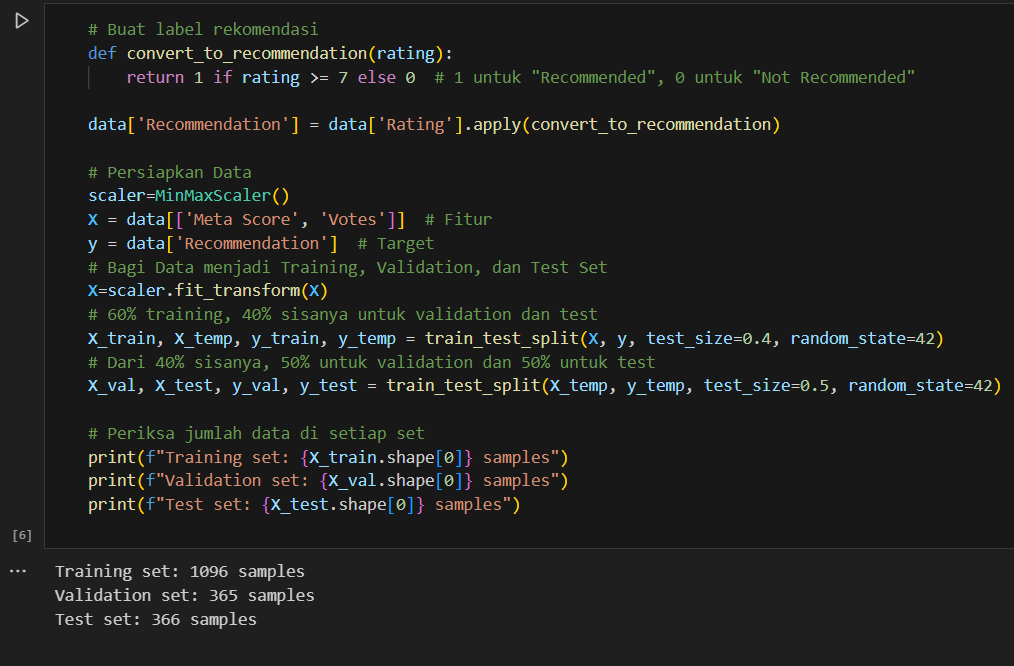
****

**Bab V**

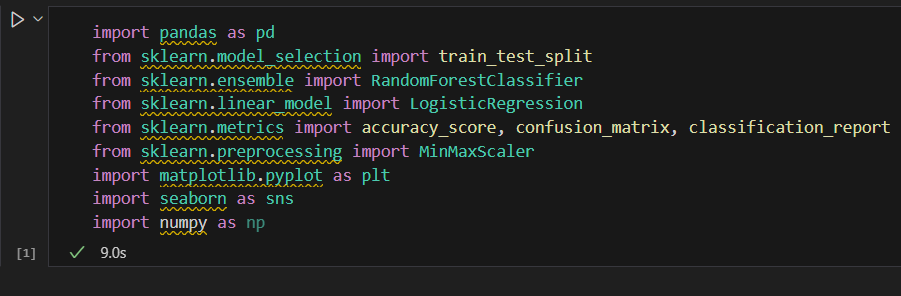
**Kesimpulan**

Dalam melakukan prediksi Logistic Regression dan Random Forest memiliki akurasi yang berbeda. Random Forest memiliki akurasi yang lebih tinggi dibanding dengan Logistic Regression. Ini juga bisa dilihat dalam Confusion Matrix Random Forest yang lebih tepat dalam menebak apakah movie direkomendasikan atau tidak. Jadi, Random Forest adalah algoritma yang lebih baik dalam memprediksi apakah Movie direkomendasi atau tidak.

Lampiran:



Sample Training, Validation, Testing



Library Python yang digunakan

Referensi:

Dataset:

<https://www.kaggle.com/datasets/kianindeed/imdb-movie-dataset-dec-2023/data>

Referensi Literatur:

<https://www.datacamp.com/blog/classification-machine-learning>

<https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

<https://builtin.com/data-science/random-forest-algorithm>

Herlawati, H., & Handayanto, R. T. (2020). Penggunaan Matlab dan Python dalam Klasterisasi Data. *Jurnal Kajian Ilmiah*, *20*(1), 103-118.Herlawati, H., & Handayanto, R. T. (2020). Penggunaan Matlab dan Python dalam Klasterisasi Data. *Jurnal Kajian Ilmiah*, *20*(1), 103-118.