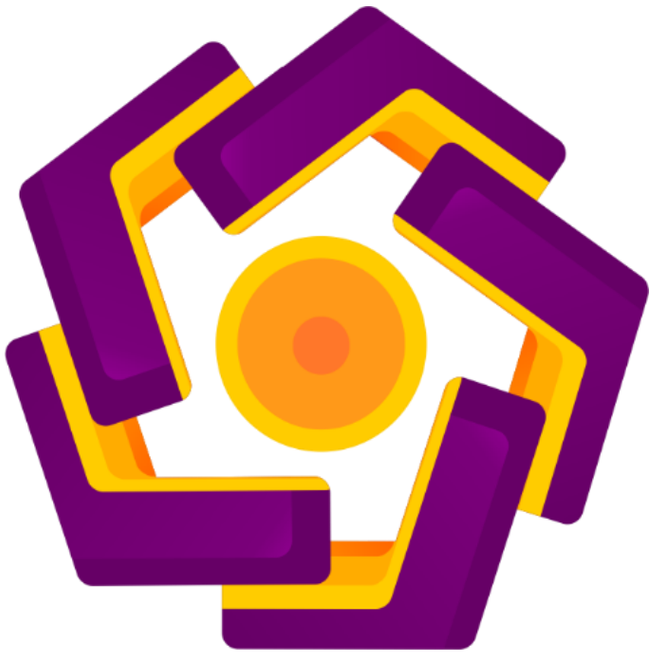
UJIAN AKHIR SEMESTER

BIG DATA & PREDICTIVE ANALYTICS LANJUT

(KLASIFIKASI POPULER ANIME)



DOSEN PENGAMPU:

Enda Putri Atika, M.Kom

disusun oleh:

Ixvannando Asdik Prasetyawan 22.11.5188

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM

YOGYAKARTAYOGYAKARTA

2025

Soal Ujian *(disesuaikan dengan sifat ujian)*

1. Berdasarkan apa yang sudah Anda pelajari, silahkan gunakan kemampuan anda untuk menyelesaikan sebuah menggunakan *classification* yang melibatkan penggunaan Machine Learning. (SCPMK 1534113, 50 Poin)

   a Pilih satu bidang yang Anda beserta rekan tim minati, jabarkan alasan pemilihan bidang tersebut dan jelaskan apa yang ingin dicapai dengan memilih topik ini.

   Saya memilih klasifikasi popular anime karena anime merupakan salah satu bentuk seni dan hiburan yang memiliki basis penggemar yang sangat luas di seluruh dunia. Dalam dunia yang semakin terhubung, pemahaman tentang faktor-faktor yang memengaruhi popularitas anime dapat membantu berbagai pihak, seperti platform streaming, studio produksi, dan komunitas penggemar, untuk menyajikan konten yang lebih relevan dan sesuai dengan preferensi audiens.

   Apa yang Ingin Dicapai :

   - Identifikasi Faktor Penentu Popularitas Menggunakan teknik machine learning untuk memahami hubungan antara berbagai fitur (seperti skor, jumlah episode, genre, dan jumlah penonton) terhadap popularitas suatu anime.
   - Peningkatan Rekomendasi Konten Dengan memahami tren popularitas, kami ingin membantu platform streaming atau pengembang aplikasi untuk memberikan rekomendasi anime yang lebih sesuai dengan minat pengguna.
   - Dukungan pada Produksi Konten Memberikan wawasan kepada studio anime tentang elemen-elemen yang perlu dipertimbangkan dalam menciptakan anime yang lebih diminati.
   - Peningkatan Pengalaman Pengguna Membantu penggemar menemukan anime yang sesuai dengan preferensi mereka, berdasarkan pola popularitas yang dianalisis.

   b Ceritakan proses mendapatkan data dan informasi lengkap mengenai data tersebut (seperti waktu, penjelasan setiap kolom, sumber dll). Data yang digunakan harus data terbaru dengan range 1-4 tahun kebelakang.

   Pengambilan dataset dilakukan pada 17/1/2025.

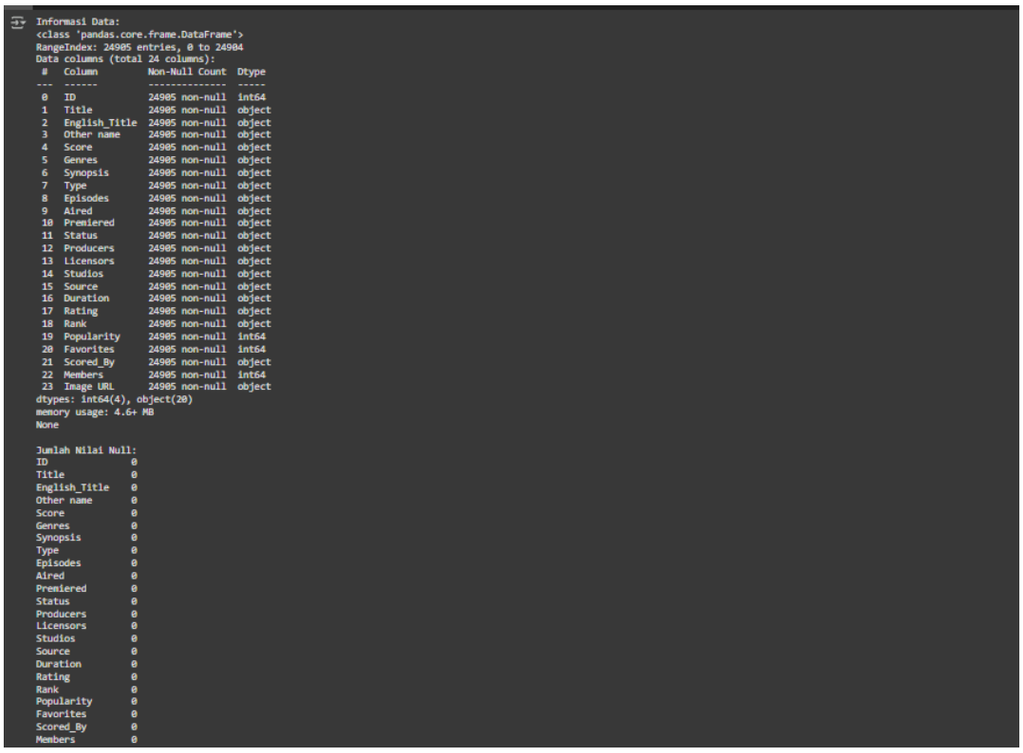
   Dataset ini memiliki 24 kolom:

   1. ID
   2. Title
   3. Englist_Title
   4. Other name
   5. Score
   6. Genres
   7. Synopsis
   8. Type
   9. Episodes

10. Aired
11. Premiered
12. Status
13. Producers
14. Licensors
15. Studios
16. Source
17. Duration
18. Rating
19. Rank
20. Popularity
21. Favorites
22. Scores_By
23. Members
24. Image URL

Link Dataset : https://www.kaggle.com/datasets/dbdmobile/myanimelist-dataset?select=anime-dataset-2023.csv

c Lakukan pre-processing data dengan memeriksa tipe data, mengganti nama kolom, memeriksa nilai null, mengubah tipe data (agar bisa di proses), menampilkan summary, dan menampilkan matriks korelasinya menggunakan metode metode yang pernah dipelajari.

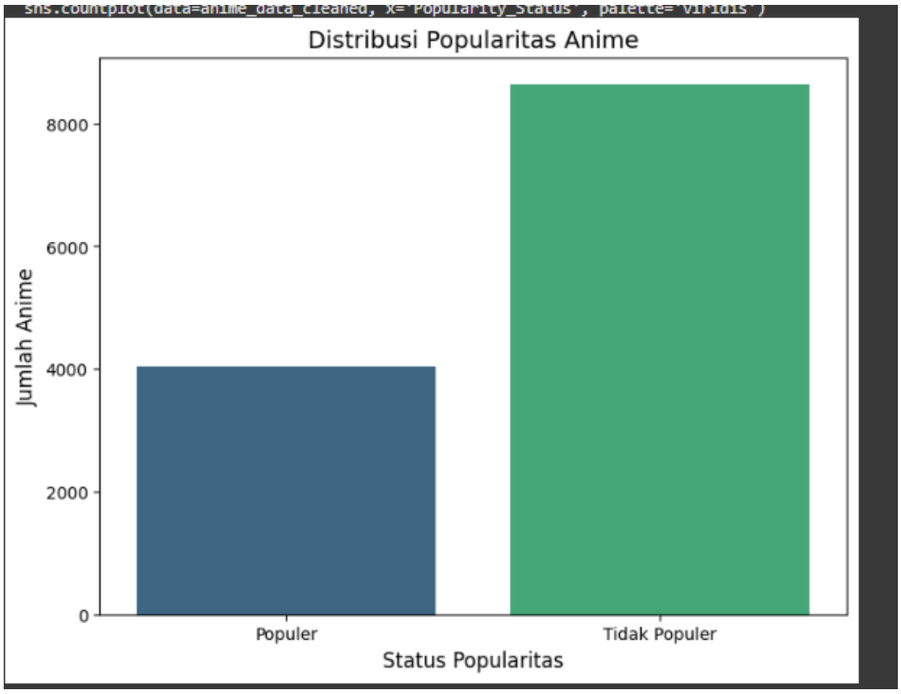


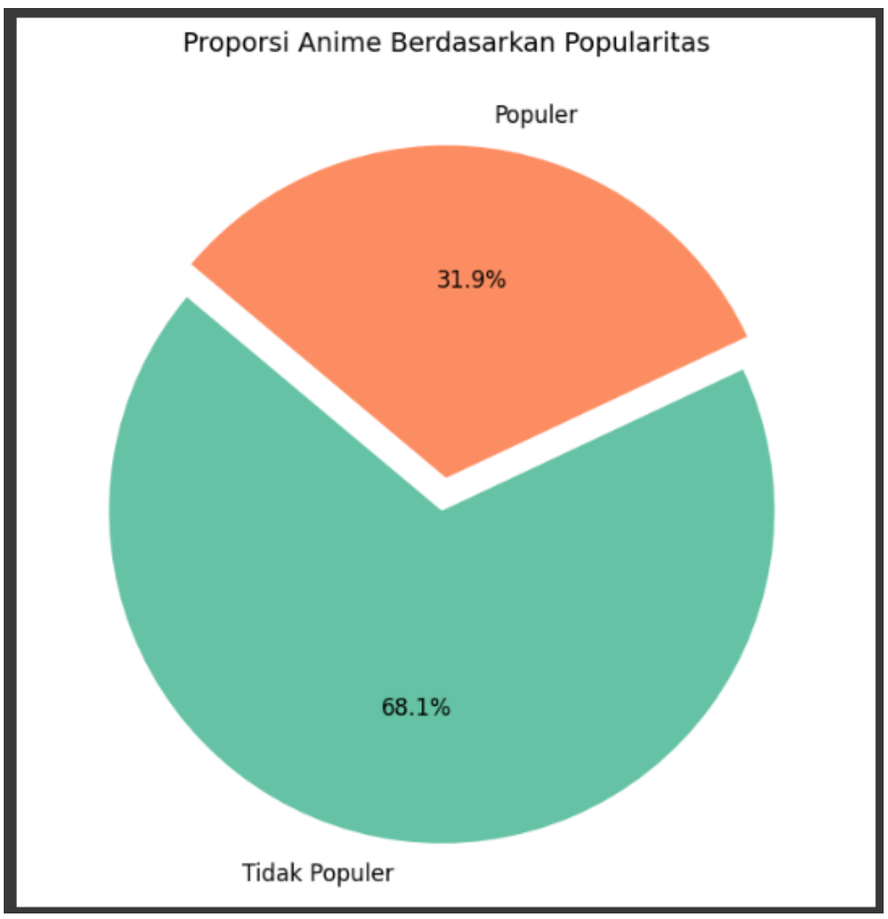
```
Informasi Data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24905 entries, 0 to 24904
Data columns (total 24 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   ID             24905 non-null  int64
 1   Title          24905 non-null  object
 2   English_Title  24905 non-null  object
 3   Other name     24905 non-null  object
 4   Score          24905 non-null  object
 5   Genres         24905 non-null  object
 6   Synopsis       24905 non-null  object
 7   Type           24905 non-null  object
 8   Episodes       24905 non-null  object
 9   Aired          24905 non-null  object
 10  Premiered      24905 non-null  object
 11  Status         24905 non-null  object
 12  Producers      24905 non-null  object
 13  Licensors      24905 non-null  object
 14  Studios        24905 non-null  object
 15  Source         24905 non-null  object
 16  Duration       24905 non-null  object
 17  Rating         24905 non-null  object
 18  Rank           24905 non-null  object
 19  Popularity     24905 non-null  int64
 20  Favorites      24905 non-null  int64
 21  Scored_By      24905 non-null  object
 22  Members        24905 non-null  int64
 23  Image URL      24905 non-null  object
dtypes: int64(4), object(20)
memory usage: 4.6+ MB
None

Jumlah Nilai Null:
ID               0
Title            0
English_Title    0
Other name       0
Score            0
Genres           0
Synopsis         0
Type             0
Episodes         0
Aired            0
Premiered        0
Status           0
Producers        0
Licensors        0
Studios          0
Source           0
Duration         0
Rating           0
Rank             0
Popularity       0
Favorites        0
Scored_By        0
Members          0
```

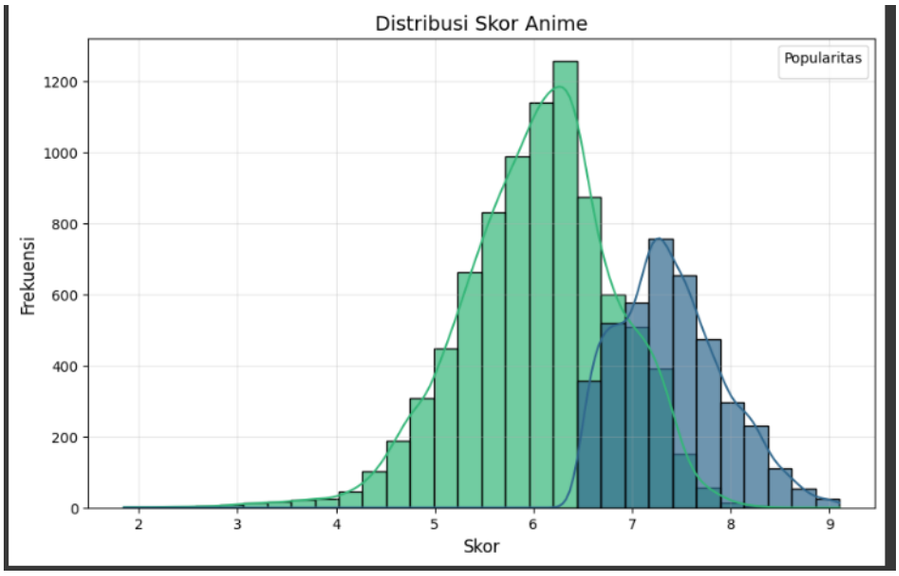
```
Ringkasan Statistik:
                ID          Score     Episodes          Rank    Popularity  \
count  12701.000000  12701.000000  12650.000000  12701.000000  12701.000000
mean   21102.176364      6.478799     13.547826   6350.514920   7491.516180
std    16646.530305      0.941088     53.413329   3666.272186   4769.959409
min        1.000000      1.850000      1.000000      1.000000      1.000000
25%     4522.000000      5.840000      1.000000   3175.000000   3258.000000
50%    18703.000000      6.510000      3.000000   6349.000000   7162.000000
75%    36027.000000      7.170000     13.000000   9525.000000  11532.000000
max    55647.000000      9.100000   3057.000000  12701.000000  19191.000000

           Favorites     Scored_By       Members
count   12701.000000  1.270100e+04  1.270100e+04
mean      838.445162  3.648718e+04  7.084740e+04
std      6067.451606  1.290340e+05  2.139781e+05
min         0.000000  1.000000e+02  1.800000e+02
25%         1.000000  4.530000e+02  1.458000e+03
50%        10.000000  2.390000e+03  6.562000e+03
75%        92.000000  1.685300e+04  4.058400e+04
max    217606.000000  2.660903e+06  3.744541e+06
```

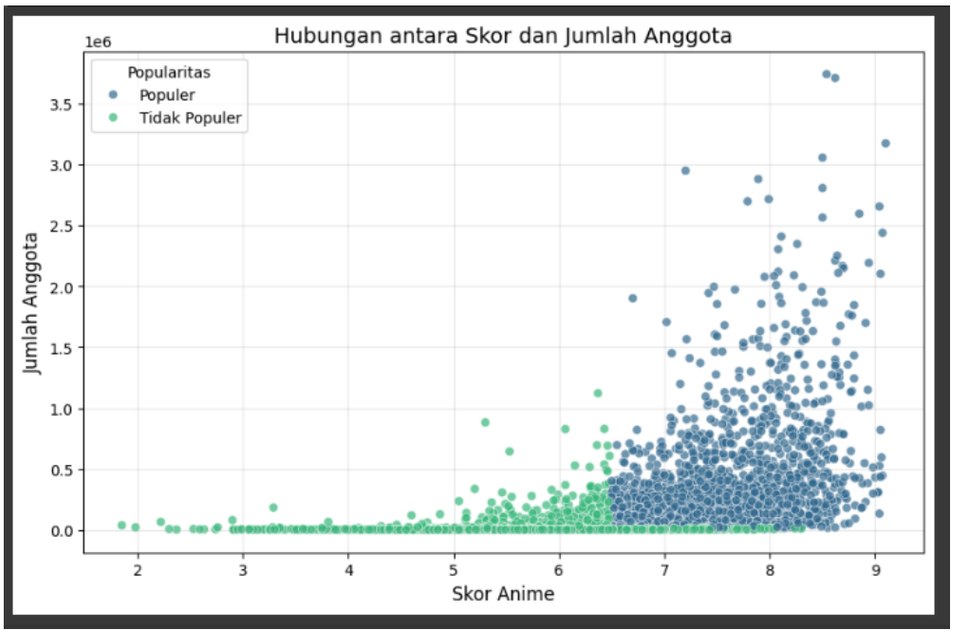
```
Matriks Korelasi:
                   ID     Score  Episodes      Rank  Popularity  Favorites  \
ID           1.000000 -0.088173 -0.061964  0.103634    0.101149  -0.011417
Score       -0.088173  1.000000  0.063445 -0.976923   -0.706196   0.243995
Episodes    -0.061964  0.063445  1.000000 -0.058765   -0.011963   0.060565
Rank         0.103634 -0.976923 -0.058765  1.000000    0.737103  -0.205851
Popularity   0.101149 -0.706196 -0.011963  0.737103    1.000000  -0.204764
Favorites   -0.011417  0.243995  0.060565 -0.205851   -0.204764   1.000000
Scored_By    0.018909  0.357896  0.042992 -0.335063   -0.389794   0.785321
Members      0.029228  0.397911  0.044609 -0.375835   -0.444417   0.774898

             Scored_By   Members
ID            0.018909  0.029228
Score         0.357896  0.397911
Episodes      0.042992  0.044609
Rank         -0.335063 -0.375835
Popularity   -0.389794 -0.444417
Favorites     0.785321  0.774898
Scored_By     1.000000  0.989127
Members       0.989127  1.000000
```

d Gunakan exploratory data analysis (EDA) untuk melihat sudut padang yang ada mengenai data (minimal 4) dua diantaranya bar dan pie chart, 2 diantaranya bebas. Berikan penjelasan.









e Berdasarkan analisis data tersebut, jelaskan alasan pemilihan kolom/fitur yang relevan untuk menyelesaikan permasalahan yang ingin dicapai.

Penjelasan alasan pemilihan kolom/fitur yang relevan dalam konteks klasifikasi popularitas anime:

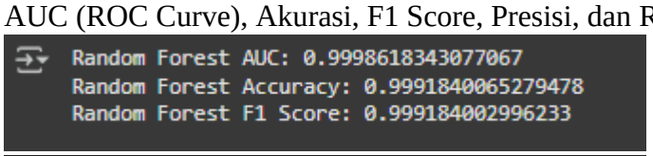
1. Fitur yang Dipilih
    - Score: Menggambarkan penilaian rata-rata dari pengguna. Penilaian ini menjadi indikator penting untuk mengetahui kualitas anime dan pengaruhnya terhadap popularitas.
    - Members: Menunjukkan jumlah anggota yang menambahkan anime tersebut ke dalam daftar mereka. Semakin tinggi jumlah anggota, semakin besar kemungkinan anime tersebut populer.
    - Rank: Posisi ranking anime berdasarkan skor. Anime dengan peringkat tinggi cenderung lebih populer.

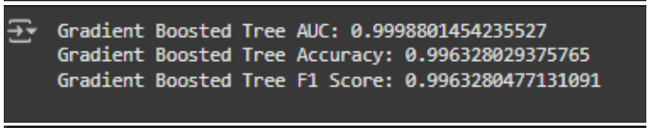
- Scored_By: Jumlah pengguna yang memberikan skor. Semakin banyak orang yang memberikan skor, semakin tinggi kemungkinan bahwa anime tersebut populer.
- Episodes: Jumlah episode dari anime. Anime dengan jumlah episode yang lebih banyak atau lebih sedikit dapat menarik audiens tertentu.

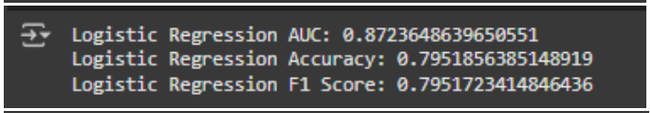
2. Alasan Pemilihan Kolom
   - Fitur-fitur ini secara langsung atau tidak langsung berhubungan dengan persepsi pengguna, jumlah audiens, dan minat komunitas, yang semuanya menjadi faktor penting dalam menentukan popularitas.
   - Kolom yang dipilih memiliki keterkaitan logis dengan popularitas, misalnya, anime dengan skor tinggi atau jumlah anggota yang besar biasanya memiliki tingkat popularitas yang tinggi.
3. Mengapa Tidak Semua Kolom Digunakan? Karena
   - Beberapa kolom seperti Genres, Synopsis, atau Studios memiliki nilai teks atau kategori yang sulit diolah tanpa teknik pemrosesan tambahan (misalnya, encoding atau NLP). Penggunaan kolom ini membutuhkan waktu lebih lama untuk proses analisis.
   - Beberapa kolom, seperti English_Title atau Image_URL, tidak relevan untuk menentukan popularitas karena tidak memberikan informasi tambahan terkait minat audiens.

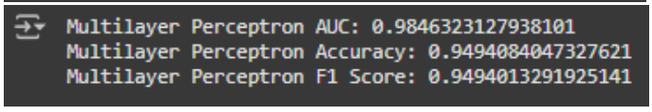
2. Pengembangan model machine learning. (SCPMK 1534114, 50 Poin)

a. Gunakan minimal 4 model Machine Learning dari library Spark untuk menyelesaikan masalah yang Anda pilih. 2 Model sesuai dengan instruksi (Random Forest, Gradient Boost Tree) dan dua model lain bebas (belum pernah dibahas). Lalu bandingkan hasilnya menggunakan metriks seperti AUC (ROC Curve), Akurasi, F1 Score, Presisi, dan Recall.









Random Forest

- **AUC:** 0.9999 (sangat tinggi)
- **Accuracy:** 0.9992 (sangat akurat)
- **F1 Score:** 0.9992 (sangat baik dalam keseimbangan presisi dan recall)

- **Kesimpulan:** Random Forest memiliki performa sangat baik dan hampir sempurna dalam memprediksi popularitas anime.

Gradient Boosted Tree

- **AUC**: 0.9999 (sangat tinggi, sedikit lebih baik dari Random Forest)
- **Accuracy**: 0.9963 (sangat baik)
- **F1 Score**: 0.9963 (sedikit di bawah Random Forest)
- **Kesimpulan**: Performa Gradient Boosted Tree mendekati Random Forest, tetapi lebih lambat dalam pelatihan dan prediksi.

Logistic Regression

- **AUC**: 0.8724 (cukup baik)
- **Accuracy**: 0.7952 (kurang akurat dibandingkan Random Forest dan GBT)
- **F1 Score**: 0.7952 (rendah untuk data tidak seimbang)
- **Kesimpulan**: Logistic Regression lebih sederhana tetapi tidak cocok untuk dataset ini karena performanya jauh lebih rendah.

Multilayer Perceptron (MLP)

- **AUC**: 0.9846 (sangat baik)
- **Accuracy**: 0.9494 (sangat baik)
- **F1 Score**: 0.9494 (cukup tinggi)
- **Kesimpulan**: MLP memiliki performa baik tetapi tidak seefisien Random Forest atau Gradient Boosted Tree.

b. Dari ke-4 model classification tersebut, pilih dua model dengan performa terbaik dan lakukan hyperparameter tuning untuk melihat perubahan performa yang dihasilkan. Lalu tentukan model terbaik yang bisa menjadi solusi pada masalah yang Anda tetapkan diawal.

```
Tuning model Random Forest dimulai...
Tuning Random Forest selesai!
Tuning model Gradient Boosted Tree dimulai...
Tuning Gradient Boosted Tree selesai!
Random Forest AUC setelah tuning: 0.9994271390281501
Gradient Boosted Tree AUC setelah tuning: 0.9993472856805589
```

```
Final Model Comparison:
Random Forest (Tuned) AUC: 0.9994271390281501
Gradient Boosted Tree (Tuned) AUC: 0.9993472856805589
Random Forest is the best model.
```

c. Jabarkan karakteristik model terbaik yang Anda dapatkan terhadap korelasinya dengan data. Apakah ada sifat tertentu dari data yang ternyata cocok dengan model dan sebaliknya?

Random Forest

Karakteristik Model:

Random Forest adalah ensemble model berbasis pohon keputusan. Model ini bekerja dengan membuat banyak pohon keputusan independen dan menggabungkan hasilnya untuk menghasilkan prediksi yang kuat.

Korelasinya dengan Data:

- Cocok dengan data yang memiliki banyak fitur numerik: Dataset ini memiliki fitur numerik utama seperti Score, Members, Scored_By, dan Popularity. Random Forest secara alami mampu menangani data numerik dan mengidentifikasi hubungan non-linear di antara fitur.

- Tahan terhadap outlier: Random Forest tidak terlalu terpengaruh oleh outlier dalam data, sehingga cocok untuk dataset yang mungkin memiliki distribusi nilai yang ekstrem.
- Stabilitas tinggi terhadap data yang tidak seimbang: Meskipun dataset ini mungkin memiliki distribusi yang tidak seimbang pada Popularity_Status, Random Forest tetap memberikan prediksi yang akurat dengan AUC dan akurasi yang hampir sempurna.

Kelemahan:

Karena Random Forest membuat banyak pohon, model ini bisa memakan waktu lebih lama untuk prediksi pada dataset besar jika jumlah pohon sangat besar.

Link Colab:

https://colab.research.google.com/drive/1wjmQhBdaT8KYv1SZMPzur3pfvckT3eZI?usp=sharing