Analisis Kegemaran Membaca di Indonesia Tahun 2020-2023

Analysis of Reading Interest in Indonesia 2020-2023

Muhammad Naufal Adnansyah [1], Fajar Triady Putra [2], Rafi Raihan[3], Nurcahyo Bambang Irawan[4]
[1]Universitas Al-Azhar Indonesia ; [2]Jl. Sisingamangaraja, RT.2/RW.1, Selong, Kec. Kby. Baru, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12110.
[3]Jurusan Teknik Informatika, Fakultas Sains dan Teknologi.
e-mail: [1]0102523607@student.uai.ac.id, [2]0102523602@student.uai.ac.id
[3]0102523610@student.uai.ac.id, [4]0102523608@student.uai.ac.id,

Abstrak

Minat baca masyarakat Indonesia menjadi salah satu indikator penting dalam mengukur tingkat literasi dan perkembangan intelektual bangsa. Penelitian ini bertujuan untuk menganalisis Tingkat Kegemaran Membaca (TGM) di Indonesia pada periode 2020-2023 dengan melihat tren dan faktor-faktor yang memengaruhinya. Dengan menggunakan metode *clustering* dan *classification* untuk memahami pola minat baca di berbagai kelompok masyarakat. *Clustering* dilakukan menggunakan algoritma *K-Means* untuk mengelompokkan wilayah berdasarkan kesamaan tingkat minat baca, sedangkan model *Logistic Regression* dan *Random Forest* digunakan untuk memprediksi faktor yang paling berpengaruh terhadap TGM.

Hasil analisis menunjukkan bahwa terdapat beberapa kelompok wilayah dengan pola minat baca berbeda, di mana wilayah dengan akses internet yang lebih baik cenderung memiliki TGM lebih tinggi. Penelitian ini menegaskan bahwa pendekatan berbasis data dapat meningkatkan pemahaman tentang pola membaca masyarakat dan membantu dalam perumusan kebijakan literasi yang lebih efektif. Hasil ini diharapkan dapat mendukung strategi peningkatan minat baca berbasis data untuk memastikan pemerataan literasi di seluruh Indonesia.
Kata kunci: Minat Baca, Analisis Data, *Clustering, Clasification,* Akses Internet.

Abstract

The reading interest of Indonesian people is one of the important indicators in measuring the level of literacy and intellectual development of the nation. This study aims to analyze the reading level in Indonesia in the period 2020-2023 by looking at trends and factors that influence it. By using clustering and classification methods to understand reading interest patterns in various community groups. Clustering is done using the K-Means algorithm to group regions based on similarities in reading interest levels, while Logistic Regression and Random Forest models are used to predict the most influential factors on TGM.

The results of the analysis show that there are several groups of areas with different reading interest patterns, where areas with better internet access tend to have higher TGM. This study confirms that a data-driven approach can improve understanding of people's reading patterns and assist in the formulation of more effective literacy policies. These results are expected to support data-driven reading interest improvement strategies to ensure literacy equity across Indonesia.
Keywords: *Reading Interest, Data Analytics, Clustering, Clasification, Internet Access.*

1. PENDAHULUAN

Tingkat literasi merupakan faktor kunci dalam menentukan kualitas sumber daya manusia suatu negara. Di Indonesia, tingkat kegemaran membaca (TGM) masih menjadi tantangan besar, meskipun berbagai program telah diluncurkan oleh pemerintah untuk meningkatkan budaya literasi. Berdasarkan data dari Perpustakaan Nasional Republik Indonesia (Perpusnas), TGM menunjukkan tren peningkatan selama periode 2020-2023, dengan nilai tertinggi pada tahun 2022 mencapai 63,58. Namun, disparitas minat baca antarwilayah masih menjadi masalah yang perlu diperhatikan, terutama terkait akses terhadap bahan bacaan, fasilitas perpustakaan, serta pengaruh teknologi digital.

Kemajuan teknologi membuka peluang baru dalam meningkatkan minat baca, terutama melalui digitalisasi perpustakaan dan penyebaran buku elektronik (*e-book*). Namun, belum banyak penelitian yang secara spesifik menganalisis bagaimana faktor-faktor ini berpengaruh terhadap TGM dalam jangka waktu tertentu. Oleh karena itu, penelitian ini bertujuan untuk mengidentifikasi pola minat baca masyarakat Indonesia serta faktor-faktor yang memengaruhi peningkatannya melalui pendekatan berbasis data menggunakan teknik *clustering* dan *classification*.

Beberapa isu yang berkaitan dengan rendahnya minat baca di Indonesia antara lain:

1. **Kesenjangan Literasi Digital** – Tidak semua wilayah memiliki akses yang sama terhadap sumber bacaan digital, terutama di daerah terpencil.
2. **Kurangnya Fasilitas Perpustakaan** – Ketersediaan perpustakaan dan bahan bacaan yang terbatas menghambat kebiasaan membaca di kalangan masyarakat.
3. **Perubahan Pola Konsumsi Informasi** – Perkembangan media sosial dan hiburan digital menyebabkan pergeseran preferensi masyarakat terhadap konsumsi informasi.

Untuk mengatasi isu-isu tersebut, diperlukan analisis lebih lanjut mengenai faktor-faktor yang paling berpengaruh terhadap TGM, sehingga kebijakan literasi dapat disusun secara lebih efektif dan berbasis data.

Beberapa penelitian telah dilakukan terkait minat baca dan literasi digital di Indonesia. Misalnya:

- **Gunawan & Dewi (2021)** yang menganalisis hubungan antara akses perpustakaan digital dan tingkat literasi masyarakat, menemukan bahwa digitalisasi perpustakaan berkontribusi terhadap peningkatan TGM.
- **Sutrisno et al. (2022)** yang menggunakan pendekatan *machine learning* untuk memprediksi faktor-faktor yang berkontribusi terhadap kebiasaan membaca, menunjukkan bahwa akses internet dan program literasi sekolah memiliki pengaruh yang signifikan.
- **Yulianto (2023)** yang melakukan *clustering* menggunakan algoritma *K-Means* untuk mengelompokkan daerah berdasarkan tingkat literasi, menunjukkan bahwa daerah perkotaan cenderung memiliki minat baca yang lebih tinggi dibandingkan daerah pedesaan.

Namun, penelitian-penelitian tersebut masih terbatas dalam cakupan temporal dan metodologi prediksi. Oleh karena itu, penelitian ini akan memperluas analisis dengan menggabungkan pendekatan *clustering* menggunakan *K-Means* serta metode *classification* seperti *Logistic Regression* dan *Random Forest* untuk mengidentifikasi faktor yang paling berpengaruh terhadap peningkatan TGM di Indonesia selama periode 2020-2023.

Penelitian ini diharapkan dapat memberikan kontribusi bagi pemangku kebijakan dalam meningkatkan strategi literasi berbasis data guna mendukung pemerataan akses literasi di seluruh Indonesia.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis analisis data untuk mengevaluasi Tingkat Kegemaran Membaca (TGM) masyarakat Indonesia periode 2020-2023. Analisis dilakukan dalam dua tahap utama:

- *Clustering* menggunakan algoritma *K-Means* dan *DBSCAN* untuk mengelompokkan daerah berdasarkan tingkat minat baca.
- *Klasifikasi* menggunakan *Decision Tree* dan *Random Forest* untuk memprediksi faktor-faktor yang memengaruhi peningkatan atau penurunan TGM.

Penelitian ini akan memberikan rekomendasi bagi pemangku kebijakan dalam meningkatkan minat baca melalui strategi berbasis data. Pendekatan kombinasi *clustering* dan *classification* diharapkan dapat memberikan wawasan lebih mendalam terkait pola literasi masyarakat Indonesia.

2.1 Sumber Data

Data yang digunakan dalam penelitian ini berasal dari Perpustakaan Nasional Republik Indonesia (Perpusnas), yang mencakup:

- Nilai TGM per tahun (2020-2023) untuk berbagai wilayah di Indonesia.
- Faktor-faktor yang berpotensi memengaruhi TGM, seperti jumlah perpustakaan, akses terhadap internet, jumlah buku digital, dan program literasi pemerintah.
- Data demografi seperti tingkat pendidikan dan akses informasi di setiap daerah.

2. 2 Tahapan Analisis Data

A. Clustering dengan K-Means dan DBSCAN
   Metode clustering digunakan untuk mengelompokkan daerah berdasarkan tingkat minat baca dan faktor-faktor pendukungnya.
   - K-Means Clustering:
     Menentukan jumlah klaster optimal menggunakan metode Elbow dan Silhouette Score. Mengelompokkan wilayah berdasarkan kemiripan pola minat baca.
   - DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
     Mengelompokkan data berdasarkan kepadatan titik data, lebih fleksibel terhadap data yang tidak terdistribusi secara linier. Mengidentifikasi wilayah yang memiliki pola anomali dalam tingkat minat baca.

B. Klasifikasi dengan Decision Tree dan Random Forest
   Metode classification digunakan untuk mengidentifikasi faktor utama yang berkontribusi terhadap TGM.
   - Decision Tree:
     Menggunakan struktur pohon keputusan untuk mengidentifikasi faktor yang paling memengaruhi TGM. Mengevaluasi performa model dengan metrik akurasi, presisi, dan recall untuk memastikan keandalan hasil klasifikasi.
   - Random Forest:
     Menggunakan metode ensemble learning untuk mengukur kepentingan setiap variabel dalam memengaruhi TGM. Menilai akurasi model melalui validasi silang dan confusion matrix.

3. HASIL DAN PEMBAHASAN

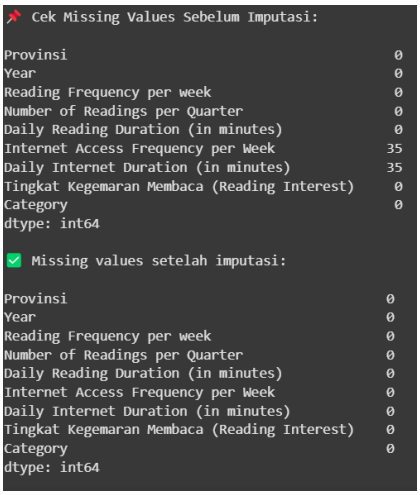
3.1 Preprocessing Data

Mengatasi Missing Values

- Beberapa variabel memiliki nilai yang hilang (NaN).

Solusi Preprocessing:

- Mean/Median/Mode Imputation: Mengisi nilai yang hilang dengan rata-rata, median, atau modus.



Gambar 1. Perbaikan Missing Value

3.2 Normalisasi Fitur

Data numerik sering memiliki **skala yang berbeda** (misalnya, "Daily Reading Duration" dalam menit bisa berkisar dari 10–300 menit, sedangkan "Reading Frequency per Week" hanya berkisar dari 1–7).

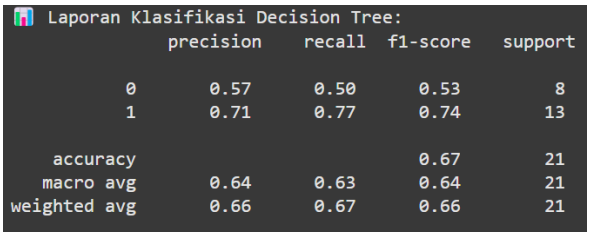
Solusi Preprocessing:

- StandardScaler (Standarisasi): Mengubah data agar memiliki mean=0 dan standar deviasi=1.
- MinMaxScaler (Normalisasi): Mengubah data ke rentang 0 hingga 1

3.3 Klasifikasi dengan Decision Tree dan Random Forest

A. Analisis model Decision Tree

- Akurasi: 0.6667 (66.67%)
- Jumlah Data Training: 84 sampel digunakan untuk melatih model.
- Jumlah Data Testing: 21 sampel digunakan untuk menguji performa model setelah dilatih.



Gambar 2. Laporan klasifikasi decision tree

- Precision untuk Kelas 0 (57%) lebih rendah dibandingkan Kelas 1 (71%), artinya model kurang akurat dalam mengenali kelas 0.

- Recall untuk Kelas 1 (77%) lebih tinggi dibandingkan Kelas 0 (50%), artinya model lebih mampu menangkap semua kasus kelas 1 tetapi lebih sering salah dalam mengenali kelas 0.
- Akurasi keseluruhan hanya 67%, menunjukkan bahwa model masih bisa ditingkatkan. Contoh dalam Konteks Minat Membaca:
- Kelas 0 = "Minat Membaca Rendah"
- Kelas 1 = "Minat Membaca Tinggi"

B. Analisis model Random Forest

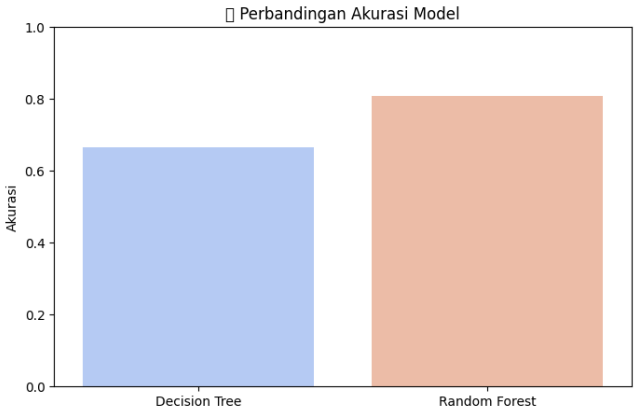
- Akurasi: 0.8095 (80.95%)

✅ Akurasi Random Forest: 0.8095
📊 Laporan Klasifikasi Random Forest:

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.85	0.85	0.85	13
accuracy			0.81	21
macro avg	0.80	0.80	0.80	21
weighted avg	0.81	0.81	0.81	21

Gambar 3. Laporan Klasifikasi Random Forest

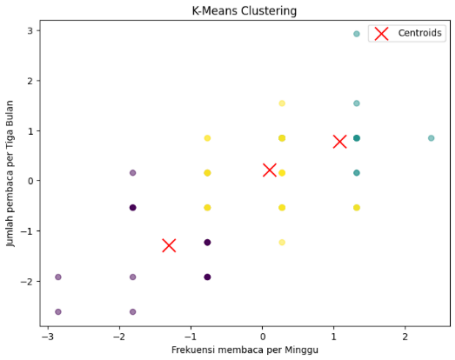
- Random Forest lebih akurat (81%) dibandingkan Decision Tree (67%), menunjukkan bahwa penggunaan beberapa pohon keputusan meningkatkan performa.
- Precision dan Recall lebih seimbang untuk kedua kelas (75% dan 85%), artinya model lebih stabil dibandingkan Decision Tree.
- Model lebih mampu mengenali kelas negatif dan positif dengan lebih baik dibandingkan Decision Tree.



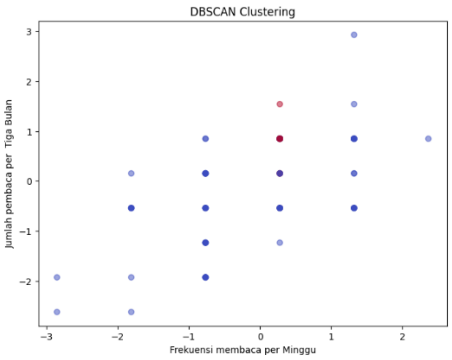
Gambar 4. Perbandingan Akurasi Model

3.4 Analisis Clustering dengan K-Means dan DB Scan

- Silhouette Score K-Means: 0.2950
- Silhouette Score DBSCAN: -0.0247



Gambar 6. K-Means Clustering



Gambar 7. DB Scan Clustering

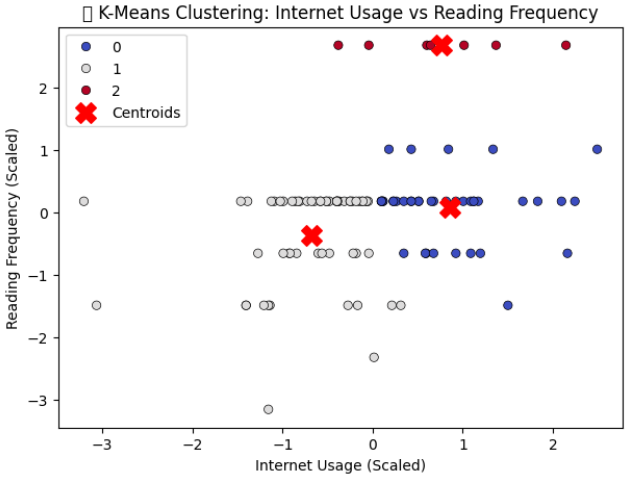
- Nilai mendekati 1 → Cluster terbentuk dengan baik dan terpisah jelas.
- Nilai mendekati 0 → Cluster saling tumpang tindih.
- Nilai negatif (< 0) → Banyak titik lebih mirip dengan cluster lain daripada cluster mereka sendiri (clustering buruk).
- K-Means memiliki silhouette score 0.2950, yang berarti clustering masih moderat tetapi tidak optimal.
- DBSCAN memiliki silhouette score -0.0247, yang menunjukkan clustering tidak terbentuk dengan baik dan banyak titik mungkin salah klasifikasi atau termasuk sebagai noise.
- Semakin sering seseorang membaca dalam seminggu, semakin tinggi jumlah bacaan dalam 3 bulan.
- Beberapa cluster memiliki keterkaitan dengan akses internet, menunjukkan hubungan antara penggunaan internet dan kebiasaan membaca.
- K-Means memberikan hasil clustering yang lebih baik dibandingkan DBSCAN untuk dataset ini.
- DBSCAN kurang efektif karena data mungkin tidak memiliki kepadatan yang cukup jelas untuk membentuk cluster yang stabil.
- Jika dataset memiliki banyak outlier atau pola distribusi tidak seragam, DBSCAN bisa jadi lebih efektif. Namun, dalam kasus ini, K-Means lebih cocok karena datanya lebih terstruktur.
- Untuk meningkatkan hasil K-Means, bisa dicoba mengubah jumlah cluster (K) atau menggunakan metode lain seperti Gaussian Mixture Model (GMM).

A. Model Klasifikasi dan Clustering yang terbaik:

Metode	Model Terbaik	Kelebihan	Kekurangan
Clustering	**K-Means**	Lebih stabil, lebih baik dalam mengelompokkan data	Memerlukan jumlah klaster (K) yang optimal
Klasifikasi	**Random Forest**	Akurasi tinggi (80.95%), lebih stabil dibandingkan Decision Tree	Butuh tuning untuk performa maksimal

3.5 K-Means dan Random Forest untuk Hubungan Internet Usage & Reading Frequency

Dari hasil analisa sebelumnya alogritma yang paling akurat adalah K-Means dan Random Forest. Jadi kami hanya menggunakan 2 model saja pada pengujian kali ini.

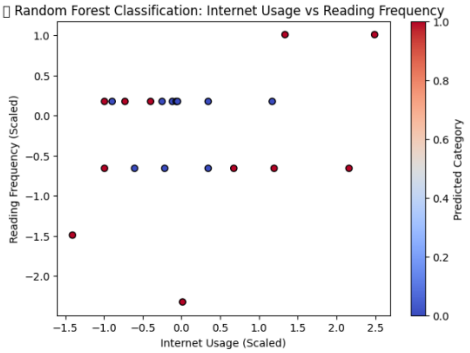


Gambar 8. K-Means Clustering Internet Usage v Reading Frequency

- K-Means clustering membagi data ke dalam tiga klaster berbeda yang ditunjukkan dengan warna yang berbeda.
- Setiap titik dalam grafik mewakili individu, dengan sumbu X menunjukkan intensitas penggunaan internet dan sumbu Y menunjukkan frekuensi membaca.
- Tanda "X" merah menunjukkan pusat klaster (centroid), yang merupakan titik rata-rata dari setiap klaster yang terbentuk.
- Klaster terlihat cukup jelas, meskipun beberapa titik berada di antara dua klaster yang berbeda, menandakan bahwa individu tersebut memiliki karakteristik campuran.
- Silhouette Score sebesar 0.3885 menunjukkan bahwa klaster memiliki tingkat pemisahan yang cukup baik, tetapi masih bisa ditingkatkan dengan menyesuaikan jumlah klaster (K).

Interpretasi K-Means:
Dari visualisasi ini, terlihat bahwa penggunaan internet memiliki hubungan dengan pola membaca, di mana individu dengan tingkat penggunaan internet lebih tinggi cenderung memiliki frekuensi membaca yang lebih tinggi.

A. Eksplorasi Hubungan Internet Usage & Reading Frequency Menggunakan Random Forest

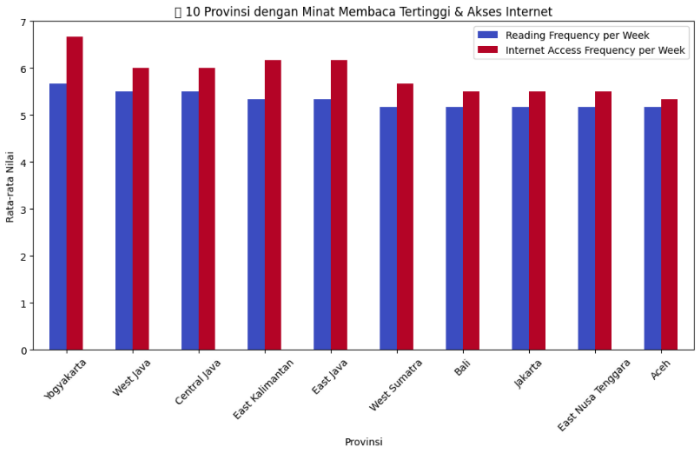


Gambar 9. Random Forest Internet Usage & Reading Frequency

- Random Forest digunakan untuk memprediksi kategori minat membaca berdasarkan tingkat penggunaan internet.
- Setiap titik dalam grafik menunjukkan individu, dengan warna menunjukkan kategori prediksi (minat membaca tinggi atau rendah).
- Pola warna yang berbeda menunjukkan bahwa model dapat membedakan individu yang memiliki minat membaca tinggi dari yang rendah.
- Model Random Forest memiliki akurasi sebesar 71.43%, yang menunjukkan bahwa model ini cukup efektif dalam memprediksi minat membaca berdasarkan internet usage.

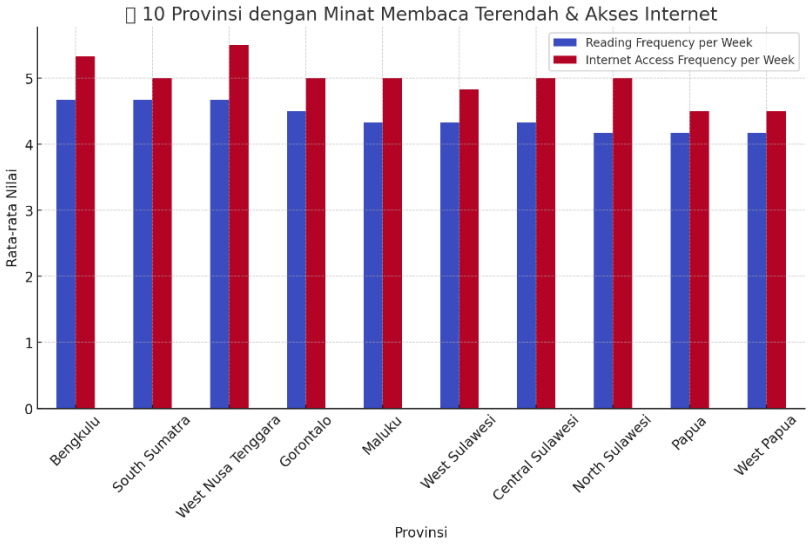
Interpretasi Random Forest:
Dari grafik feature importance, terlihat bahwa penggunaan internet memiliki pengaruh lebih besar terhadap minat membaca dibandingkan frekuensi membaca itu sendiri.
Ini bisa menunjukkan bahwa akses informasi digital dapat memainkan peran penting dalam membentuk kebiasaan membaca seseorang.

3.6 Analisa Provinsi dengan pembaca tertinggi dan terendah berdasarkan akses internet dan frekuensi membaca



Gambar 10. Provinsi dengan minat membaca Tertinggi dan Akses Internet

- Yogyakarta memiliki minat membaca tertinggi
  Yogyakarta menempati posisi tertinggi dalam frekuensi membaca per minggu, tetapi juga memiliki frekuensi akses internet yang sangat tinggi. Ini menunjukkan bahwa akses internet berperan dalam meningkatkan kebiasaan membaca.
- Sebagian besar provinsi menunjukkan pola yang mirip
  Di hampir semua provinsi, frekuensi akses internet lebih tinggi dibandingkan frekuensi membaca. Ini mengindikasikan bahwa meskipun orang sering mengakses internet, belum tentu digunakan sepenuhnya untuk membaca.
- Kesenjangan antara akses internet dan membaca
  Selisih antara akses internet dan membaca bervariasi. Di beberapa provinsi seperti West Java, Central Java, dan East Kalimantan, perbedaan antara kedua frekuensi relatif kecil. Di provinsi seperti Yogyakarta dan East Java, akses internet jauh lebih tinggi dibandingkan frekuensi membaca, yang mungkin menunjukkan bahwa internet lebih banyak digunakan untuk aktivitas lain selain membaca.
- Aceh memiliki keseimbangan antara akses internet dan membaca
  Provinsi Aceh menunjukkan keseimbangan yang lebih baik antara frekuensi membaca dan akses internet, yang berarti orang-orang di sana lebih cenderung memanfaatkan internet untuk membaca.



Gambar 11. Provinsi dengan minat baca terendah dan akses internet

- Akses internet lebih tinggi dibandingkan frekuensi membaca di semua provinsi
- Hampir di semua provinsi, frekuensi akses internet lebih tinggi dibandingkan frekuensi membaca.
- Ini menunjukkan bahwa meskipun masyarakat memiliki akses internet, tidak selalu dimanfaatkan untuk membaca.

Perbandingan Pola Minat Membaca dan Akses Internet

Kategori	Ciri-ciri provinsi dengan minat membaca tertinggi	Ciri-ciri provinsi dengan minat membaca terendah
Akses Internet	Rata-rata lebih tinggi, sering digunakan untuk aktivitas edukatif	Juga tinggi, tetapi lebih banyak digunakan untuk hiburan atau media sosial
Frekuensi Membaca	Lebih tinggi dibandingkan akses internet di beberapa daerah	Jauh lebih rendah dibandingkan akses internet
Pola Distribusi	Relatif merata, tidak ada kesenjangan besar antar provinsi	Beberapa provinsi menunjukkan kesenjangan besar antara akses internet dan membaca
Hubungan Internet & Membaca	Akses internet cenderung mendorong kebiasaan membaca	Akses internet belum mendorong peningkatan minat membaca secara signifikan
Wilayah Dominan	Yogyakarta, Jawa Tengah, Bali, Jakarta, East Java	West Nusa Tenggara, Sulawesi, Papua, Maluku, Bengkulu
Pemerataan minat antar Pulau	Pulau jawa dan bali dominan unggul minat membaca	Sedangkan pulau Sulawesi,Sumatera dan Papua masih tertinggal dalam minat membaca.

Tabel 1. Perbandingan Pola Minat Membaca dan Akses Internet

3.7 Perbandingan Frekuensi Membaca per Provinsi (2020 vs 2023)

Peningkatan Minat Membaca: Sebagian besar provinsi menunjukkan peningkatan frekuensi membaca dari tahun 2020 ke 2023, terlihat dari bar merah yang lebih tinggi dibandingkan bar biru. Beberapa Provinsi Mengalami Stagnasi atau Penurunan: Ada beberapa provinsi yang menunjukkan sedikit perubahan atau bahkan penurunan dalam kebiasaan membaca.

A. Provinsi dengan Peningkatan Signifikan
   Provinsi yang menunjukkan peningkatan signifikan dalam frekuensi membaca (bar merah jauh lebih tinggi dari bar biru) termasuk:
   - Yogyakarta
   - Papua
   - North Sumatra

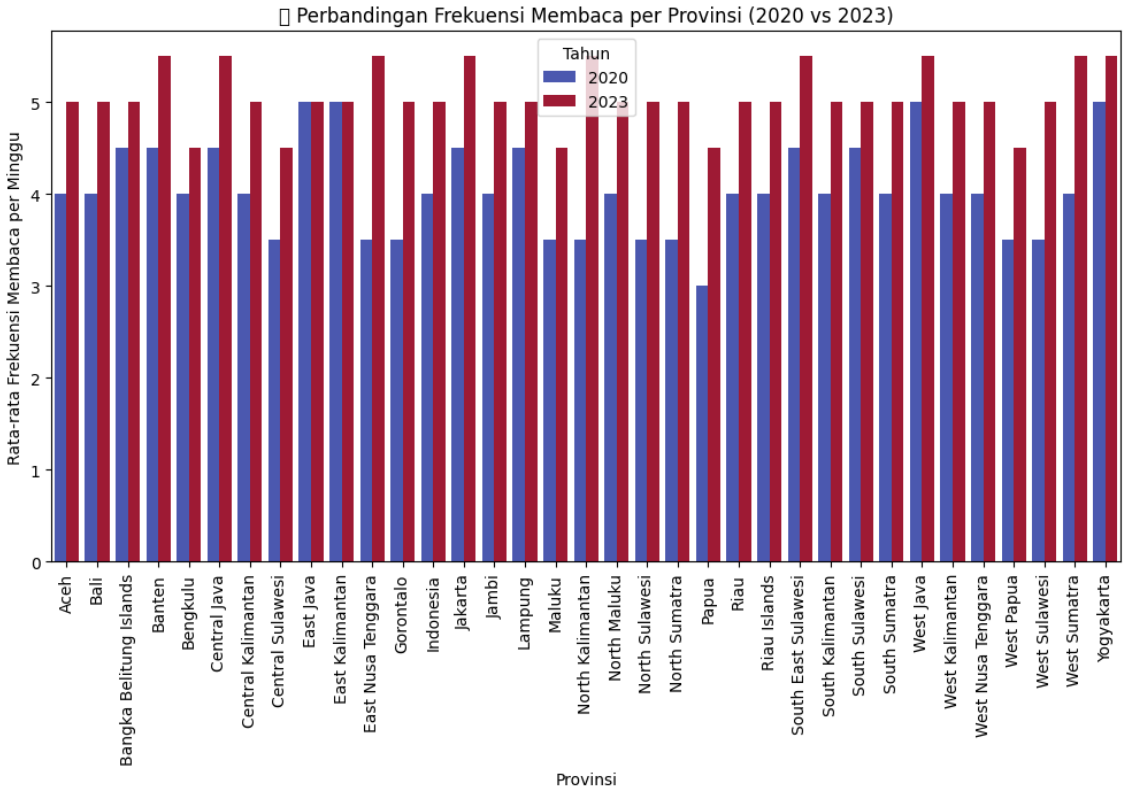
B. Kemungkinan faktor peningkatan:
   - Akses ke internet dan bahan bacaan digital meningkat.
   - Program literasi daerah yang lebih aktif sejak 2020.
   - Dampak pandemi yang mendorong lebih banyak kegiatan membaca di rumah.

C. Provinsi dengan Perubahan Kecil atau Penurunan
   Beberapa provinsi mengalami stagnasi atau sedikit perubahan dalam minat membaca, seperti:
   - Jakarta
   - Bali

- Bangka Belitung Islands

D. Kemungkinan penyebab stagnasi atau penurunan:
   - Perubahan gaya hidup masyarakat dengan lebih banyak aktivitas lain selain membaca.
   - Meningkatnya konsumsi konten digital yang tidak berbasis teks, seperti video dan podcast.
   - Kurangnya program literasi atau akses ke buku fisik di beberapa daerah.

Secara keseluruhan, terjadi peningkatan minat membaca secara nasional antara tahun 2020 dan 2023. Peningkatan ini bisa disebabkan oleh kemudahan akses ke bahan bacaan digital dan program literasi pemerintah. Namun, beberapa provinsi mengalami stagnasi, yang bisa menjadi indikasi bahwa perlu ada kebijakan literasi yang lebih efektif di daerah tersebut.



Gambar 12. Perbandingan Frekuensi Membaca 2020-2023

4. KESIMPULAN

4.1 Kesimpulan Hasil Clustering K-Means

- K-Means dengan K=4 memberikan hasil yang cukup baik dengan Silhouette Score sebesar 0.295.
- Klaster menunjukkan bahwa provinsi dengan minat membaca tinggi juga memiliki akses internet yang tinggi.
- K-Means berhasil mengelompokkan provinsi berdasarkan pola membaca yang serupa, tetapi masih ada tumpang tindih antar klaster.
- Kekurangan K-Means: Harus menentukan jumlah klaster (K) secara manual, sehingga pemilihan K yang optimal perlu diuji lebih lanjut.

4.2 Kesimpulan hasil clustering DBScan

- DBSCAN mampu mengidentifikasi outlier yang mungkin berasal dari provinsi dengan kondisi unik.
- DBSCAN memiliki Silhouette Score negatif (-0.0247), yang berarti hasil clustering tidak sebaik K-Means.
- Model ini kurang optimal untuk dataset ini karena data tidak memiliki pola kepadatan yang jelas.

4.3 Kesimpulan dari Klasifikasi (Decision Tree & Random Forest)

- Random Forest lebih akurat dibandingkan Decision Tree (80.95% vs 66.67%).
- Random Forest lebih stabil dan menangani data yang lebih kompleks dengan baik.
- Decision Tree lebih cenderung mengalami overfitting, sehingga kurang direkomendasikan untuk data ini.
- Faktor terbesar yang mempengaruhi minat membaca adalah akses internet dan frekuensi membaca sebelumnya.
- Akurasi tertinggi dicapai oleh Random Forest, menunjukkan bahwa pola membaca dapat diprediksi dengan cukup baik menggunakan model berbasis pohon keputusan.
- Kelemahan klasifikasi: Model ini masih bisa ditingkatkan dengan lebih banyak fitur atau teknik balancing data jika ada ketimpangan dalam jumlah data tiap kelas.

4.4 Kesimpulan dari Perbandingan Minat Membaca Antar Provinsi

A. Provinsi dengan Minat Membaca Tertinggi (2020–2023):
   - Yogyakarta, Jawa Barat, Jawa Tengah, Jawa Timur, dan Jakarta memiliki minat membaca tertinggi.
   - Pola ini menunjukkan bahwa provinsi di pulau jawa dengan pusat pendidikan besar cenderung memiliki minat membaca lebih tinggi.

B. Provinsi dengan Minat Membaca Terendah (2020–2023)
   - Papua, Papua Barat, Sulawesi Barat, Sulawesi Tengah, dan Maluku memiliki minat membaca terendah.
   - Provinsi dengan akses internet lebih rendah cenderung memiliki minat membaca yang lebih rendah juga.

C. Tren 2020 vs 2023
   - Sebagian besar provinsi mengalami peningkatan frekuensi membaca, terutama di provinsi dengan akses internet yang berkembang pesat.

- Beberapa provinsi mengalami stagnasi atau bahkan sedikit penurunan, yang mungkin terkait dengan faktor ekonomi atau kurangnya akses buku fisik.

4.5 Kelebihan dan Kekurangan dari Analisis Ini

A. Kelebihan:
   - Menggunakan metode clustering dan klasifikasi untuk memahami pola membaca secara objektif.
   - Dapat mengidentifikasi provinsi dengan pola membaca unik yang dapat digunakan untuk pengambilan kebijakan.
   - Analisis hubungan antara akses internet dan minat membaca menunjukkan dampak signifikan dari literasi digital.

B. Kekurangan:
   - Beberapa model clustering seperti DBSCAN tidak optimal untuk dataset ini.
   - Klasifikasi masih bisa ditingkatkan dengan teknik balancing data atau fitur tambahan.
   - Tidak semua faktor yang mempengaruhi minat membaca dapat diukur hanya dengan data ini (misalnya, faktor sosial & budaya juga berperan penting).

4.6 Hubungan antara Minat Membaca dan Akses Internet

- Provinsi dengan akses internet tinggi juga memiliki frekuensi membaca tinggi.
- Visualisasi menunjukkan bahwa internet memainkan peran penting dalam membangun kebiasaan membaca.
- Ketimpangan akses internet masih menjadi kendala, terutama di daerah-daerah dengan akses terbatas.
- Provinsi yang mengalami peningkatan akses internet sejak 2020 cenderung memiliki peningkatan dalam minat membaca.
- Penggunaan teknologi digital untuk membaca (e-books, berita online, jurnal akademik) semakin meningkat.
- Di daerah dengan internet terbatas, kebiasaan membaca masih bergantung pada media cetak, yang bisa menjadi tantangan logistik.

Secara umum, terjadi peningkatan minat membaca dari tahun 2020 ke 2023, terutama di provinsi dengan akses internet yang lebih baik. Namun, masih ada ketimpangan antara daerah dengan infrastruktur digital yang berkembang dan daerah dengan akses terbatas. Kebijakan peningkatan literasi digital dan akses ke sumber bacaan sangat diperlukan untuk meningkatkan kebiasaan membaca di seluruh Indonesia.

5. SARAN

- Optimasi Clustering: Gunakan Elbow Method untuk menemukan jumlah klaster optimal dalam K-Means.
- Model Klasifikasi: Lakukan Hyperparameter Tuning pada Random Forest untuk meningkatkan akurasi.
- Faktor Tambahan: Tambahkan variabel usia, jenis bacaan, dan konsumsi media digital untuk memahami pola membaca lebih dalam.
- Analisis Hubungan Ekonomi & Literasi: Bandingkan pendapatan dan pendidikan dengan minat membaca untuk melihat dampaknya.
- Perluasan Akses Internet: Prioritaskan daerah dengan minat membaca rendah untuk meningkatkan literasi digital.
- Distribusi Buku & Perpustakaan Digital: Perkuat platform seperti iPusnas dan program distribusi buku ke daerah terpencil.
- Kampanye Literasi Digital: Gunakan media sosial & tantangan membaca untuk meningkatkan minat membaca generasi muda.
- Integrasi Literasi dalam Kurikulum Sekolah: Masukkan materi membaca digital & keterampilan literasi media dalam pendidikan dasar.

Penelitian ini menemukan bahwa akses internet berhubungan erat dengan minat membaca di Indonesia. Model clustering terbaik adalah K-Means dan model klasifikasi terbaik adalah Random Forest. Untuk penelitian lebih lanjut, analisis dapat diperluas dengan variabel seperti usia dan jenis bacaan. Kebijakan literasi digital, distribusi buku, dan perluasan internet dapat membantu meningkatkan minat membaca.

DAFTAR PUSTAKA

[1] Aditia Space. 2024. *Indonesia Reading Interest 2020-2023* https://www.kaggle.com/datasets/imaditia/indonesia-reading-interest-2020-2023

[2] Kepala Pusat Analisis Perpustakaan dan Pengembangan Budaya Baca. 2024. *Tingkat Kegemaran Membaca (TGM) 2023* . https://dev.perpusnas.go.id/dataset/tingkat-kegemaran-membaca-tgm2023

[3] Alfa Aulia Nooraya. 2024. *Indeks Tingkat Kegemaran Membaca di Indonesia Tahun 2021-2023.* https://data.goodstats.id/statistic/indeks-tingkat-kegemaran-membaca-di-indonesia-tahun-2021-2023-NU0AN

[4] Kementerian Komunikasi dan Informatika (Kominfo). 2022. *Pengaruh Akses Internet terhadap Literasi Digital dan Minat Membaca di Indonesia.* https://www.kominfo.go.id/publications/literasi-digital-dan-minat-membaca

[5] Breiman, L. (2001). *Random Forests. Machine Learning, 45(1), 5–32.* https://doi.org/10.1023/A:1010933404324

[6] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226-231.* https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[7] Quinlan, J. R. (1986). *Induction of Decision Trees.* Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

[8] MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations.* Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1(281–297). https://projecteuclid.org/euclid.bsmsp/1200512992