LAPORAN AKHIR PROYEK

BIG DATA DAN VISUALISASI

Judul

Analisis Perilaku Pelanggan untuk Rekomendasi Produk Berdasarkan Dataset Amazon

# **ABSTRAK**

Pertumbuhan pesat sektor *e-commerce* telah menghasilkan volume data yang sangat besar, memberikan peluang untuk menggali wawasan berharga mengenai perilaku pelanggan. Proyek ini bertujuan untuk mengimplementasikan konsep Big Data dan visualisasi untuk menganalisis dataset penjualan dan ulasan produk Amazon. Metodologi yang digunakan meliputi pra-pemrosesan data untuk membersihkan dan mentransformasi data mentah, diikuti dengan analisis data eksploratif (EDA) untuk mengidentifikasi pola, seperti kategori produk terpopuler dan distribusi rating. Berdasarkan analisis ini, sebuah sistem rekomendasi sederhana berbasis popularitas dikembangkan untuk menyarankan produk yang memiliki rating tinggi dan jumlah ulasan signifikan. Seluruh hasil analisis dan sistem rekomendasi diintegrasikan ke dalam sebuah aplikasi web interaktif yang dibangun menggunakan *framework* Flask. Hasil proyek menunjukkan bahwa kategori seperti *Electronics* dan *Computers & Accessories* mendominasi pasar, dan mayoritas produk memiliki rating positif. Aplikasi web yang dihasilkan berhasil menyajikan visualisasi data dan rekomendasi produk secara efektif, membuktikan kelayakan pemanfaatan data untuk pengambilan keputusan bisnis strategis di lingkungan *e-commerce*.

**BAB 1: PENDAHULUAN**

**1.1 Latar Belakang**

Era digital telah mengubah cara konsumen berbelanja, dengan *e-commerce* menjadi salah satu sektor industri dengan pertumbuhan tercepat di dunia. Platform seperti Amazon, sebagai pemimpin pasar global, memproses jutaan transaksi dan interaksi pengguna setiap hari. Setiap klik, pencarian, ulasan, dan pembelian menghasilkan jejak data yang, jika dianalisis dengan benar, dapat menjadi aset strategis yang sangat berharga. Fenomena ini dikenal sebagai Big Data, yang ditandai oleh volume (jumlah data), *velocity* (kecepatan data dihasilkan), dan *variety* (keragaman jenis data).

Bagi perusahaan *e-commerce*, kemampuan untuk memahami perilaku pelanggan adalah kunci untuk meningkatkan kepuasan, loyalitas, dan pada akhirnya, pendapatan. Dengan menganalisis data transaksi dan ulasan, perusahaan dapat mempersonalisasi pengalaman belanja, mengoptimalkan inventaris, dan memberikan rekomendasi produk yang relevan. Proyek ini berfokus pada pemanfaatan dataset publik dari Amazon untuk mensimulasikan proses analisis tersebut, dari data mentah hingga menjadi sebuah aplikasi yang memberikan wawasan bisnis.

**1.2 Rumusan Masalah**

Berdasarkan latar belakang di atas, rumusan masalah untuk proyek ini adalah sebagai berikut:

1. Bagaimana pola perilaku pelanggan dapat diidentifikasi dari data penjualan dan ulasan produk pada dataset Amazon?
2. Bagaimana cara membangun sistem rekomendasi produk yang sederhana namun efektif berdasarkan metrik popularitas seperti rating dan jumlah ulasan?
3. Bagaimana hasil analisis data dan rekomendasi produk dapat disajikan secara efektif dan interaktif kepada pemangku kepentingan bisnis melalui sebuah aplikasi web?

**1.3 Tujuan Proyek**

Proyek ini bertujuan untuk:

1. Melakukan analisis data eksploratif (EDA) pada dataset Amazon untuk mengidentifikasi kategori produk terlaris, distribusi rating, dan pola-pola lainnya.
2. Mengimplementasikan sistem rekomendasi produk sederhana berbasis popularitas.
3. Mengembangkan aplikasi web menggunakan Python dan Flask sebagai *dashboard* untuk menampilkan visualisasi data yang informatif dan hasil rekomendasi produk.
4. Mendemonstrasikan penerapan praktis dari konsep Big Data dan visualisasi dalam konteks *e-commerce*.

**BAB 2: TINJAUAN PUSTAKA**

**2.1 Konsep Big Data di E-commerce**

Big Data merujuk pada kumpulan data yang sangat besar dan kompleks sehingga sulit untuk ditangani dengan perangkat lunak pengolah data tradisional. Dalam konteks *e-commerce*, Big Data mencakup data transaksi, data demografis pengguna, data navigasi situs (jejak klik), ulasan produk, dan interaksi media sosial. Pemanfaatan Big Data memungkinkan perusahaan *e-commerce* untuk melakukan:

1. **Segmentasi Pelanggan:** Mengelompokkan pelanggan berdasarkan perilaku pembelian untuk kampanye pemasaran yang lebih tertarget.
2. **Analisis Sentimen:** Menganalisis teks ulasan untuk memahami persepsi publik terhadap suatu produk atau layanan.
3. **Optimisasi Harga:** Menyesuaikan harga secara dinamis berdasarkan permintaan, persaingan, dan perilaku pelanggan.
4. **Sistem Rekomendasi:** Menyarankan produk yang relevan kepada pengguna untuk meningkatkan *cross-selling* dan *up-selling*.

**2.2 Sistem Rekomendasi**

Sistem rekomendasi adalah subkelas dari sistem penyaringan informasi yang berupaya memprediksi 'peringkat' atau 'preferensi' yang akan diberikan pengguna terhadap suatu item. Terdapat beberapa pendekatan utama:

1. **Collaborative Filtering:** Pendekatan ini merekomendasikan item berdasarkan preferensi dari pengguna lain yang memiliki selera serupa.
2. **Content-Based Filtering:** Merekomendasikan item yang memiliki atribut serupa dengan item yang disukai pengguna di masa lalu.
3. **Hybrid Filtering:** Menggabungkan beberapa pendekatan (misalnya, *collaborative* dan *content-based*) untuk mengatasi kelemahan masing-masing.
4. **Popularity-Based Filtering:** Pendekatan non-personalisasi yang merekomendasikan item paling populer kepada semua pengguna. Meskipun sederhana, metode ini efektif untuk pengguna baru (*cold start problem*) dan sebagai *baseline*.

**2.3 Algoritma yang Digunakan**

Dalam proyek ini, sistem rekomendasi yang dibangun menggunakan pendekatan **Popularity-Based Filtering**. Ini bukan merupakan algoritma *machine learning* yang kompleks, melainkan sebuah metode heuristik yang efektif. Logikanya adalah produk yang disukai oleh banyak orang (memiliki rating tinggi dan banyak ulasan) kemungkinan besar juga akan disukai oleh pengguna lain. Pemeringkatan popularitas dilakukan dengan mengurutkan produk berdasarkan dua kriteria utama:

1. **Rating (rating):** Nilai rata-rata yang diberikan oleh pengguna.
2. **Jumlah Ulasan (rating\_count):** Jumlah total pengguna yang telah memberikan ulasan, yang berfungsi sebagai proksi untuk volume penjualan dan kepercayaan.

Produk diurutkan secara menurun berdasarkan rating dan kemudian rating\_count untuk memastikan produk yang direkomendasikan tidak hanya memiliki rating tinggi tetapi juga telah terbukti populer di kalangan banyak pembeli.

**BAB 3: METODOLOGI PENELITIAN**

**3.1 Deskripsi Dataset**

Dataset yang digunakan dalam proyek ini adalah "Amazon Sales Dataset" yang diperoleh dari platform Kaggle. Dataset ini berisi sekitar 1,4 juta baris data (setelah pembersihan) yang mencakup berbagai produk yang dijual di Amazon India. Kolom-kolom utama yang digunakan dalam analisis adalah:

1. product\_name: Nama produk.
2. category: Kategori produk, seringkali dalam format bersarang.
3. discounted\_price: Harga produk setelah diskon.
4. actual\_price: Harga asli produk.
5. rating: Rata-rata rating produk (skala 1-5).
6. rating\_count: Jumlah ulasan yang diterima produk.
7. user\_id: ID unik untuk pengguna yang memberikan ulasan.

Tahap pra-pemrosesan data mencakup pembersihan kolom harga dari simbol mata uang ('₹') dan koma, konversi tipe data ke format numerik, penanganan nilai yang hilang, dan ekstraksi kategori utama dari kolom category.

**3.2 Arsitektur Sistem**

Arsitektur sistem proyek ini dirancang secara modular dan dapat diuraikan dalam alur kerja berikut:

1. **Pengambilan & Pra-pemrosesan Data:** Skrip Python (analysis\_lib.py) memuat file amazon.csv menggunakan Pandas, melakukan semua langkah pembersihan data yang diperlukan, dan menyiapkan DataFrame yang siap untuk dianalisis.
2. **Analisis & Visualisasi:** Skrip yang sama melakukan analisis data eksploratif untuk menemukan wawasan. Hasilnya divisualisasikan menggunakan Matplotlib dan Seaborn, kemudian disimpan sebagai file gambar statis (PNG).
3. **Mesin Rekomendasi:** Sebuah fungsi di analysis\_lib.py mengimplementasikan logika rekomendasi berbasis popularitas dengan memfilter dan mengurutkan DataFrame, lalu mengembalikan daftar produk teratas.
4. **Backend Aplikasi Web:** File app.py berfungsi sebagai server web Flask. Saat aplikasi dijalankan, ia memanggil fungsi-fungsi dari analysis\_lib.py untuk memproses data, menghasilkan visualisasi, dan mendapatkan rekomendasi.
5. **Frontend Aplikasi Web:** app.py meneruskan hasil (path gambar dan daftar rekomendasi) ke template index.html. Template ini menggunakan sintaks Jinja2 untuk secara dinamis menampilkan visualisasi dan tabel rekomendasi di browser pengguna.

**3.3 Alat dan Teknologi yang Digunakan**

1. **Bahasa Pemrograman:** Python 3.9
2. **Library Analisis & Manipulasi Data:** Pandas
3. **Library Visualisasi Data:** Matplotlib, Seaborn
4. **Framework Web:** Flask
5. **Lingkungan Pengembangan:** Visual Studio Code
6. **Browser Web:** Google Chrome (untuk menampilkan hasil akhir)

**BAB 4: HASIL DAN ANALISIS**

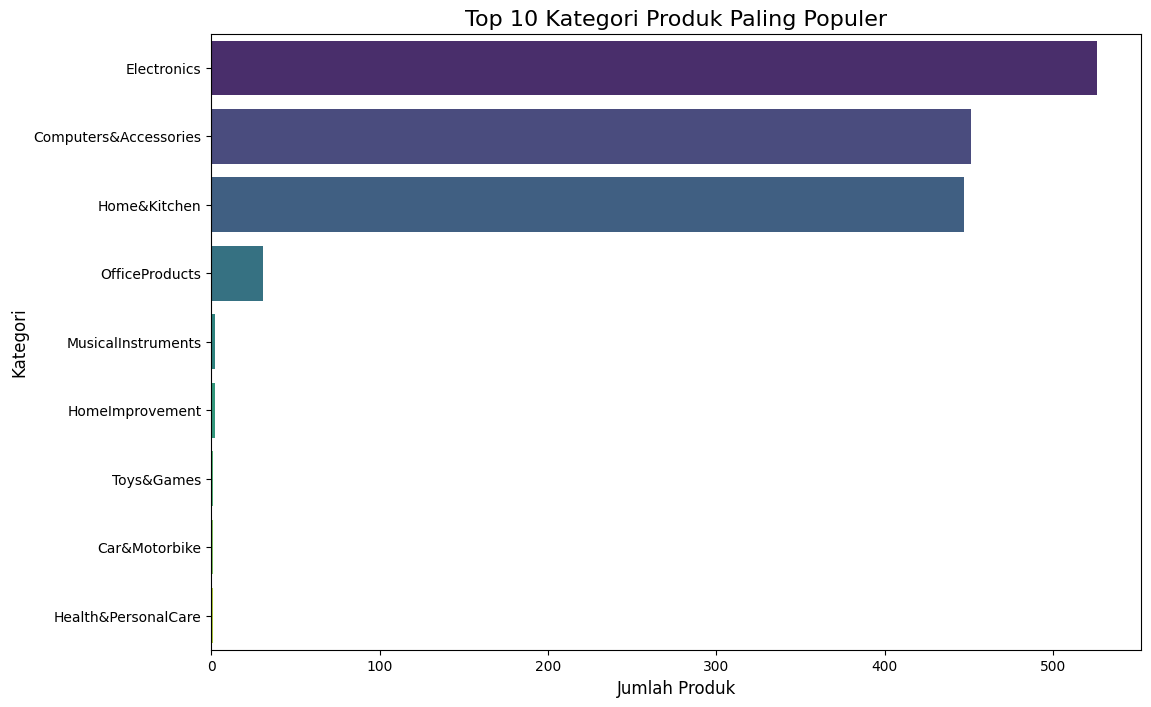
**4.1 Hasil Analisis Data Eksploratif (EDA)**

Analisis pada dataset mengungkapkan beberapa wawasan kunci:

1. **Dominasi Kategori Tertentu:** Ditemukan bahwa kategori seperti Electronics, Computers & Accessories, dan Home & Kitchen merupakan tiga kategori dengan jumlah produk terbanyak di dalam dataset.
2. **Distribusi Rating Positif:** Mayoritas produk memiliki rating yang sangat baik. Analisis distribusi menunjukkan puncak frekuensi berada di sekitar rating 4.0 hingga 4.5, mengindikasikan kepuasan pelanggan yang umumnya tinggi atau potensi bias dalam pengumpulan data ulasan.
3. **Korelasi Harga dan Rating:** Tidak ditemukan korelasi yang kuat antara harga produk (actual\_price) dan ratingnya. Produk dengan harga lebih tinggi tidak selalu mendapatkan rating yang lebih baik, dan sebaliknya.

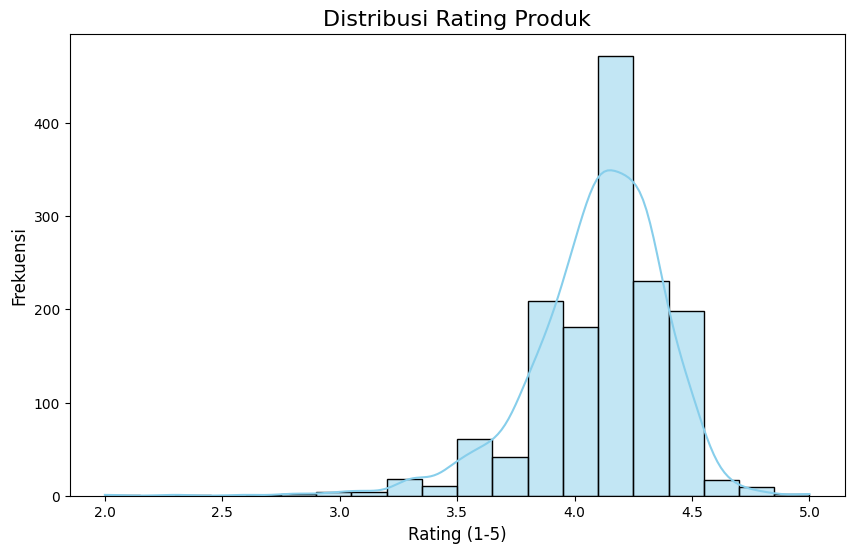
**4.2 Visualisasi Data**

Untuk menyajikan temuan EDA secara visual, dua grafik utama dibuat:



**Gambar 4.1**  *Grafik batang Top 10 Kategori Produk*

**Grafik 1: Top 10 Kategori Produk:** Grafik batang ini (Gambar 4.1) secara visual mengkonfirmasi kategori produk mana yang paling dominan, memberikan gambaran cepat mengenai fokus pasar dalam dataset ini.

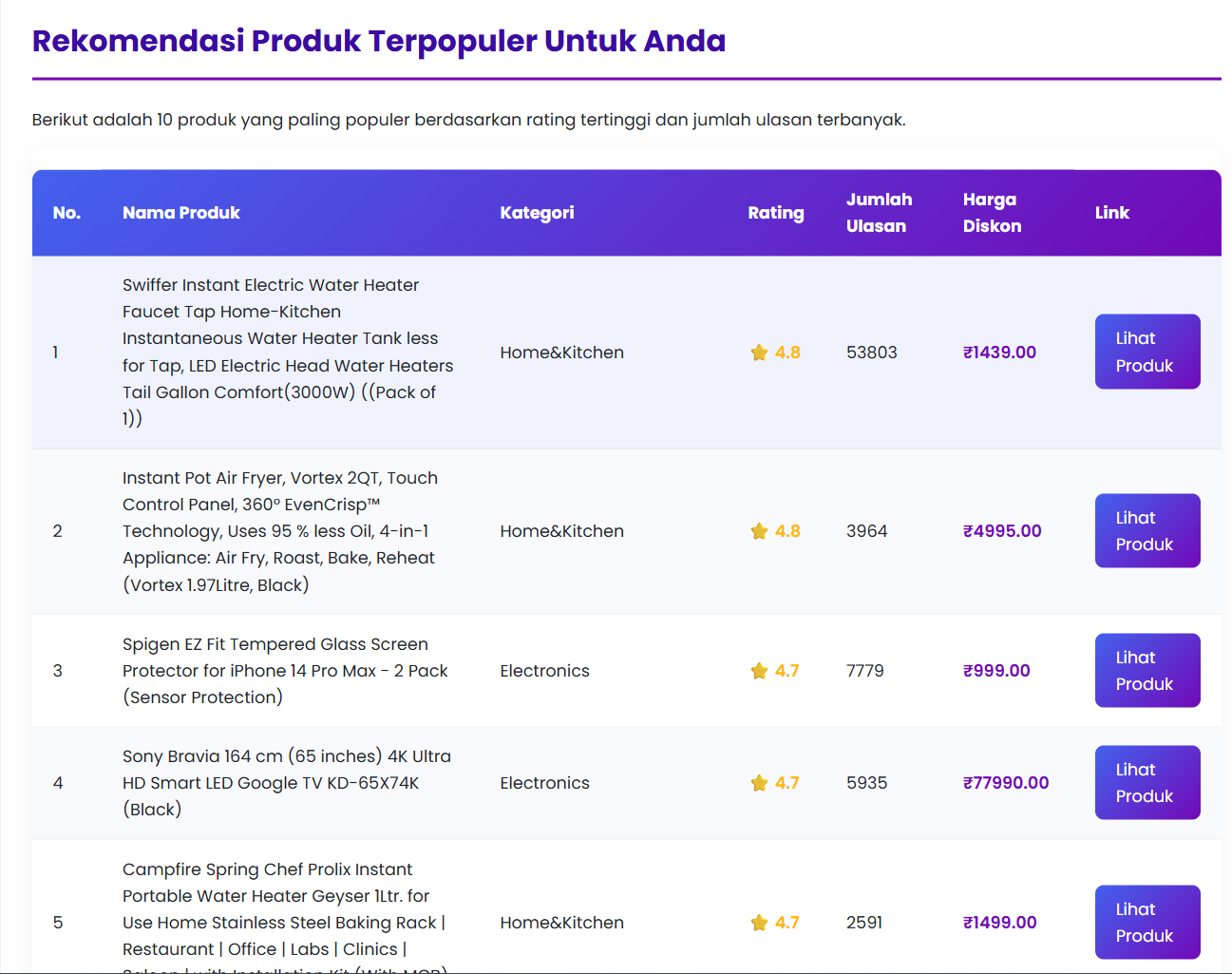


**Gambar 4.2** *Histogram Distribusi Rating Produk*

**Grafik 2: Distribusi Rating Produk:** Histogram ini (Gambar 4.2) menunjukkan sebaran rating di semua produk. Puncak yang tajam di sisi kanan skala menggarisbawahi tren rating positif yang telah disebutkan.

**4.3 Hasil Implementasi Sistem Rekomendasi**

Sistem rekomendasi berbasis popularitas berhasil diimplementasikan. Sistem ini memfilter produk yang memiliki setidaknya 100 ulasan, kemudian mengurutkannya berdasarkan rating tertinggi dan jumlah ulasan terbanyak. Hasilnya adalah daftar 10 produk yang paling "aman" untuk direkomendasikan karena telah terbukti disukai oleh banyak orang.



**Gambar 4.3** *Output Tabel Rekomendasi Produk Terpopuler*

Contoh output rekomendasi disajikan dalam format tabel di aplikasi web, menampilkan nama produk, kategori, rating, jumlah ulasan, harga, dan tautan langsung ke halaman produk.

**4.4 Tampilan Aplikasi Web**

Hasil akhir dari proyek ini adalah sebuah *dashboard* web interaktif (Gambar 4.3). Halaman web dirancang dengan struktur yang bersih dan profesional, terdiri dari tiga bagian utama:

1. **Header:** Judul proyek dan deskripsi singkat.
2. **Bagian Analisis Visual:** Menampilkan dua plot yang dihasilkan dari EDA, memberikan ringkasan visual instan kepada pengguna.
3. **Bagian Rekomendasi:** Menampilkan tabel berisi 10 produk yang direkomendasikan, lengkap dengan detail relevan dan tautan yang dapat diklik.

Aplikasi ini berhasil membuktikan bagaimana data mentah dapat diubah menjadi alat pendukung keputusan yang interaktif dan mudah dipahami.

**BAB 5: KESIMPULAN DAN SARAN**

**5.1 Ringkasan Hasil**

Proyek ini telah berhasil mencapai semua tujuan yang ditetapkan. Kami berhasil melakukan pra-pemrosesan dan analisis pada dataset Amazon yang besar, mengidentifikasi pola perilaku pelanggan melalui EDA, dan membuat visualisasi data yang informatif. Sistem rekomendasi sederhana berbasis popularitas berhasil dikembangkan dan diintegrasikan ke dalam sebuah aplikasi web fungsional menggunakan Flask. Proyek ini secara efektif mendemonstrasikan alur kerja lengkap dari pemrosesan Big Data, mulai dari analisis hingga presentasi hasil dalam format yang dapat diakses oleh pemangku kepentingan bisnis.

**5.2 Keterbatasan Proyek**

Meskipun berhasil, proyek ini memiliki beberapa keterbatasan:

1. **Rekomendasi Non-Personalisasi:** Sistem rekomendasi yang digunakan bersifat umum dan memberikan hasil yang sama untuk semua pengguna. Sistem ini tidak mempertimbangkan preferensi individu.
2. **Analisis Teks Terbatas:** Proyek ini tidak menggali lebih dalam analisis sentimen dari konten ulasan (review\_content), yang dapat memberikan wawasan lebih kaya tentang "mengapa" pelanggan menyukai atau tidak menyukai suatu produk.
3. **Dataset Statis:** Dataset yang digunakan bersifat statis dan mungkin tidak mencerminkan tren pasar terbaru.

**5.3 Saran untuk Pengembangan Selanjutnya**

Berdasarkan keterbatasan yang ada, pengembangan di masa depan dapat berfokus pada:

1. **Implementasi Rekomendasi Personal:** Mengembangkan sistem rekomendasi yang lebih canggih menggunakan *Collaborative Filtering* (berbasis kemiripan pengguna) atau *Content-Based Filtering*.
2. **Analisis Sentimen (NLP):** Menerapkan teknik *Natural Language Processing* (NLP) pada kolom ulasan untuk mengklasifikasikan sentimen (positif, negatif, netral) dan mengekstrak topik utama.
3. **Pengembangan Dashboard yang Lebih Interaktif:** Menggunakan library visualisasi berbasis JavaScript (seperti D3.js atau Chart.js) yang terintegrasi dengan Flask untuk memungkinkan pengguna memfilter data secara *real-time*.
4. **Skalabilitas:** Memigrasikan sistem ke arsitektur berbasis cloud dan menggunakan alat seperti Apache Spark untuk memproses data yang jauh lebih besar dan mendekati *real-time*.

**DAFTAR PUSTAKA**

1. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
2. Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.
3. Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media.
4. Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

**LAMPIRAN**

**Kode Program Utama**

* app.py

(Salin-tempel kode lengkap dari file app.py di sini)

* analysis\_lib.py

(Salin-tempel kode lengkap dari file analysis\_lib.py di sini)

* templates/index.html

(Salin-tempel kode lengkap dari file index.html di sini)

* static/css/style.css

(Salin-tempel kode lengkap dari file style.css di sini)