PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 1 No. 1 (2025) pp: 108-112
P-ISSN: XXXX-XXXX, e-ISSN: XXXX-XXXX

Analisis Sentimen Risiko Serangan Jantung Menggunakan *K-Means Clustering* Dengan *RapidMiner*

[1]Ade Kurniaty, [2]Adi Muslim, [3]Aryazeyla Rachayudiza, [4]Bima Aditiya Milano, [5]Diana Manullang, [6]Maulana Fansyuri

[1]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[2]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[3]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[4]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[5]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
[6]Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

E-mail: adekurniaty.ak@gmail.com*, adhimuslim201@gmail.com, zeylarachayudiza05@gmail.com, bima.aditiya30@gmail.com, dianamanullang06@gmail.com.

Abstract

This study discusses the analysis of patient grouping based on heart attack risk by applying the K-Means Clustering algorithm using RapidMiner software. In this modern era, patient health data is very important for early identification and prevention of serious diseases such as heart attack. To understand the patterns of patient characteristics related to this risk, a clustering process was carried out on a heart attack risk dataset obtained from Kaggle, consisting of 8,763 patient data entries. The research stages began with data collection, data preprocessing, and the implementation of the K-Means algorithm with a certain number of clusters (e.g., three), which will group patients based on their risk profiles (e.g., low, moderate, and high risk). The research results are expected to show the distribution of patient data into these clusters, for example, how many patients fall into the high, moderate, and low-risk clusters. With these results, the K-Means algorithm proves effective in identifying groups of patients with similar characteristics, as well as providing useful insights for early detection and intervention of heart attack risk automatically. This research is expected to serve as a basis for the development of a more accurate and adaptive risk identification system for the dynamics of health data.

Keywords: Data Mining, K-Means, RapidMiner, Heart Attack Risk, Patient Grouping

Abstrak

Penelitian ini membahas analisis pengelompokan pasien berdasarkan risiko serangan jantung dengan menerapkan algoritma *K-Means* Clustering menggunakan perangkat lunak *RapidMiner*. Di era modern ini, data kesehatan pasien menjadi sangat penting untuk identifikasi dini dan pencegahan risiko penyakit serius seperti serangan jantung. Untuk memahami pola-pola karakteristik pasien yang berkaitan dengan risiko ini, dilakukan proses clustering terhadap dataset risiko serangan jantung yang diperoleh dari *Kaggle*, terdiri dari 8.763 entri data pasien. Tahapan penelitian dimulai dari pengumpulan data, pra-pemrosesan data, hingga implementasi algoritma *K-Means* dengan jumlah klaster tertentu (misalnya, tiga), yang akan mengelompokkan pasien berdasarkan profil risiko mereka (misalnya, risiko rendah, sedang, dan tinggi). Hasil penelitian diharapkan dapat menunjukkan distribusi data pasien ke dalam klaster-klaster tersebut, misalnya berapa banyak pasien yang tergolong dalam klaster risiko tinggi, sedang, dan rendah. Dengan hasil ini, algoritma *K-Means* terbukti efektif dalam mengidentifikasi kelompok pasien dengan karakteristik serupa, serta memberikan gambaran yang bermanfaat dalam upaya deteksi dini dan intervensi terhadap risiko serangan jantung secara otomatis. Penelitian ini diharapkan dapat menjadi dasar untuk pengembangan sistem identifikasi risiko yang lebih akurat dan adaptif terhadap dinamika data kesehatan.

Kata Kunci: *Data Mining*, *K-Means*, *RapidMiner*, Risiko Serangan Jantung, Pengelompokan Pasien

A. PENDAHULUAN

Diterima Redaksi: 19-06-2025 | Selesai Revisi: 25-06-2025 | Diterbitkan Online: 30-06-2025

Di era modern saat ini, data menjadi aset yang sangat berharga di berbagai sektor, termasuk dalam bidang kesehatan. Kemajuan teknologi memungkinkan pengumpulan data medis yang masif dari berbagai sumber, seperti rekam medis elektronik, perangkat wearable, hingga hasil pemeriksaan laboratorium. Data ini bukan lagi sekadar catatan historis, melainkan telah menjadi kunci untuk memahami pola penyakit, mengidentifikasi faktor risiko, dan meningkatkan kualitas pelayanan kesehatan.

Penyakit jantung, khususnya serangan jantung, masih menjadi salah satu penyebab utama morbiditas dan mortalitas global. Pencegahan dini dan manajemen risiko yang efektif sangat krusial untuk mengurangi dampak penyakit ini. Namun, identifikasi individu yang berisiko tinggi seringkali kompleks karena melibatkan interaksi berbagai faktor demografi, gaya hidup, dan kondisi klinis. Memahami kelompok-kelompok pasien dengan profil risiko yang serupa dapat membantu tenaga medis dalam merancang intervensi yang lebih personal dan tepat sasaran. (Rahayu et al., 2020)

Dalam konteks inilah, analisis pengelompokan (clustering) menjadi sangat penting untuk menggali wawasan tersembunyi dari data pasien. Salah satu metode yang efektif untuk mengelompokkan data berdasarkan kemiripan karakteristik adalah algoritma *K-Means* Clustering. Algoritma ini memungkinkan identifikasi kelompok-kelompok pasien yang memiliki kesamaan dalam fitur-fitur risiko serangan jantung. Proses analisis ini akan dilakukan menggunakan *RapidMiner*, sebuah platform data science berbasis visual yang memudahkan pengguna dalam melakukan pra-pemrosesan data, clustering, hingga visualisasi hasil analisis. Penelitian ini bertujuan untuk mengidentifikasi dan mengelompokkan pasien berdasarkan profil risiko serangan jantung mereka, memberikan dasar yang lebih kuat untuk strategi pencegahan dan penanganan.

B. METODE

Dalam konteks penelitian ini, teknik clustering digunakan untuk mengelompokkan data pasien risiko serangan jantung berdasarkan kemiripan karakteristik mereka, di mana algoritma *K-Means* menjadi salah satu metode yang paling umum digunakan karena efisiensinya dalam pengelompokan data berskala besar. Penelitian ini didasarkan pada beberapa konsep dan teori yang relevan, antara lain mengenai data mining, clustering, algoritma *K-Means*, serta pemanfaatan platform *Kaggle* sebagai sumber data. Berikut teori metode yang akan digunakan untuk penelitian ini.

1. Data Mining

Data mining adalah proses penggalian informasi atau pengetahuan yang bermanfaat dari kumpulan data dalam jumlah besar dengan cara menemukan pola, hubungan, atau tren yang tersembunyi di dalamnya. Data mining, atau terkadang dikenal dengan *knowledge discovery in databases* (KDD), yakni aktivitast erkait pengumpulan data, penggunaan data historis guna menemukan pengetahuan, informasi, pola, ataupun kaitan di data besar (Handayani, 2022)

2. *Kaggle*

Kaggle adalah sebuah platform online berbasis komunitas yang bergerak di bidang data science dan machine learning, yang memungkinkan para penggunanya untuk menemukan dan berbagi kumpulan data, membangun model prediktif, serta mengikuti kompetisi analisis data secara global. Didirikan pada tahun 2010 dan diakuisisi oleh Google pada tahun 2017, *Kaggle* telah menjadi ruang belajar dan eksperimen bagi jutaan ilmuwan data, peneliti, praktisi, dan mahasiswa dari berbagai belahan dunia.

3. *K-Means*

K-Means merupakan teknik clustering yang diperoleh dari sebuah dataset dengan cara menghitung jarak dari setiap titik ke pusat cluster secara iterative (Hani, 2022). Proses ini dimulai dengan menentukan jumlah klaster (K) yang diinginkan, kemudian memilih pusat awal (centroid) secara acak. Setiap data kemudian dikaitkan ke centroid terdekat berdasarkan jarak Euclidean atau metrik lainnya. Setelah semua data terkelompok, centroid akan diperbarui berdasarkan rata-rata posisi data dalam setiap klaster.

4. Clustering

Clustering adalah salah satu teknik dalam data mining dan machine learning yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok (klaster) berdasarkan (Kodinariya & Makwana, 2013) kemiripan atau kesamaan karakteristik antar data. Tujuannya adalah agar data dalam satu klaster memiliki tingkat kemiripan yang tinggi satu sama lain, sementara data yang berada di klaster yang berbeda memiliki perbedaan yang signifikan.

5. *K-Means* Clustering

K-Means Clustering adalah algoritma unsupervised learning yang digunakan untuk mengelompokkan data ke dalam beberapa klaster berdasarkan kemiripan. Algoritma ini bekerja dengan menentukan jumlah klaster (K), lalu secara iteratif membagi data ke klaster terdekat berdasarkan jarak ke pusat klaster (centroid), dan memperbarui posisi centroid hingga hasilnya stabil. Tujuannya adalah agar data dalam satu klaster saling mirip dan berbeda dengan data di klaster lain. *K-Means* sering digunakan dalam analisis data pelanggan, dokumen, dan media sosial.

Dalam penelitian ini, tahapan prosedural digambarkan melalui diagram alur yang merepresentasikan rangkaian kegiatan penelitian dari tahap awal hingga akhir. Diagram ini bertujuan untuk memberikan gambaran yang lebih sistematis dan mempermudah pemahaman terhadap setiap langkah yang dilakukan selama proses penelitian.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

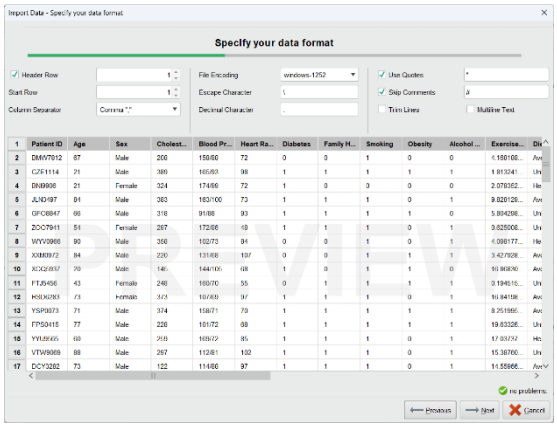
Gambar 1 Alur Metodologi Penelitian

C. HASIL DAN PEMBAHASAN

Setelah melalui rangkaian tahapan penelitian yang dimulai dari studi literatur hingga implementasi algoritma *K-Means* menggunakan *RapidMiner*, diperoleh hasil pengelompokan data pasien risiko serangan jantung yang telah dianalisis lebih lanjut. Data yang diperoleh dari *Kaggle* terlebih dahulu diproses sebelum dimasukkan ke dalam model clustering. Proses ini menghasilkan beberapa klaster yang merepresentasikan pola-pola karakteristik pasien berdasarkan kemiripan atribut dan profil kesehatan. Pada tahap ini, hasil *clustering* dianalisis untuk melihat kecenderungan fitur pada masing-masing klaster, serta dilakukan interpretasi visual melalui grafik atau diagram untuk mendukung pemahaman yang lebih komprehensif terhadap hasil yang diperoleh.

Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 8.763 data entri, memuat informasi komprehensif terkait profil kesehatan pasien yang relevan dengan risiko serangan jantung. Setiap entri data pasien dicirikan oleh sejumlah atribut utama. Atribut-atribut tersebut meliputi ID Pasien (Patient ID) sebagai identifikasi unik untuk setiap individu, Usia (Age) pasien dalam tahun, serta Jenis Kelamin (Sex) biologis. Data klinis penting lainnya mencakup Kolesterol (Cholesterol), Tekanan Darah (Blood Pressure), dan Detak Jantung (Heart Rate). Selain itu, dataset ini juga merekam informasi mengenai Diabetes (indikator keberadaan penyakit diabetes), Riwayat Keluarga (Family History) terkait penyakit jantung, dan kebiasaan Merokok (Smoking). Faktor gaya hidup dan kondisi fisik juga tercatat, seperti tingkat Obesitas (Obesity), Konsumsi Alkohol (Alcohol Consumption), durasi Jam Olahraga Per Minggu (Exercise Hours Per Week), jenis Diet yang dijalani, serta Riwayat Masalah Jantung Sebelumnya (Previous Heart Problems). Data lain yang relevan meliputi Penggunaan Obat (Medication Use), Tingkat Stres (Stress Level), Jam Duduk Per Hari (Sedentary Hours Per Day), dan Pendapatan (Income). Informasi tambahan terkait pengukuran tubuh dan biokimia adalah IMT (BMI) dan Trigliserida (Triglycerides). Aktivitas fisik juga dipantau melalui Hari Aktivitas Fisik Per Minggu (Physical Activity Days Per Week), serta pola tidur dengan Jam Tidur Per Hari (Sleep Hours Per Day). Terakhir, informasi geografis pasien dicatat melalui atribut Negara (Country), Benua (Continent), dan Belahan Bumi (Hemisphere). Seluruh informasi ini mencerminkan beragam faktor yang berkontribusi pada profil risiko kesehatan pasien. Dataset ini sangat berguna dalam analisis pengelompokan pasien, khususnya untuk mengidentifikasi pola-pola karakteristik pasien yang berkaitan dengan risiko serangan jantung, dan dapat pula menjadi dasar bagi pengembangan model prediksi di masa mendatang. Atribut Risiko Serangan Jantung (Heart Attack Risk) merupakan kolom target yang akan digunakan untuk validasi hasil pengelompokan di masa depan, namun tidak diikutsertakan dalam proses *K-Means* Clustering karena sifatnya sebagai algoritma *unsupervised learning*.

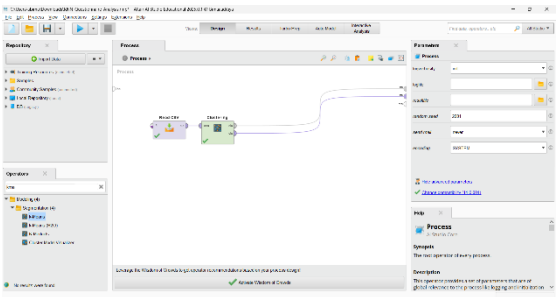


Gambar 2 Dataset

Implementasi *K-Means*

Setelah data berhasil dimasukkan dan ditinjau, tahap berikutnya adalah melakukan pemodelan terhadap dataset pasien risiko serangan jantung. Algoritma yang diterapkan yaitu *K-Means*, dengan bantuan perangkat lunak *RapidMiner*. Tahapan diawali dengan mengimpor data ke dalam *RapidMiner*, lalu data dipra-proses dan dianalisis menggunakan algoritma *K-Means* dengan parameter jumlah klaster (K) yang telah ditentukan (misalnya, sebanyak 3) serta jenis pengukuran yang sesuai untuk data numerik dan kategorikal

.



Gambar 3 Implementasi *K-Means*

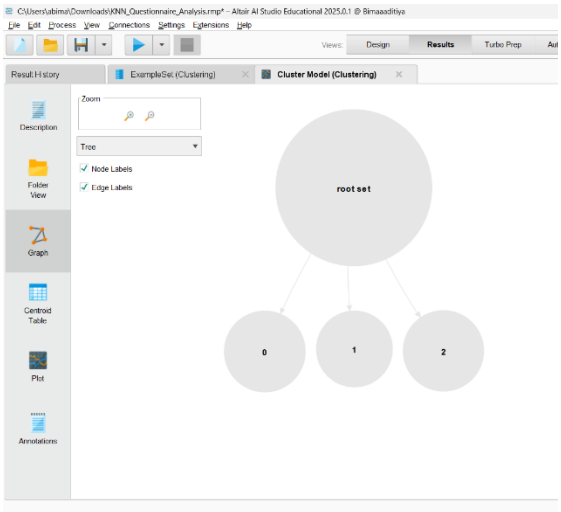
Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

Proses dimulai dengan mengimpor data ke dalam *RapidMiner*, lalu data tersebut dianalisis menggunakan algoritma *K-Means* dengan parameter k = 3 dan jenis pengukuran yang digunakan adalah tipe campuran. Pendekatan ini dipilih karena dalam analisis Hate Speech Twitter, tujuan utamanya adalah untuk mengelompokkan data ke dalam tiga kategori, yaitu C0 untuk komentar netral, C1 untuk komentar negatif, dan C2 untuk komentar positif. Setelah proses pemodelan selesai, *RapidMiner* menghasilkan tiga kluster, yaitu kluster 0 dengan 2.997 data, kluster 1 berjumlah 2.787 data, dan kluster 2 terdiri dari 2.979 data.

Tabel 1 Clustering

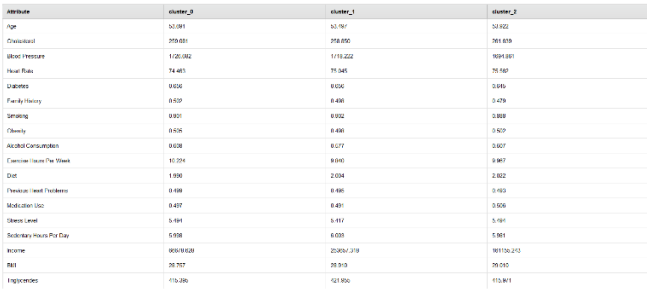
Cluster 0	2.997
Cluster 1	2.787
Cluster 2	2.979
Total	8.763

Adapun hasil yang berupa pohon dalam pilihan yang diberikan oleh *RapidMiner*s memberikan gambaran mengenai seberapa besar perbedaan antara cluster dengan seluruh data. Dengan opsi ini, kita dapat melihat bagian dari data utama atau root set, yang terbagi menjadi beberapa cabang yaitu sebuah cluster.



Gambar 4 Diagram Tree

Dari hasil pengelompokan yang telah dilakukan menggunakan algoritma *K-Means*, data kemudian diklasifikasikan ke dalam tiga kategori utama. Hal ini menunjukkan bahwa kluster 0, yang merepresentasikan komentar netral, berisi sebanyak 2.997 data. Selanjutnya, kluster 1 yang menggambarkan komentar negatif terdiri dari 2.787 data, sedangkan kluster 2 yang mencerminkan komentar positif mencakup 2.979 data. Pembagian ini membantu dalam memahami persebaran karakteristik pasien yang berkaitan dengan risiko serangan jantung.



Gambar 5 Table Centroid

D. PENUTUP

Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa algoritma *K-Means* Clustering mampu mengelompokkan data pasien risiko serangan jantung ke dalam tiga kategori utama. Dengan menggunakan dataset sebanyak 8.763 entri dari platform *Kaggle*, proses clustering melalui *RapidMiner* menghasilkan tiga klaster, di mana:

- Klaster 0 profil pasien risiko rendah memiliki jumlah data sebanyak dari 2.997 data pasien.
- Klaster 1 profil pasien risiko sedang memiliki jumlah data sebanyak 2.787 data pasien.
- Klaster 2 profil pasien risiko tinggi memiliki jumlah data sebanyak 2.979 data pasien.

Hasil ini menunjukkan bahwa algoritma *K-Means* efektif dalam mengidentifikasi kelompok-kelompok pasien dengan karakteristik serupa yang berkorelasi dengan risiko serangan jantung. Penggunaan algoritma *K-Means* terbukti efektif dalam membantu mengidentifikasi dan mengelompokkan profil risiko, sehingga dapat digunakan sebagai dasar analisis dalam upaya deteksi dini dan intervensi pada pasien berisiko serangan jantung.

SARAN

1. Pengembangan Metode Lanjutan: Disarankan untuk menggunakan kombinasi algoritma *clustering* lain atau metode *machine learning* seperti Hierarchical Clustering, DBSCAN, atau algoritma klasifikasi (misalnya, SVM, Random Forest) setelah *clustering* untuk validasi atau prediksi, guna memperoleh pemahaman yang lebih mendalam tentang kelompok risiko.

2. Peningkatan Pra-pemrosesan Data: Disarankan untuk melakukan pra-pemrosesan data yang lebih mendalam, seperti penanganan *outlier*, *feature scaling* (normalisasi/standardisasi) yang lebih kompleks, serta rekayasa fitur baru jika memungkinkan, guna meningkatkan kualitas dan representasi data untuk *clustering*.

3. Penggunaan Dataset Lebih Dinamis / Lengkap : Penelitian ini menggunakan data statis dari *Kaggle*. Untuk mendapatkan hasil yang lebih relevan dan kontekstual, sebaiknya data dikumpulkan dari sumber medis yang lebih dinamis atau menyertakan

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

lebih banyak variabel klinis dan demografi yang bervariasi.

4. Penerapan Nyata di Lingkungan Klinis: Hasil analisis ini memiliki potensi untuk digunakan oleh praktisi kesehatan, rumah sakit, atau lembaga penelitian medis untuk mendeteksi kelompok pasien berisiko tinggi secara lebih cepat dan sistematis, serta merancang program intervensi preventif yang lebih personal.

5. Validasi Eksternal dan Ekspertise Domain: Disarankan untuk melakukan validasi eksternal terhadap klaster yang terbentuk dengan data independen atau melibatkan ahli medis (dokter, kardiolog) untuk memverifikasi interpretasi dan relevansi klinis dari setiap klaster.

E. DAFTAR PUSTAKA

Anwar, C. (2022). Application of Academic Information System With Extreme Programming Method (Case Study: Jakarta International Polytechnic).

Anwar, C. (2024). Rekomendasi Teknis Untuk Pengolahan Data Berbasis *Web*. Jurnal Informatika Utama, 2(1), 50-54.

Anwar, C., & Riyanto, J. (2019). Perancangan Sistem Informasi Human Resources Development Pada PT. Semacom Integrated. International Journal of Education, Science, Technology, and Engineering (IJESTE), 2(1), 19-38.

Anwar, C., Jagat, L. S., Yanti, I., Anjarsari, E., & Sholihah, N. A. (2023). Pengembangan Media Pembelajaran Berbasis Teknologi Untuk Meningkatkan Kemampuan Anak. Caruban: Jurnal Ilmiah Ilmu Pendidikan Dasar, 6(2), 154-163.

Anwar, C., Kom, S., Kom, M., Santiari, C. N. P. L., & Sitorus, Z. (2023). Buku Referensi Sistem Informasi Berbasis Kearifan Lokal.

Anwar, C., Nurhasanah, M., Aflaha, D. S. I., & Handayani, S. (2023). DEVELOPMENT OF INFORMATION TECHNOLOGY-BASED LEARNING MEDIA FOR EDUCATORS IN ELEMENTARY SCHOOLS. Jurnal Konseling Pendidikan Islam, 4(2), 345-353.

Anwar, Chairul, et al. "The Application of Mobile Security Framework (MOBSF) and Mobile Application Security Testing Guide to Ensure the Security in Mobile Commerce Applications." Jurnal Sistim Informasi dan Teknologi (2023): 97-102.

Rahayu, S., Subekhi, A., Astuti, D., Widaningsih, I., Sartika, I., Nurhayani, N., ... & Rafidah, R. (2020). Upaya mewaspadai serangan jantung melalui pendidikan kesehatan. *JMM (Jurnal Masyarakat Mandiri)*, *4*(2), 163-171.

Handayani, F. (2022). Aplikasi Data MiningMenggunakan Algoritma *K-Means* Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar. *Jurnal Teknologi dan Informasi (JATI)*, 2088-2270.

Hani, J. E. (2022). Implementasi Data Mining Untuk Menentukan Persediaan Stok Barang Di Mini Market Menggunakan Metode *K-Means* Clustering. *Jurnal Informatika dan rekayasa komputer (JAKAKOM)*, 2808-5469.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in *K-Means* Clustering. International Journal. *International Journal*, *1*(6), 90–95.

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)