

SISTEM REKOMENDASI KATEGORI PASAR *ABALONE* (*PREMIUM, REGULER, REJECT*) UNTUK PASAR KONSUMEN MENGGUNAKAN DATA MINING

Delvita Aulia Artika¹, Bunga Dwi Febrianti^{2*}, Sirus Daniel Nababan³, Arnita⁴,
Fanny Ramadhani⁵

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email: delvitaartika@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia²

Email*: bungadwifeb@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email: sirusnababan20@gmail.com

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email: arnita@unimed.ac.id

Universitas Negeri Medan, Medan, Sumatera Utara, Indonesia¹

Email: fannyr@unimed.ac.id

ABSTRAK

Abalone atau kerang mata tujuh, merupakan komoditas laut bernilai tinggi yang banyak diminati di sektor kuliner premium, namun proses klasifikasi kualitasnya masih dilakukan secara manual dan kurang efisien. Penelitian ini bertujuan mengembangkan sistem rekomendasi kategori pasar abalone (*Premium, Regular, Reject*) berbasis data mining untuk membantu proses penilaian kualitas secara otomatis, objektif, dan akurat. Metode yang digunakan mencakup segmentasi data dengan algoritma *K-Means* berdasarkan karakteristik fisik abalone, lalu klasifikasi menggunakan *Decision Tree* untuk menentukan kategori pasar. Dataset berasal dari *UCI Machine Learning Repository* dan telah melalui tahapan *pre-processing* seperti normalisasi, *label encoding*, dan penanganan *outlier* sebelum pelatihan model. Hasil penelitian menunjukkan bahwa sistem klasifikasi yang dibangun mampu memprediksi kategori pasar dengan akurasi 99%, serta *precision*, *recall*, dan *F1-score* yang tinggi di semua kelas. Penelitian ini menggabungkan segmentasi *K-Means* dan klasifikasi *Decision Tree* untuk mengembangkan sistem penilaian kategori pasar abalone secara otomatis dan akurat untuk menggantikan proses manual yang subjektif.

Kata Kunci: *Abalone, Decision Tree, Data Mining, K-Means, Sistem Rekomendasi*

ABSTRACT

Abalone (Haliotis spp.), known as "kerang mata tujuh" in Indonesia, is a highly valuable marine commodity widely sought after in the premium culinary sector, yet its quality classification is still commonly performed manually and inefficiently. This study aims to develop a market category recommendation system for abalone (Premium, Regular, Reject) based on data mining to assist in quality assessment in an automated, objective, and accurate manner. The methods used include data segmentation with the K-Means algorithm based on the physical characteristics of abalone, then classification using Decision Tree to determine market categories. The dataset comes from the UCI Machine Learning Repository and has gone through pre-processing stages such as normalization, label encoding, and outlier handling before model training. The results show that the

built classification system can predict market categories with 99% accuracy, along with high precision, recall, and F1-scores across all classes, indicating the system's strong potential for implementation in the fisheries industry to support fast, data-driven decision-making. This study combines K-Means segmentation and Decision Tree classification to develop an automatic and accurate abalone market category assessment system to replace subjective manual processes

Keywords: Abalone, K-Means, Decision Tree, Classification, Data Mining, Recommendation System

1. PENDAHULUAN

Abalone, yang dikenal di Indonesia sebagai kerang mata tujuh, merupakan komoditas laut bernilai ekonomis tinggi dari kelas *Gastropoda. Moluska* bercangkang tunggal ini memiliki cita rasa khas, kandungan gizi yang tinggi, dan dipercaya mampu meningkatkan vitalitas, sehingga sering dijadikan hidangan mewah [1]. Di sejumlah negara Asia seperti Tiongkok dan Korea Selatan, industri budidaya abalone mengalami pertumbuhan yang sangat pesat dan telah mendominasi pasar global, dengan Tiongkok mencatat produksi lebih dari 139.000-ton pada tahun 2017. Kualitas abalone menjadi faktor utama yang menentukan harga jualnya, sehingga proses klasifikasi yang tepat sangat penting dalam distribusi dan pemasaran. Namun, karena pertumbuhan abalone berlangsung sangat lambat sementara permintaannya terus meningkat, hal ini berpotensi mendorong praktik penangkapan abalone liar secara intensif [2].

Perkembangan teknologi, khususnya dibidang data mining, memungkinkan untuk membangun sistem klasifikasi otomatis berbasis data yang lebih akurat dan efisien. Tujuan data mining mencari pola hubungan berdasarkan dari sekumpulan data yang kompleks, proses data mining terdiri dari beberapa tahapan yakni mengidentifikasi masalah yang sering terjadi, mendeskripsikan masalah tersebut, kemudian dilakukan tahap analisis dengan teknik data mining menggunakan metode algoritma apriori untuk mendapatkan hasil [3]. Untuk itu, penelitian ini mengembangkan sistem berbasis data mining yang mampu membantu proses klasifikasi secara otomatis, objektif, dan akurat.

Dalam penelitian ini, dikembangkan sistem berbasis data mining untuk membantu proses klasifikasi kualitas abalone secara otomatis, objektif, dan akurat. Algoritma *K-Means* dipilih karena kemampuannya dalam melakukan *clustering* secara efisien [4]. *K-Means* adalah algoritma partisional yang bekerja secara iteratif dengan mengelompokkan data berdasarkan jarak terdekat ke pusat kluster (*centroid*), sehingga sangat cocok untuk menemukan segmentasi alami pada data abalone tanpa memerlukan label awal. Proses ini sangat membantu dalam mengidentifikasi kelompok pasar abalone berdasarkan karakteristik yang serupa secara otomatis [5].

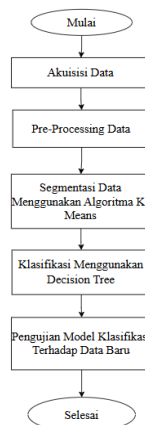
Setelah proses pengelompokan dilakukan, algoritma *Decision Tree* digunakan sebagai metode klasifikasi untuk menentukan kategori pasar abalone, yaitu *Premium*, *Reguler*, dan *Reject* [6]. *Decision Tree* mampu membangun model prediktif yang mudah dipahami, dengan aturan keputusan yang jelas dan logis. Algoritma ini cepat dalam pelatihan dan prediksi, serta dapat menangani berbagai jenis variabel, sehingga cocok untuk integrasi praktis dalam industri. [7]. Kombinasi kedua algoritma ini diharapkan dapat meningkatkan akurasi sistem rekomendasi serta mempercepat proses pengambilan keputusan dalam klasifikasi kualitas abalone.

Seperti penelitian yang dilakukan oleh [8], *Algoritma K-Means* dan *Decision Tree* telah berhasil digunakan dalam analisis pola konsumsi BBM, menunjukkan relevansi penerapannya dalam segmentasi pasar dan pengambilan keputusan berbasis data.

Berdasarkan pendekatan tersebut, penelitian ini bertujuan mengembangkan sistem rekomendasi kategori pasar abalone (*Premium, Regular, Reject*) menggunakan data mining, dengan proses otomatis pengelompokan data oleh *K-Means* dan klasifikasi oleh *Decision Tree* [9]. Dengan teknik ini, sistem rekomendasi diharapkan mampu mendukung pengambilan keputusan yang lebih objektif dan efisien dalam menilai kualitas abalone. Selain itu metode seperti *Naïve Bayes* juga dapat digunakan pada penelitian ini karena kecepatan dan kesederhanaannya. Namun, *Naïve Bayes* mengasumsikan independensi antar fitur, yang kurang cocok untuk data abalone dengan fitur saling mempengaruhi [8]. Sehingga *Decision Tree* lebih sesuai dalam penelitian ini. Sistem ini siap diterapkan di industri perikanan untuk mendukung otomatisasi dan pengelolaan pasok yang lebih baik.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksplorasi data untuk mengembangkan sistem klasifikasi kategori pasar Abalone (*Premium, Regular, Reject*) berdasarkan fitur-fitur fisik abalone. Penelitian ini menggunakan bahasa pemrograman *python* pada platform *Google Colab*. Karena data mentah yang digunakan tidak memiliki label kategori pasar, dilakukan segmentasi menggunakan algoritma *K-Means Clustering*, dilanjutkan dengan pelatihan model klasifikasi menggunakan *Decision Tree Classifier* pengujian model pada data baru. Metode ini terdiri atas beberapa tahapan, yaitu:



Gambar 1. Flowchart penelitian

Akuisisi Dataset

Ini merupakan tahapan awal penelitian, dimulai dengan akuisisi (pengambilan) dan eksplorasi awal dataset abalone. Dimana kami mengunduh dataset “Abalone” dari sumber terpercaya yaitu web: <https://archive.ics.uci.edu/> (*UCI Machine Learning Repository*). Dataset ini berisi data tentang karakteristik fisik abalone, sejenis moluska laut, yang dapat digunakan untuk memperkirakan usia dan kualitasnya. Dimana dataset Abalone ini merupakan jenis numerik, multivariat, dan kategorikal. Dataset ini terdiri dari 4.177 baris data dengan 8 fitur input dan 1 target.

Tabel 1. Fitur dalam dataset

No	Nama Fitur	Tipe Data	Keterangan
1	<i>Sex</i>	Object	Jenis abalone (M, F, I)
2	<i>Length</i>	Float	Panjang tubuh(tidak termasuk cangkang)
3	<i>Diameter</i>	Float	Diameter tubuh
4	<i>Height</i>	Float	Tinggi abalone

5	<i>Whole_Weight</i>	Float	Berat total
6	<i>Shucked_Weight</i>	Float	Berat daging
7	<i>Viscera_Weight</i>	Float	Berat organ dalam
8	<i>Shell_Weight</i>	Float	Berat cangkang setelah dikeringkan
9	<i>Rings</i>	Int	Jumlah cincin(indikassi umur abalone)

Setelah data diunduh dalam format .csv maka data dapat di impor ke dalam ke lingkungan kerja Google Colab menggunakan library pandas yang selanjutnya dapat melakukan eksplorasi. Eksplorasi data memberikan pemahaman yang lebih dalam tentang karakteristik abalone dan menunjukkan bahwa dataset ini layak digunakan untuk: *Clustering (unsupervised)* untuk segmentasi pasar berdasarkan kesamaan fisik. Dan Klasifikasi (*supervised*) untuk pemetaan hasil cluster ke label pasar (*Reject, Regular, Premium*).

Pra-Processing Data

Tahap ini penting untuk mempersiapkan data mentah sebelum segmentasi dan klasifikasi. Langkah awal adalah memeriksa missing value di semua fitur untuk menghindari gangguan saat pelatihan. Penanganannya bisa dengan menghapus baris sedikit yang bermasalah atau imputasi menggunakan mean, median, modus, atau metode berbasis model [10]. Selanjutnya tahapan untuk menangani *outlier*. *Outlier* adalah nilai-nilai yang signifikan berbeda dari sebagian data dalam dataset dan mempengaruhi hasil analisis statistik serta merusak performa model, salah satunya pada algoritma berbasis jarak *K-means*. Metode yang digunakan adalah *Interquartile Range (IQR)* yang menghitung selisih persebaran nilai tengah dari dataset. Data yang berada di luar bawah dan atas dianggap sebagai outlier [11]. Persamaan matematis untuk mengecek outlier dengan metode IQR:

$$IQR = Q3 - Q1 \quad (1)$$

$$Lower\ Bound = Q1 - 1.5 \times IQR \quad (2)$$

$$Upper\ Bound = Q3 + 1.5 \times IQR \quad (3)$$

Dimana $Q1$ = Kuartil pertama (25%) dari data terendah), $Q3$ = Kuartil ketiga (75% data terendah). Sebuah nilai dianggap outlier jika $x < Lower\ Bound$ atau $x > Upper\ Bound$. Namun untuk menjaga integritas dataset tidak perlu dihapus dan bisa digantikan dengan nilai median dari fitur terkait. Dengan persamaan matematis:

$$x'_i = \begin{cases} \frac{median(X)}{x_i}, & \text{jika } x_i < Q1 - 1.5 \times IQR \text{ atau } x_i > Q3 + 1.5 \times IQR \\ 1, & \text{jika } Q1 - 1.5 \times IQR \leq x_i \leq Q3 + 1.5 \times IQR \end{cases} \quad (4)$$

Selanjutnya adalah tahap untuk mengubah kolom *Sex* yang awalnya kategorial menjadi angka dengan metode label encoding. Dimana metode label encoding adalah metode untuk mengubah data kategorikal yang direpresentasikan sebagai label teks menjadi format numerik. Ini dilakukan karena algoritma K-Means hanya bisa mengolah data dalam bentuk numerik [12]. Dalam dataset kolom *Sex* berisi data kategorikal berupa jenis kelamin abalone *Male*(jantan), *Female*(betina), *Infant*(anak-anak). Dengan metode label encoding setiap kategori pada kolom *Sex* di beri label angka unik yaitu M=2, F=0, dan I= 1. Proses ini dilakukan otomatis berdasarkan urutan label saat di scan oleh encoder.

Langkah selanjutnya adalah melakukan standarisasi pada fitur numerik seperti *Length*, *Diameter*, *Height*, dan *Whole_weight* karena memiliki satuan dan rentang

berbeda. Tanpa standarisasi, model cenderung memprioritaskan fitur berskala besar. Proses ini menyelaraskan data agar setiap fitur memberi kontribusi setara dalam model. [3]. Pada penelitian ini metode standarisasi yang digunakan adalah z-score standardisation. Z- score bekerja dengan cara menghitung rata-rata (mean) dan standar deviasi dari masing-masing fitur, kemudian mengubah nilai aslinya menjadi nilai standar. Nilai z-score merupakan suatu seberapa besar jarak suatu nilai terhadap rata-ratanya dalam satuan standar deviasinya. Dengan persamaan matematis:

$$Z = \frac{x-\mu}{\sigma} \quad (5)$$

Dimana z adalah nilai z-score, x adalah nilai yang akan dinormalisasikan, μ adalah nilai rata-rata seluruh kolom dan baris, dan σ adalah nilai standar deviasi. Apabila z-score bernilai negative (-) maka ada di sebelah kiri rata-rata nya dalam kurva normal. Sementara bila bernilai positive(+), maka ada sebelah kanan rata-ratanya [5]. Setiap fitur numerik dalam dataset diubah nilai-nilai sehingga memiliki nilai mean nya 0, dan simpangan bakunya 1.

Selanjutnya adalah menambah kolom *Age* ke dalam dataset. Penambahan kolom ini dilakukan berdasarkan informasi dari kolom *Rings*. Dalam dataset abalone jumlah cincin (*Rings*) pada cangkang abalone digunakan untuk memperkirakan usia abalone tersebut. Berdasarkan studi biologis pada penelitian [13], umur abalone diperkirakan dengan rumus :

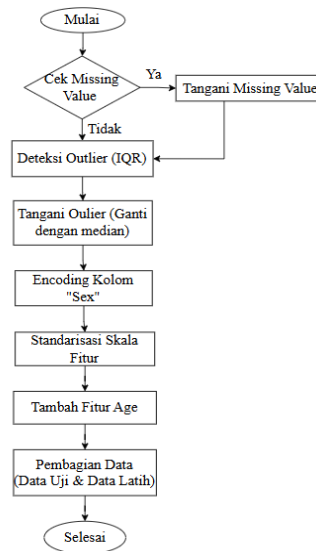
$$Age = Rings + 1.5 \quad (6)$$

Angka 1.5 tahun merupakan estimasi waktu pertumbuhan awal abalone sebelum cincin pertama terbentuk. Penambahan kolom *Age* bertujuan menyediakan informasi yang lebih bermakna daripada jumlah cincin, dengan estimasi pertumbuhan awal abalone. Kolom ini juga digunakan sebagai label dalam pemodelan serta mempermudah interpretasi dan visualisasi data.

Kemudian dataset dibagi menjadi dua subset utama. Pembagian ini bertujuan untuk mengevaluasi kemampuan model dimana dataset latih (data train) digunakan untuk melatih model agar dapat mengenali pola yang ada, sedangkan dataset uji (data test) digunakan untuk menguji kinerja model terhadap data yang belum pernah dilihat sebelumnya. Proporsi pembagian 80% untuk data latih dan 20% untuk data uji. Proporsi ini merupakan standar dalam machine learning untuk menguji model pada data yang belum pernah dilihat sebelumnya. Proporsi ini menghasilkan keseimbangan yang baik tanpa terlalu mengorbankan jumlah data untuk pelatihan maupun pengujian. Dengan memberikan 80% data untuk pelatihan model memiliki cukup banyak contoh untuk mempelajari pola dan hubungan antara fitur dan target. Dan dengan 20% data disisihkan untuk pengujian untuk mengukur performa model [14].

Selain itu, parameter `random_state` diset sebesar 42 untuk menjamin konsistensi pembagian data yang sama saat kode dijalankan berulang kali. Penggunaan angka 42 tidak memiliki makna khusus secara teknik namun angka 42 populer sebagai konvensi dalam pemrograman, terinspirasi dari referensi humor dalam buku *The Hitchhiker's Guide to the Galaxy*.

Data fitur (X) dan label target (y) dipisahkan terlebih dahulu, kemudian masing-masing dibagi menjadi: X_{train} dan y_{train} : data fitur dan label yang digunakan untuk melatih model. Data latih yang digunakan yaitu 3341 data. Dan X_{test} dan y_{test} : data fitur dan label yang digunakan untuk menguji performa model setelah pelatihan. Data uji yang digunakan yaitu 836 data. flow chart tahapan pre-processing data yang dilakukan adalah sebagai berikut:



Gambar 2. Flowchart pre-processing

Segmentasi Data Menggunakan K-Means

Segmentasi data dilakukan menggunakan algoritma *K-Means Clustering*, metode partisi dalam data mining yang membagi data ke dalam (*K klaster*) berdasarkan kemiripan karakteristik. *Clustering* berfungsi mengelompokkan data agar kesamaan dalam klaster tinggi dan antar klaster rendah. Metode ini termasuk *unsupervised learning* yang secara otomatis menemukan pola untuk membentuk klaster dengan kesamaan tertentu [15].

Cluster ini bertujuan mengelompokkan data abalone berdasarkan kemiripan fitur untuk menemukan pola tersembunyi tanpa label. Hasil klastering dapat membantu pelatihan model klasifikasi dengan mengidentifikasi batas antar kelas. Sebelum proses clustering, jumlah cluster optimal ditentukan menggunakan metode *Elbow*. Metode ini menentukan jumlah cluster optimal dengan mengamati titik siku (elbow) pada grafik penurunan inertia (WCSS) seiring bertambahnya jumlah cluster (*K*). *Within-Cluster Sum of Squares* (WCSS) menghitung total jarak kuadrat antara data dan centroid-nya; nilai yang lebih kecil menunjukkan cluster yang lebih homogen. Titik tekukan pada grafik menunjukkan jumlah cluster paling efektif untuk membagi data.

$$WCSS = \sum_{k=1}^K \sum_{i \in C_k} \|X_i - \mu_k\|^2 \quad (7)$$

Di mana *K* adalah jumlah kluster, *C_k* adalah cluster ke-*k*, *X_i* adalah titik data ke-*i* dalam cluster *C_k* dan μ_k adalah centroid (pusat) dari cluster *C_k* serta $\|X_i - \mu_k\|^2$ adalah jarak kuadratt (biasanya Euclidean) antara data *x_i* dan centroid cluster-nya [16].

Setelah nilai *K* ditetapkan, algoritma *K-Means* dengan *n_clusters=3* dan *random_state=42* dijalankan untuk mengelompokkan data latih (*X_train*) berdasarkan fitur panjang, diameter, umur, dan berat total. Hasil clustering ditambahkan sebagai kolom "*Cluster*" dengan label sesuai pusat klaster: cluster 0 (nilai fitur rendah), cluster 1 (menengah), dan cluster 2 (tinggi). Label ini digunakan sebagai dasar pemberian kategori pasar secara tidak langsung dan menjadi input untuk klasifikasi lebih lanjut menggunakan *Decision Tree*.

Klasifikasi Menggunakan Decision Tree

Setelah segmentasi dengan *K-Means*, data diklasifikasikan menggunakan *Decision Tree*, metode machine learning untuk klasifikasi dan regresi yang memodelkan

keputusan dalam bentuk pohon. Setiap node mewakili fitur, cabang keputusan, dan daun hasil akhir atau kategori. Metode ini mudah dipahami, visual, serta dapat menangani data numerik dan kategorikal [6].

Kategori pasar abalone (*Reject*, *Reguler*, *Premium*) ditentukan dari hasil segmentasi *K-Means* dengan mengacu pada rata-rata umur tiap cluster. Cluster 0 (umur terendah) diberi label *Reject*, cluster 2 (umur menengah) *Reguler*, dan cluster 1 (umur tertinggi) *Premium*. Model dilatih menggunakan fitur *Length*, *Diameter*, *Age*, dan *Whole_weight* pada data berlabel, lalu dievaluasi dengan membandingkan prediksi terhadap label pasar dari hasil clustering data uji.

Pengujian Model Klasifikasi Terhadap Data Baru

Setelah model *Decision Tree* dibangun berdasarkan segmentasi dan pelabelan *K-Means*, dilakukan pengujian pada data baru untuk menilai kemampuannya dalam mengklasifikasikan kategori pasar abalone (*Premium*, *Reguler*, *Reject*). Data uji berupa nilai fitur *length*, *diameter*, *age*, dan *whole weight* diproses oleh model untuk menghasilkan label kategori pasar. Tujuan pengujian ini adalah mengevaluasi kinerja model terhadap data yang belum pernah dilihat sebelumnya, guna memastikan aplikabilitasnya di dunia nyata. Evaluasi dilakukan menggunakan *confusion matrix*, dengan fokus pada akurasi, yakni persentase prediksi yang benar oleh model, yang dihitung dengan rumus berikut [17] :

$$Akurasi = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100\% \quad (8)$$

Presisi yaitu tingkat ketepatan antara informasi yang diminta oleh pengguna yang diminta dengan jawaban yang diberikan oleh sistem, berikut rumus untuk menghitung presisi:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

Recall yaitu persentase sebuah program memprediksi sebuah data ke bukan kelas aktualnya, recall biasa disebut juga dengan sensitifitas, berikut rumus untuk menghitung Recall:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F1-Score merupakan salah satu perhitungan evaluasi dalam informasi yang mengkombinasikan recall dan presisi. Nilai F1- Score menunjukan perbandingan rata-rata recall dan presisi yang dibobotkan, berikut rumus untuk menghitung F1-score:

$$F1 = 2x \frac{precision \times recall}{precision + recall} \quad (11)$$

3. HASIL DAN PEMBAHASAN

Pada tahap pre-processing didapat hasil data mentah yang optimal sehingga layak untuk dilakukan klasifikasi dengan algoritma machine learning. Dataset mentah abalone yang terdiri 8 fitur input dan 1 fitur target (*Rings*) yang kemudian fitur *Rings* dikonversi menjadi *Age*. Dan fitur ini menjadi fitur target baru sehingga isi dari dataset abalone yang baru adalah:

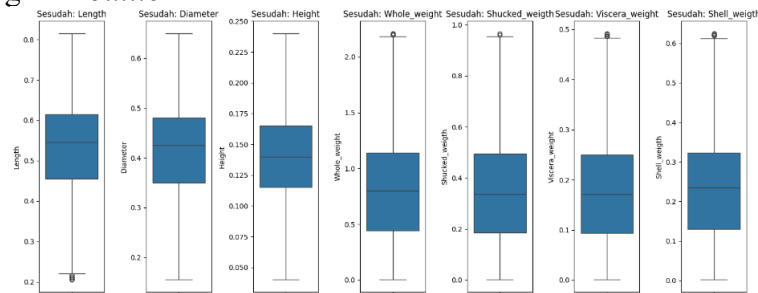
Tabel 2. Dataset setelah ditambah kolom *Age*

No	Nama Fitur	Tipe Data	Keterangan
1	<i>Sex</i>	Object	Jenis abalone (M, F, I)
2	<i>Length</i>	Float	Panjang tubuh(tidak termasuk cangkang)
3	<i>Diameter</i>	Float	Diameter tubuh
4	<i>Height</i>	Float	Tinggi abalone
5	<i>Whole Weight</i>	Float	Berat total

No	Nama Fitur	Tipe Data	Keterangan
6	<i>Shucked_Weight</i>	Float	Berat daging
7	<i>Viscera_Weight</i>	Float	Berat organ dalam
8	<i>Shell_Weight</i>	Float	Berat cangkang setelah dikeringkan
9	<i>Age</i>	Int	Umur abalone

Setelah itu untuk *missing value* hasil pengecekan menunjukkan bahwa tidak ada nilai kosong atau hilang pada semua fitur sehingga tidak diperlukan proses imputasi.

Beberapa *outlier* terdeteksi dengan metode IQR dan diganti dengan median fitur terkait agar data tetap utuh. Setelah penanganan, distribusi data lebih normal dan representatif, penting untuk algoritma sensitif skala seperti K-Means. Berikut boxplot sesudah penanganan *outlier* :



Gambar 3. Bosplot setelah penanganan *outlier*

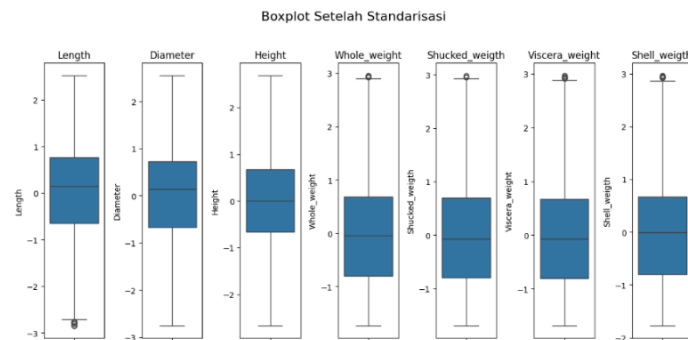
Boxplot di atas menunjukkan distribusi data setelah *outlier* diatasi dengan mengganti nilainya menggunakan median. Hasilnya, jumlah *outlier* berkurang, distribusi data lebih rapi, dan nilai ekstrim di atas 2,5 sudah hilang.

Selanjutnya, kolom *Sex* diubah dari kategorikal (M, F, I) menjadi numerik (0, 1, 2) melalui label encoding, berikut contoh data *Sex* sebelum dan sesudah encoding:

Tabel 3. Kolom *Sex* setelah encoding

No	<i>Sex</i>	<i>Sex</i> setelah encode
1	M	2
2	M	2
3	F	0
4	M	2
5	I	1

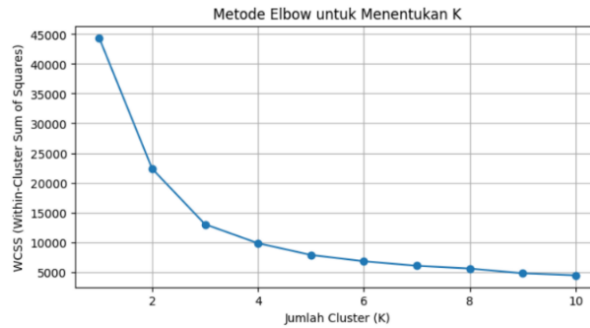
Pada tahapan standarisasi skala dengan metode z-score standarisation untuk menyeragamkan skala semua fitur. Dengan skala data diubah mejadi mean= 0 dan standar deviasi= 1. Supaya saat diolah dalam algoritma statistik atau machine learning, semuanya punya kontribusi yang seimbang. Berikut boxplot untuk melihat data sesudah di standarisasi:



Gambar 4. Boxplot setelah standarisasi

Pada boxplot ini median sudah berada disekitar 0, whisker dan box menunjukkan penyebaran data dalam skala standar deviasi. Sehingga semua fitur punya skala yang sama dan seimbang.

Segmentasi data abalone menggunakan algoritma K-Means untuk mengelompokkan data abalone berdasarkan kemiripan fitur utama yang telah dipilih yaitu: panjang (*Length*), diameter (*Diameter*), umur (*Age*), dan berat keseluruhan (*Whole_weight*). Berdasarkan hasil metode Elbow, nilai optimal untuk jumlah cluster (*k*) adalah 3.

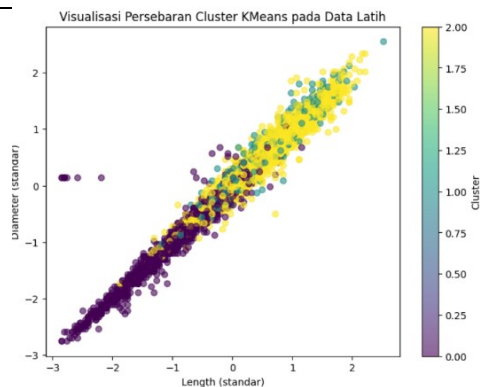


Gambar 5. Grafik menentukan K

Dapat dilihat pada grafik bahwa saat nilai K meningkat dari 1 sampai ke 10, nilai WCSS (*Within-Cluster Sum of Squares*) menurun tajam hingga K=3, lalu melambat, membentuk titik siku (elbow) pada K=3. Ini menunjukkan bahwa 3 adalah jumlah cluster optimal untuk algoritma K-Means. Dengan demikian, data abalone dibagi menjadi 3 cluster: Cluster 0: Rata-rata panjang, diameter, umur, dan berat total paling rendah. Cluster 1: Rata-rata fitur berada di tingkat menengah. Cluster 2: Rata-rata panjang, diameter, umur, dan berat total tertinggi.

Tabel 4. Contoh label data per cluster

ID	Length	Diameter	Age	Whole_weight	Cluster
4038	0.190054	0.352555	12.5	-0.308016	2
1272	-0.469586	-0.611449	9.5	-0.665406	0
3863	1.113550	0.941669	19.5	1.268956	1
3160	-2.228627	-2.271678	8.5	-1.486237	0
3894	-0.293682	-0.343670	13.5	-0.398159	2



Gambar 6. Visualisasi persebaran data uji cluster k-means

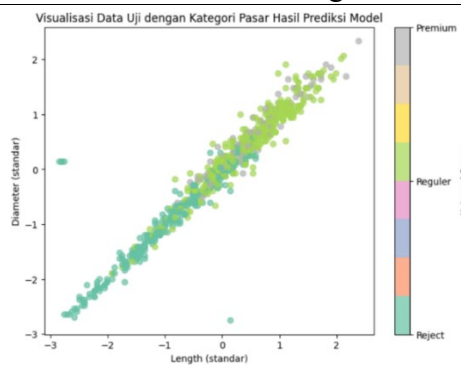
Visualisasi cluster K-Means menunjukkan korelasi positif antara panjang dan diameter abalone. Setiap warna mewakili cluster dengan ukuran fisik berbeda yang diasosiasikan

dengan kategori pasar: Warna ungu (cluster 0), warna hijau kebiruan (cluster 1), dan warna kuning (cluster 2).

Setelah pelabelan hasil segmentasi K-Means, model klasifikasi *Decision Tree* dibangun dan dilatih. Data uji tanpa label kemudian diklasifikasikan berdasarkan pola dari data latih, sehingga model dapat memprediksi kategori pasar abalone.

Tabel 5. Contoh label kategori pasar

ID	Cluster	Market_Label
4038	2	Reguler
1272	0	Reject
3384	0	Reject
3863	1	Premium
3894	2	Reguler



Gambar 7. Visualisasi persebaran data latih model klasifikasi

Tabel di atas menunjukkan hasil klasifikasi *Decision Tree* yang memberi label kategori pasar (Reject, Reguler, Premium) berdasarkan hasil clustering K-Means. Misalnya, ID 4038 berada di Cluster 2 dan dilabeli Reguler karena memiliki karakteristik serupa dengan kelompok pasar tersebut. ID 3863 masuk Cluster 1 dan diberi label Premium karena termasuk kelompok dengan rata-rata umur tertinggi. Sementara ID 1272, 3384, dan 3160 berada di Cluster 0 dan dilabeli Reject, yang mencerminkan kualitas terendah berdasarkan umur.

Visualisasi grafik menunjukkan hasil prediksi model terhadap data uji abalone. Setiap titik merepresentasikan satu abalone, dengan posisi ditentukan oleh panjang dan diameter (yang berkorelasi positif), dan warna menunjukkan prediksi kategori pasar oleh *Decision Tree*: biru (Reject), hijau (Reguler), dan abu-abu (Premium).

Pengujian juga dilakukan untuk menganalisis dan mengevaluasi model klasifikasi ini untuk melihat seberapa optimal. Pengujian dilakukan dengan menginput data baru dengan nilai fitur length, Diameter, Age, dan Whole_weight untuk di klasifikasikan dengan model *Decision Tree*.

Hasil Sistem Prediksi kategori Abalone:

Masukkan data fitur abalone untuk prediksi kategori pasar:

Length : 0.8; Diameter : 0.3; Age : 11; Whole_weight : 0.6

Prediksi kategori pasar abalone berdasarkan input Anda : Reguler.

Dari pengujian menunjukkan Klasifikasi dengan algoritma *Decision Tree* pada model ini mampu menghasilkan performa yang sangat baik. Dimana model dapat memprediksi kategori pasar dari data baru yang belum pernah dilihat sebelumnya.

Tabel 6. Hasil evaluasi model klasifikasi

Label	Precision	Recall	F1-Score	Support
Premium	1.00	1.00	1.00	101
Reguler	0.99	1.00	0.99	428
Reject	0.99	0.99	0.99	307
Accuracy			0.99	836
Macro avg	0.99	0.99	0.99	836
Weightedavg	0.99	0.99	0.99	836

Hasil evaluasi menunjukkan bahwa model klasifikasi memiliki performa sangat baik dalam mengelompokkan data ke tiga kelas (Premium, Reguler, Reject), dengan akurasi mencapai 99%. Nilai Precision, Recall, dan F1-score untuk semua kelas berada di kisaran 0.99 hingga 1.00, mencerminkan ketepatan dan cakupan prediksi yang tinggi. Distribusi data antar kelas juga proporsional, yang ditunjukkan oleh nilai *support*. Baik *macro average* maupun *weighted average* masing-masing mencapai 0.99, menandakan performa model yang konsisten dan merata, meskipun terdapat perbedaan jumlah data antar kelas, model tetap memberikan performa tinggi secara keseluruhan.

4. KESIMPULAN DAN SARAN

Penelitian ini berhasil menerapkan metode segmentasi dan klasifikasi untuk menentukan kategori pasar abalone berdasarkan karakteristik fisik. Algoritma *K-Means* membagi data menjadi tiga cluster optimal berdasarkan panjang, diameter, umur, dan berat total, dengan jumlah cluster ditentukan menggunakan metode Elbow. Setiap cluster dianalisis dan diberi label pasar: *Premium*, *Reguler*, dan *Reject*. Hasil segmentasi ini digunakan untuk pelabelan pada tahap klasifikasi.

Algoritma *Decision Tree* kemudian digunakan untuk klasifikasi, menghasilkan performa tinggi dengan akurasi 99% serta nilai precision, recall, dan f1-score yang konsisten. Model ini terbukti andal dalam mengklasifikasi data baru. Untuk pengembangan lebih lanjut, disarankan memperluas dataset dari berbagai wilayah, mencoba metode lain seperti DBSCAN dan *Random Forest*, serta mengimplementasikan sistem ini dalam bentuk aplikasi atau platform web untuk mendukung industri perikanan secara praktis dan otomatis.

5. DAFTAR PUSTAKA

- [1] I. M. S. Arta, I. G. N. P. Dirgayusa and N. L. P. R. Puspitha, "Perbandingan Laju Pertumbuhan Aballon (*Haliotis squamata*) Menggunakan Metode Co-culture dan Monoculture di Pantai Geger, Nusa Dua, Kabupaten Bandung, BALI," *Journal of Marine and Aquatic Sciences*, vol. 7, no. 2, pp. 232-242, 2021.
- [2] K. U. Nur, "Budidaya Abalon di Asia: Teknologi dan Manajemen Budidayanya," *Media Akuatika : Jurnal Ilmiah Jurusan Budidaya Perairan*, vol. V, no. 03, pp. 95-106, 2020.
- [3] A. Nugraha, O. Nurdiawan and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, pp. 849-855, 2022.
- [4] Normah, S. Nurajizah and A. Salbinda, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," *Jurnal Teknik Komputer AMIK BSI*, vol. 7, no. 2, pp. 158-163, 2021.

- [5] Y. M. Fahmi and S. , "Penerapan Algoritma K-Means Data Mining Pada Clustering Kelayakan Penerima UKT Dengan Normalisasi Data Model Z-Score," *Building of Informatics, Technology and Science (BITS)* , vol. VI, no. 03, p. 1977–1986 , 2024.
- [6] A. H. Nasrullah, "Implementasi Algoritma *Decision Tree* Untuk Klasifikasi Produk Laris," *Jurnal Ilmiah Ilmu Komputer* , vol. 7, no. 2, pp. 45-51, 2021.
- [7] D. A. Mukhsinin, M. Rafliansyah, S. A. Ibrahim, R. and D. Wulandari, "Implementasi Algoritma *Decision Tree* untuk Rekomendasi Film dan Klasifikasi Rating pada Platform Netflix," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. IV, no. 02, pp. 570-579, 2024.
- [8] U. Arfan and N. Paraga, "Perbandingan Algoritma K-Means, Naïve Bayes dan *Decision Tree* dalam Memprediksi Penjualan Bahan Bakar Minyak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science* , vol. 4, no. 4, pp. 1379-1389 , 2024.
- [9] A. Maulana, I. M. Ashari and A. Does, "Implementasi Sistem Rekomendasi Pada Sistem Informasi Seminar," *Just IT : Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, vol. 13, no. 3, p. 151 – 156 , 2023.
- [10] S. Y. Kurniawan, S. Sanjaya, Y. Vitriani and I. Afriyanti, "Klasifikasi Kelayakan Air Minum dengan Backpropagation Neural Network Berbasis Penanganan Missing Value dan Normalisasi," *Journal of Information System Research (JOSH)*, vol. VI, no. 01, p. 76–84, 2024.
- [11] R. J. Alfirdausy, I. Aliyah and A. Fanani, "Optimasi K-Nearest Neighbor Dengan Particle Swarm Optimization Untuk Klasifikasi Idiopathic Thrombocytopenic Purpura," *Komputika: Jurnal Sistem Komputer*, vol. XIII, no. 01, pp. 113 - 120 , 2024.
- [12] C. Herdian, A. Kamila and I. G. A. M. Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technologia : Jurnal Ilmiah*, vol. XV, no. 01, pp. 93-108, 2024.
- [13] S. Guney, I. Kilinc, A. A. Hameed and A. Jamil, "Abalone Age Prediction Using Machine Learning," *ResearchGate*, vol. X, no. 01, pp. 1-11, 2022.
- [14] M. D. S. Putra, I. G. J. E. Putra and K. Q. Fredlina, "Perancangan Sistem Informasi Monitoring Perkembangan Proyek Berbasis Mobile Pada Cv. Nayana Engineering Menggunakan Framework Flutter," *Jurnal Teknologi Informasi dan Komputer*, vol. X, no. 04, pp. 142-149, 2024.
- [15] R. I. P. Siagian, E. Pratama, F. A. Lubis, S. A. Priscilia and Ramadhani Fanny, "Segmentasi Pelanggan Dengan Algoritma K-Means Untuk Strategi Pemasaran Yang Efektif," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. IX, no. 04, pp. 5615 -5620, 2025.
- [16] Y. Hariyanto, A. Promadewi and M. Hanafi , "Analisis Kepuasan Masyarakat terhadap Pelayanan Publik menggunakan K-Means Clustering," *Journal of Information System Research (JOSH)*, vol. VI, no. 02, p. 1065–1074 , 2025.
- [17] T. Arifqi, N. Suarna and W. Prihartono, "Penggunaan Naive Bayes Dalam Menganalisis Sentimen Ulasan Aplikasi Mcdonald's Di Indonesia," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. VIII, no. 02, pp. 1949-1956, 2024.