Jurnal Komputer dan Informatika
Vol 15 No 2, Oktober 2020: hlm 271-276

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

IMPLEMENTASI *FUZZY C-MEANS CLUSTERING* DALAM MENENTUKAN USIA *ABALONE*

Erwin Surya Effendy
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
Email: *erwinsuryaeffendy@gmail.com*

ABSTRAK

Clustering adalah salah satu metode pengelompokan data yang dapat memberikan solusi pada masalah yang ada pada kehidupan sehari-hari. Maka dari itu, implementasi *Machine Learning* dalam teknik clustering akan menghasilkan akurasi prediksi yang baik dari kelas data pelatihan dengan jumlah instance yang besar, tetapi memberikan akurasi yang buruk di kelas dengan jumlah instance yang kecil.

Kata kunci: *Fuzzy C-Means*, Klasterisasi, Pembelajaran Mesin, *Abalone*, Kaggle

ABSTRACT

Clustering is a data grouping method that can provide solutions to problems encountered in everyday life. Therefore, the implementation of Machine Learning in clustering techniques will produce good prediction accuracy from training data classes with a large number of instances, but will produce poor accuracy in classes with a small number of instances.

Keywords*: Fuzzy C-Means, Clustering, Machine Learning, Abalone, Kaggle*

1. PENDAHULUAN

Penelitian ini akan melakukan analisis terhadap data yang besar yaitu data *Abalone* dengan 4117 instance sampel menggunakan algoritma *Fuzzy C-Means* yang telah dioptimasi dengan menggunakan seleksi atribut agar hasil cluster lebih optimal. Menghitung dan memastikan jumlah cincin pada cangkang abalone dengan menganalisis atribut lain yang dimiliki oleh suatu abalone, yakni jenis kelamin, panjang cangkang, berat abalone utuh, berat cangkang, dan lainnya. Dan pada tahap akhir penelitian, penulis dapat menentukan usia dari abalone tertentu yang dapat dihitung dari jumlah cincin yang telah dipastikan pada cangkang abalone. Kemudian, hasil analisis data ini digunakan untuk menentukan apakah dataset ini dapat menunjukkan indikasi depopulasi abalone yang terjadi di Tasmania sejak dataset ini awalnya dikumpulkan.

Metode

Clustering merupakan salah satu metode *Unsupervised Learning* yang bertujuan untuk melakukan pengelompokan data berdasarkan kemiripan atau jarak antar data. *Clustering* adalah metode pengelompokan data yang digunakan untuk mengenali kelompok-kelompok (*cluster*) yang dihasilkan dari pengelompokkan unsur-unsur yang lebih kecil berdasarkan adanya kemiripan satu sama lain (Tan, Steinbach, & Kumar, 2016). *Clustering* memiliki karakteristik dimana anggota dalam satu cluster memiliki kemiripan yang sama atau jarak yang sangat dekat, sementara anggota antar cluster memiliki kemiripan yang sangat berbeda atau jarak yang sangat jauh (Bai, Liang, & Dang, 2012).

Algoritma *clustering* yang digunakan pada penelitian ini adalah *Fuzzy C-Means* adalah suatu teknik pengelompokan (*clustering*) data yang bersifat iteratif ditentukan oleh nilai/derajakeanggotaan tertentu, setiap data dalam beberapa cluster dapat memiliki keanggotaan yang berbeda (Bezdek, Ehrlich, & Full, 1984). Keuntungan dari FCM adalah dapat memberikan

pemisahan *instance* yang lebih baik ketika suatu objek tidak dipisahkan dengan benar. Nilai *clustering* yang optimal dapat diukur dengan menggunakan *silhouette fuzzy*. Metode FCM ini terbukti efektif untuk diterapkan dalam proses pengklasifikasian karakteristik terhadap dataset. FCM menggunakan model pengelompokan fuzzy sehingga data dapat menjadi anggota dari semua kelas atau cluster terbentuk dengan derajat atau tingkat keanggotaan yang berbeda antara 0 hingga. Tingkat keberadaan data dalam suatu kelas atau cluster ditentukan oleh derajat keanggotaannya.

2. HASIL DAN PEMBAHASAN

2.1 Data

Dataset populasi abalone pada penelitian ini diunduh dari situs web bernama Kaggle. Dataset ini berasal dari (Nash, 1995) yang pertama kali dianalisis dalam disertasi Ph.D. (Waugh, 1995). Dataset ini menunjukkan 4.177 spesimen abalon yang dikumpulkan dari Tasmania, Australia. Setiap spesimen dicirikan oleh sembilan atribut. Abalon dikumpulkan di dua wilayah, Bass Strait (1.617 sampel) dan St. Helens (2.560 sampel). Pertumbuhan abalone Bass Strait sangat lambat, sedangkan abalone St. Helens tumbuh lebih cepat (Waugh, 1995).

Nilai numerik dalam kumpulan data diskalakan dengan faktor 200, dan variabel Rings adalah proksi indikator usia dari suatu abalone, dengan usia abalone dalam satuan tahun kira-kira dapat dihitung dengan persamaan: Usia = 1,5 + jumlah Rings (Nash, Sellers, Talbot, Cawthorn, & Ford, 1994). Dataset ini dikembangkan oleh (Nash, 1995) dan kemudian diberikan kepada (Waugh, 1995) yang melakukan pra-proses dan menggunakannya dalam riset Ph.D. Pra-pemrosesan termasuk menghilangkan titik sampel dengan nilai NULL dan menghapus kolom Site ID yang dia pilih untuk tidak digunakan. Waugh memberikan kumpulan data yang telah diproses sebelumnya dengan Site ID dihapus, ke UC Irvine untuk kepentingan peneliti di masa mendatang. Repositori data UC Irvine saat ini mengutip 31 makalah yang mengambil referensi dari dataset ini.

Atribut	Deskripsi
Sex	Jenis kelamin abalone, M untuk laki-laki dewasa, F untuk perempuan dewasa, dan I untuk infant (bayi).
Length	Panjang cangkang abalone, diukur dalam satuan 200*milimeter.
Diameter	Diameter cangkang abalone, diukur dalam satuan 200*milimeter.
Height	Tinggi cangkang abalone, diukur dalam satuan 200*milimeter.
Whole_weight	Berat dari abalone utuh, diukur dalam satuan 200*gram.
Shucked_weight	Berat dari abalone yang dikupas dari cangkangnya, diukur dalam satuan 200*gram.
Viscera_weight	Berat dari jeroan/isi perut abalone, diukur dalam satuan 200*gram.
Shell_weight	Berat dari cangkang abalone yang dikeringkan, diukur dalam satuan 200*gram.
Rings	Jumlah cincin pada cangkang abalone.

2.2 Pra-Pemrosesan

Data Sebelumnya, dataset ini telah melalui tahap pra-pemrosesan data oleh (Waugh, 1995). Sehingga tidak perlu dilakukan pra-pemrosesan data lebih lanjut pada dataset ini, karena korelasi yang tinggi antara semua variabel ukuran panjang dan berat abalone, setiap variabel dapat digunakan

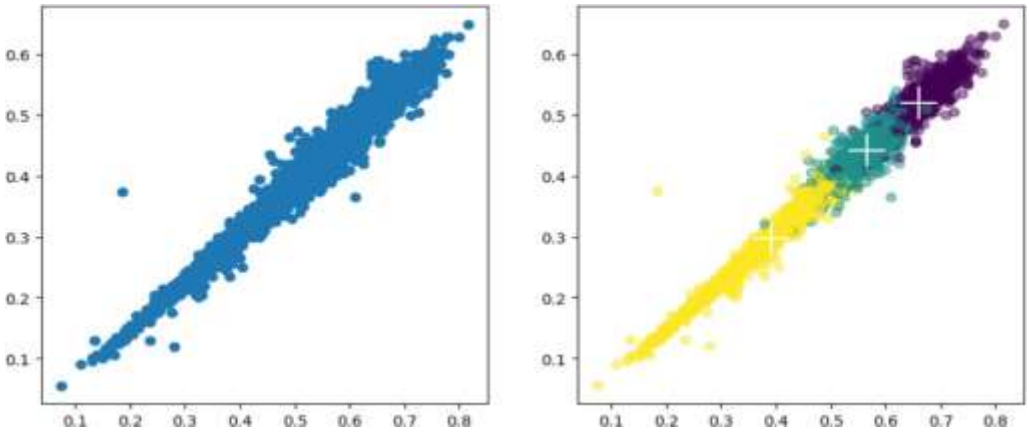
ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

dalam perhitungan dan proses clustering. Salah satu variabel yang memiliki pengaruh besar adalah *Length*, karena larangan memancing abalone di Tasmania didasarkan pada panjang abalone tersebut. Tidak ada titik sampel tertentu yang akan dihapus pada penelitian ini. Dua poin sampel dengan nilai *Height* 0,0 dipertahankan karena nilai atribut lainnya tampak masuk akal, dan juga variabel *Height* itu sendiri tetap dipertahankan.

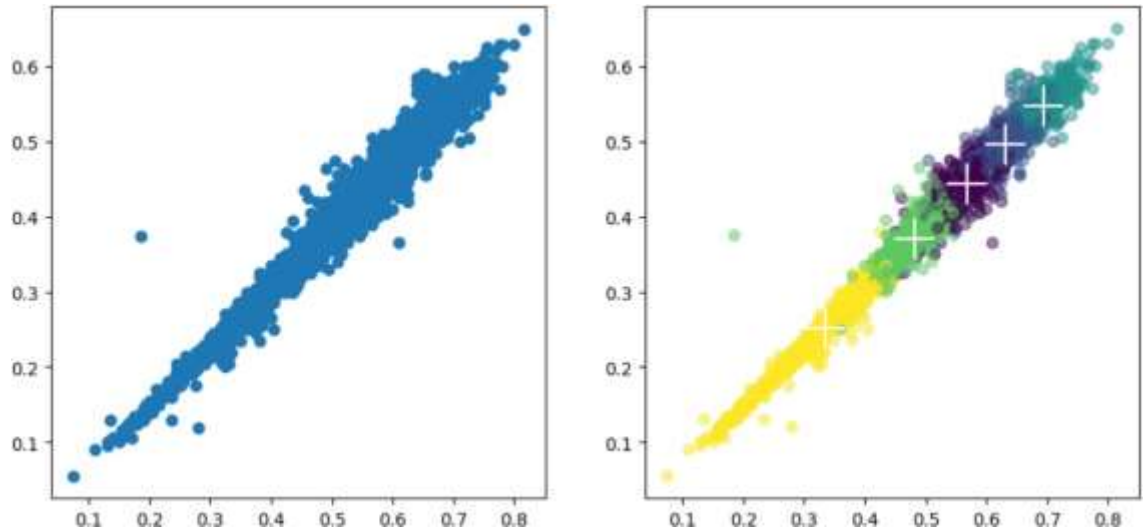
2.3 Eksperimen

Algoritma Fuzzy C-Means akan melakukan clustering pada variabel-variabel yang telah ditentukan. Dengan menggunakan library "scikit-learn", dataset dimasukkan untuk diproses lebih lanjut sehingga dapat menjalankan eksperimen clustering ini.

Pertama, peneliti melakukan pengelompokan data dengan 3 cluster yang menampilkan visualisasi data berupa scatter plot seperti berikut ini:

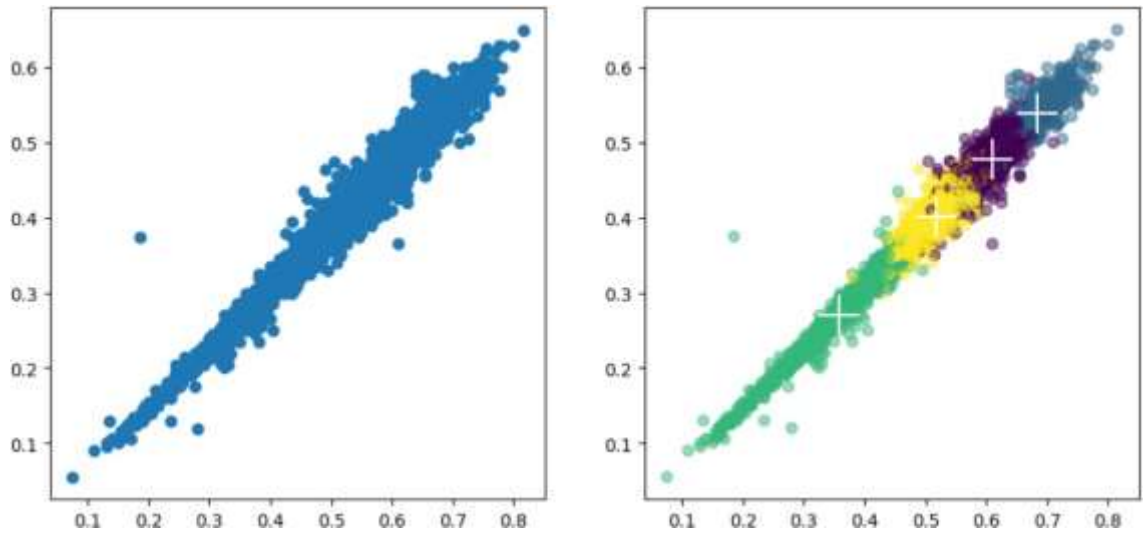


Kedua, peneliti melakukan pengelompokan data dengan 4 cluster yang menampilkan visualisasi data berupa scatter plot seperti berikut ini:

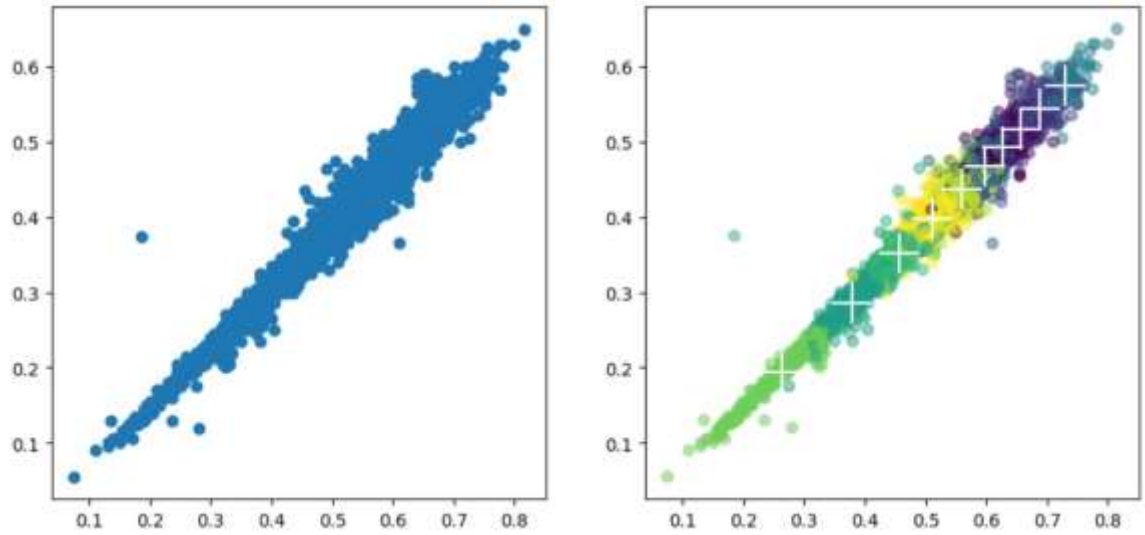


Dapat dilihat adanya perbedaan area scatter plot pada cluster yang ditunjukkan dengan warna hijau dan kuning. Tetapi, secara keseluruhan dengan ditambahnya 1 cluster tidak memengaruhi hasil eksperimen secara signifikan.

Ketiga, peneliti melakukan pengelompokan data dengan 5 cluster yang menampilkan visualisasi data berupa scatter plot seperti berikut ini:



Mulai terdapat collision yang terjadi antara beberapa cluster seperti yang ditunjukkan pada graph scatter plot.Terakhir, peneliti melakukan pengelompokan data dengan 10 cluster yang menampilkan visualisasi data berupa scatter plot seperti berikut ini:



Pada area atas-kanan graph terjadi penumpukan oleh cluster 5, 6, 7, 8, 9, dan 10, yang disebabkan oleh similaritas/kemiripan yang dimiliki oleh beberapa cluster tersebut.

2.4 Evaluasi

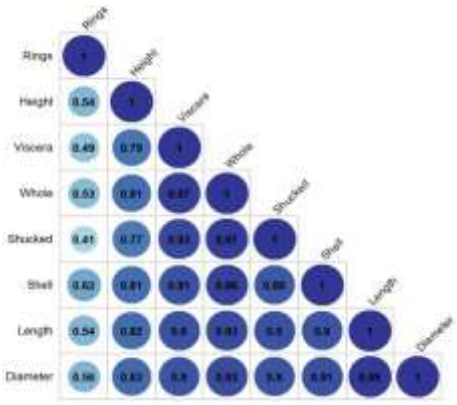
Nilai rata-rata silhouette yang dihasilkan pada proses 3 cluster adalah 0.5070986751977092 dengan nilai terbaiknya adalah 0.7354594251441374.
Nilai rata-rata silhouette yang dihasilkan pada proses 4 cluster adalah 0.4836773513086767 dengan nilai terbaiknya adalah 0.7340045286374763.
Nilai rata-rata silhouette yang dihasilkan pada proses 5 cluster adalah 0.4680399407197755 dengan nilai terbaiknya adalah 0.7297426054889078.
Nilai rata-rata silhouette yang dihasilkan pada proses 10 clusteradalah 0.38096701031011404 dengan nilai terbaiknya adalah 0.7062221556716443.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Jika melihat hasil *fuzzy silhouette* di atas, dapat disimpulkan hasil *clustering* terbaik pada dataset penelitian ini dapat diraih dengan jumlah cluster yang proporsional, pada kasus ini jumlah cluster yang lebih kecil. Semakin sedikit cluster yang diproses, semakin baik nilai *silhouette* yang dihasilkan. Sebaliknya, semakin banyak cluster yang diproses, nilai silhouette yang dihasilkan tidak sebaik sebelumnya. Namun begitu, banyaknya cluster pada eksperimen ini tidak memengaruhi hasil penelitian secara signifikan.

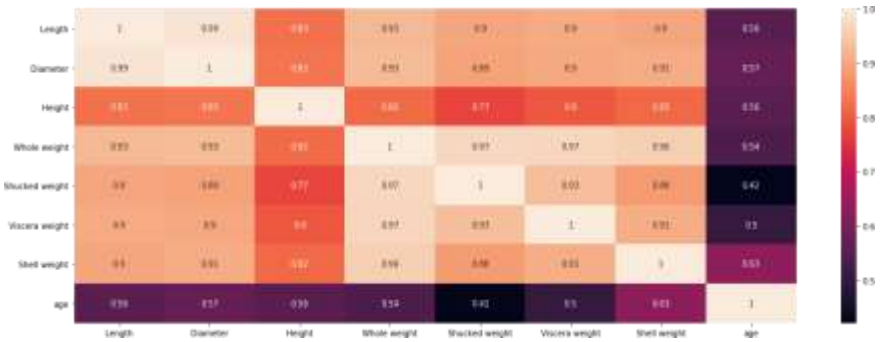
2.5 Analisis

Dari hasil evaluasi di atas, pengelompokan (clustering) data dengan menggunakan algoritma Fuzzy C-Means memberikan hasil yang lebih baik dibandingkan dengan algoritma K-Means karena memiliki ambiguitas titik data. Algoritma ini dapat menangani sebaran data yang mengalami overlap yang tidak bisa ditangani oleh K-Means di mana algoritma K-Means lebih berfokus pada meminimalisir terjadinya *class imbalance* (Hartono, S., Tulus, & Nababan, 2018).



Shell Weight memiliki korelasi terbesar dengan *Rings*, sedangkan *Shucked Weight* memiliki korelasi terkecil. Dengan menghitung berat total, berat tubuh tanpa cangkang, berat jeroan, dan berat cangkang, kita dapat memprediksi jumlah cincin pada kerang abalone. Dengan mengetahui jumlah cincinnya, kita dapat memprediksi usia dari abalone ini (Tiap 1 cincin pada kerang = 1,5 tahun usia kerang). Ditambah lagi, variabel *diameter* sangat jelas dapat membedakan jenis kelamin kerang.

Variabel *Whole weight* hampir berselang-seling secara linear dengan semua fitur lainnya kecuali usia, Variabel *Height* memiliki linearitas paling sedikit dengan fitur yang tersisa. Usia abalone dapat ditentukan secara linear dan proporsional dengan *Shell weight*, *Diameter*, dan *Length*. Sedangkan *Shucked weight* adalah variabel yang paling tidak berkorelasi dengan usia abalone.Dilihat dari hasil analisa di atas, dapat ditentukan dan dipastikan bahwa ukuran panjang dan berat kerang adalah faktor yang paling menentukan banyaknya cincin yang dapat ditemukan di cangkang abalone. Dan dari jumlah cincin abalone, peneliti dapat menghitung usia suatu abalone dengan persamaan 1,5 + jumlah cincin (dalam satuan tahun).



3. KESIMPULAN

Clustering yang merupakan salah satu teknik *Machine Learning* dapat menyelesaikan masalah yang ada pada kehidupan sehari-hari. Salah satu algoritma *clustering* yang banyak dipakai yaitu *Fuzzy C-Means* dapat memberikan hasil yang memuaskan dari proses clustering dengan adanya fitur *fuzzy silhouette.* Dengan dilakukannya eksperimen ini, dapat dipastikan bahwa jumlah cincin pada cangkang abalone sangat dipengaruhi dari ukuran panjang dan berat abalone tersebut. Serta dari jumlah cincin cangkang abalone yang ditemukan, suatu abalone dapat dipastikan usianya.

Peneliti berharap agar penelitian ini dapat memastikan pengetahuan bagi para nelayan pencari abalone agar dapat menangkap serta memperhatikan sisi kelestarian dari spesies Abalone. Dengan adanya pengetahuan mengenai populasi abalone, diharapkan pemerintah setempat dapat membuat kebijakan untuk memanen dengan usia spesies tertentu guna menjaga siklus perkembangbiakan spesies ini.

DAFTAR PUSTAKA

[1] Bai, L., Liang, J., & Dang, C. (2012). *A Modified K-Modes Clustering Algorithm.*
[2] Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: The Fuzzy C-Means Clustering Algorithm. *Computers & Geosciences*, 191-203.
[3] Hartono, S., S. O., Tulus, & Nababan, E. B. (2018). Optimization Model of K-Means Clustering Using Artificial. *IOP Conf. Series: Materials Science and Engineering*, p. 288.
[4] Nash, W., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., & Ford, W. B. (1994). The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 411.
[5] Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to Data Mining.* India: Pearson Education.