

Implementasi dan Analisis Perbandingan Algoritma *Machine Learning* dalam Penentuan Keanggotaan Gugus Bintang NGC 7790

Laporan Research Based Learning
SK5002 Algoritma Pemrograman dalam Sains

Dosen Pengampu : Dr. Wahyu Hidayat, S.Si., M.Si.



Oleh: Kelompok 2

10322015 Ulivia Embun Tresna Wardani

10322041 Fatiha Izza Tunisa

10323024 Luthfiana Sutarjo

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI BANDUNG

2025

Abstrak

Gugus bintang merupakan objek penting dalam studi astronomi karena terdiri atas bintang-bintang dengan karakteristik fisik dan kimia yang serupa serta kemungkinan besar terbentuk dari awan molekul yang sama. Salah satu di antaranya adalah NGC 7790, gugus bintang terbuka yang terletak di rasi Cassiopeia. Dalam pengamatan astronomi, data yang diperoleh tidak hanya mencakup bintang anggota gugus, tetapi juga bintang latar yang tidak berasosiasi secara fisik. Oleh karena itu, penentuan keanggotaan bintang menjadi tahap penting sebelum analisis lanjutan dilakukan. Penelitian ini menerapkan pendekatan *machine learning* untuk menentukan keanggotaan gugus NGC 7790 secara efisien. Data yang digunakan berasal dari Gaia DR3 milik European Space Agency (ESA), yang menyediakan pengukuran astrometri presisi tinggi untuk lebih dari satu miliar bintang di Galaksi Bima Sakti. Parameter yang digunakan meliputi posisi, paralaks, dan gerak diri bintang. Parameter-parameter ini menggambarkan bagaimana bintang-bintang tersebar secara spasial dan bagaimana pergerakannya. Hal itu membantu menentukan apakah mereka terikat secara gravitasi sebagai bagian dari gugus yang sama. Data diperoleh melalui perangkat lunak TOPCAT, kemudian dilakukan *pre-processing* untuk membersihkan data guna meningkatkan akurasi pengelompokan. Penelitian ini menerapkan empat algoritma pembelajaran mesin, yaitu Random Forest dan Support Vector Machine sebagai pendekatan *supervised learning*, serta K-Means dan HDBSCAN sebagai pendekatan *unsupervised learning*. Pendekatan ini tidak hanya digunakan untuk mengelompokkan bintang berdasarkan kemiripan karakteristik fisik dan kinematik, tetapi juga untuk membandingkan performa tiap metode dalam hal akurasi, efisiensi, dan ketahanan terhadap noise. Hasil penelitian ini diharapkan dapat menjadi referensi bagi pengembangan sistem otomatis dalam penentuan keanggotaan gugus bintang.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Astronomi, sebagai salah satu ilmu pengetahuan tertua, mempelajari benda-benda langit untuk memahami proses fundamental pembentukan dan evolusi bintang serta dinamika galaksi. Dalam konteks ini, gugus bintang (*star cluster*) menjadi objek studi yang krusial. Gugus bintang, sebagai kumpulan bintang yang terikat secara gravitasi, memberikan wawasan mendalam mengenai tahap awal pembentukan bintang. Dalam astrofisika, gugus bintang (*star cluster*) berfungsi sebagai laboratorium alam untuk meneliti proses pembentukan dan evolusi bintang. Melalui pengelompokan bintang ke dalam gugus, astronom dapat mempelajari karakteristik yang terbentuk dari lingkungan yang sama, seperti usia, komposisi kimia, dan dinamika gravitasi. Namun, pada masa lalu, pengenalan anggota gugus bintang dilakukan dengan cara yang sangat terbatas. Langit diamati dalam dua dimensi, dan jarak antar bintang dihitung secara manual melalui paralaks atau perbedaan kecerlangan. Akibatnya, banyak ketidakpastian muncul dalam menentukan apakah suatu bintang termasuk dalam gugus tertentu atau tidak.

Kini, dengan hadirnya misi seperti *Gaia* dari European Space Agency (ESA), astronom memiliki akses pada data posisi, jarak, dan kecepatan bintang yang sangat presisi dan berskala besar. Kemajuan teknologi ini telah merevolusi cara pengumpulan data astronomi, menghasilkan data dalam volume yang sangat masif dengan ukuran mencapai *terabyte* hingga *petabyte*. Analisis manual terhadap set data sebesar ini secara praktis tidak mungkin dilakukan, sehingga astronomi modern telah bertransformasi menjadi ilmu yang sangat bergantung pada metode analisis data komputasi untuk menyaring, memproses, dan mengekstraksi informasi ilmiah yang berharga dari lautan data tersebut. Tantangan ini bukan hanya soal volume, tetapi juga kompleksitas data yang sering kali mengandung klaster dengan berbagai bentuk, kepadatan, dan tingkat kontaminasi dari bintang latar depan atau belakang yang tidak terkait.

Machine learning (ML) hadir menawarkan kemampuan untuk mengenali pola, mengelompokkan data, dan melakukan prediksi tanpa perlu eksplisit dirumuskan dengan persamaan fisika yang kompleks. Dalam konteks klusterisasi bintang, ML dapat membantu menentukan keanggotaan gugus berdasarkan kemiripan posisi, kecepatan, dan parameter fotometrik lainnya. Dalam program ini digunakan metode *supervised learning*, yaitu dengan Random Forest dan Support Vector Machine (SVM), untuk klasifikasi keanggotaan gugus dengan data berlabel. Kedua metode *supervised learning* ini dikenal memiliki performa tinggi dalam menghadapi data non-linear dan multi-dimensi. Selain itu, digunakan juga metode *unsupervised learning*, yaitu dengan K-Means dan HDBSCAN, untuk mengelompokkan data tanpa label. Keduanya merupakan metode klusterisasi yang umum digunakan karena

K-Means unggul dalam kesederhanaan dan kecepatan, sedangkan HDBSCAN mampu mendeteksi kluster dengan bentuk tidak beraturan serta lebih robust terhadap *noise*.

Penggunaan keempat metode tersebut dilakukan bukan hanya untuk membentuk sistem identifikasi gugus bintang yang efisien, tetapi juga untuk membandingkan akurasi dan kinerja masing-masing metode dalam konteks data astronomi. Dengan begitu, penelitian ini dapat memberikan gambaran yang komprehensif tentang metode mana yang paling sesuai untuk mengklasifikasi atau mengelompokkan bintang berdasarkan karakteristiknya. Secara keseluruhan, pengembangan program ini diharapkan dapat mempermudah astronom dalam menganalisis data bintang berukuran besar, meningkatkan efisiensi proses identifikasi gugus, dan membuka peluang lebih luas bagi penerapan kecerdasan buatan dalam penelitian astrofisika masa depan.

1.2 Rumusan Masalah

1. Bagaimana performa algoritma supervised (Random Forest, SVM) dibandingkan dengan unsupervised (K-Means, HDBSCAN) dalam menentukan anggota gugus NGC 7790 berdasarkan parameter astrometri?
2. Seberapa efektif penggunaan algoritma unsupervised learning dalam menangani bintang latar belakang (field star) pada gugus NGC 7790 tanpa memerlukan training data?
3. Manakah di antara keempat metode machine learning tersebut yang paling optimal untuk diterapkan pada penentuan anggota gugus bintang dengan karakteristik seperti NGC 7790?

1.3 Tujuan Penelitian

1. Mengimplementasikan empat algoritma machine learning, yaitu Random Forest, SVM, K-Means, dan HDBSCAN untuk memisahkan anggota gugus NGC 7790 dari bintang latar belakang.
2. Mengevaluasi kinerja setiap metode berdasarkan parameter akurasi dan efisiensi waktu komputasi.
3. Menganalisis perbedaan karakteristik hasil antara pendekatan supervised dan unsupervised learning dalam menangani data astrometri.
4. Menentukan metode yang paling optimal untuk diterapkan pada kasus penentuan keanggotaan gugus bintang dengan karakteristik seperti NGC 7790.

1.4 Manfaat Penelitian

Penelitian ini diharapkan memberikan beberapa manfaat sebagai berikut:

1. Ilmiah: Menambah wawasan mengenai penerapan *machine learning* dalam astronomi serta memberikan perbandingan kinerja empat metode ML dalam klusterisasi gugus bintang.

2. Praktis: Menghasilkan program yang mempermudah proses identifikasi gugus bintang secara otomatis dan efisien.
3. Edukatif: Menjadi sarana pembelajaran penerapan *supervised* dan *unsupervised learning* dalam bidang sains, khususnya astrofisika.

1.5 Batasan Masalah

Untuk menjaga agar penelitian ini tetap fokus dan dapat diselesaikan dalam jangka waktu yang ditentukan, ditetapkan beberapa batasan sebagai berikut:

1. Gugus bintang yang menjadi objek *clustering* adalah NGC 7790
2. Analisis hanya difokuskan pada empat metode *machine learning*, yaitu dua metode *supervised learning* (Random Forest dan Support Vector Machine (SVM)) serta dua metode *unsupervised learning* (K-Means dan HDBSCAN). Metode lain di luar keempat algoritma tersebut tidak dibandingkan.
3. Parameter evaluasi yang digunakan dalam perbandingan performa metode terbatas pada metrik umum seperti akurasi dan efisiensi komputasi (waktu proses). Aspek lain seperti interpretabilitas model atau kompleksitas matematis tidak menjadi fokus utama.
4. Proses pelabelan data (untuk metode *supervised*), model ditrainning dengan dataset NGC 1624 yang memiliki karakteristik sama dengan NGC 7790.
5. Implementasi program difokuskan pada pengembangan sistem klusterisasi berbasis komputasi, bukan pada pengamatan atau pengambilan data astronomi secara langsung.
6. Analisis hasil difokuskan pada kemampuan metode dalam mengenali dan membedakan kluster bintang, bukan pada studi mendalam mengenai sifat fisik bintang itu sendiri (seperti suhu, massa, atau evolusinya).

BAB II

TINJAUAN PUSTAKA

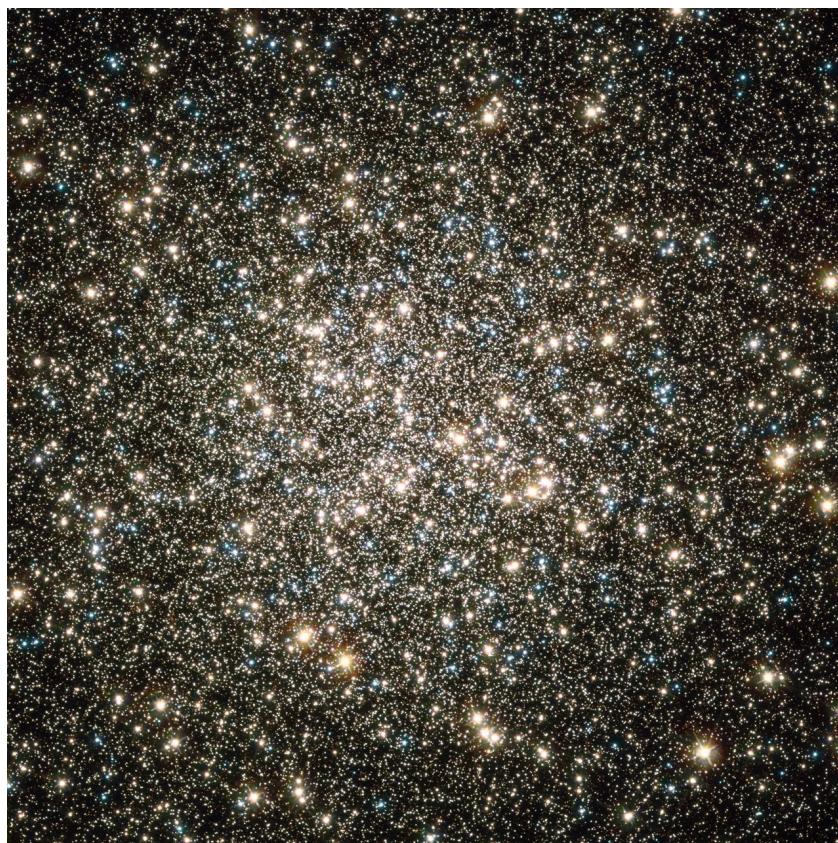
2.1 Konsep Gugus Bintang dan Signifikansinya

Gugus bintang didefinisikan sebagai sekelompok bintang yang terikat secara gravitasi satu sama lain. Objek ini memberikan wawasan yang sangat penting mengenai proses pembentukan dan evolusi bintang, serta dinamika galaksi secara keseluruhan. Karena bintang-bintang dalam satu gugus terbentuk dari awan molekuler yang sama dan pada waktu yang hampir bersamaan sehingga mereka memiliki komposisi kimia dan usia yang serupa. Hal ini menjadikan gugus bintang sebagai laboratorium alami yang ideal untuk menguji teori-teori evolusi bintang.

Tabel 2.1.1: Tipe gugus bintang dan karakteristiknya (Fraknoi dkk, 2022)

Karakteristik	Gugus Bola	Gugus Terbuka	Asosiasi
Jumlah di Galaksi	150	Ribuan	Ribuan
Lokasi di Galaksi	Halo dan <i>central bulge</i>	Piringan (dan lengan spiral)	Lengan spiral
Diameter (tahun cahaya)	50 - 450	< 30	100 - 500
Massa (M_{\odot})	$10^4 - 10^6$	$10^2 - 10^3$	$10^2 - 10^3$
Jumlah Bintang	$10^4 - 10^6$	50 - 1000	$10^2 - 10^4$
Warna	Merah	Merah atau biru	Biru
Kecerlangan (L_{\odot})	$10^4 - 10^6$	$10^2 - 10^6$	$10^4 - 10^7$
Usia (tahun)	Milyaran	Ratusan juta tahun dan beberapa berusia milyaran	Hingga puluhan juta

2.1.1 Gugus Bola

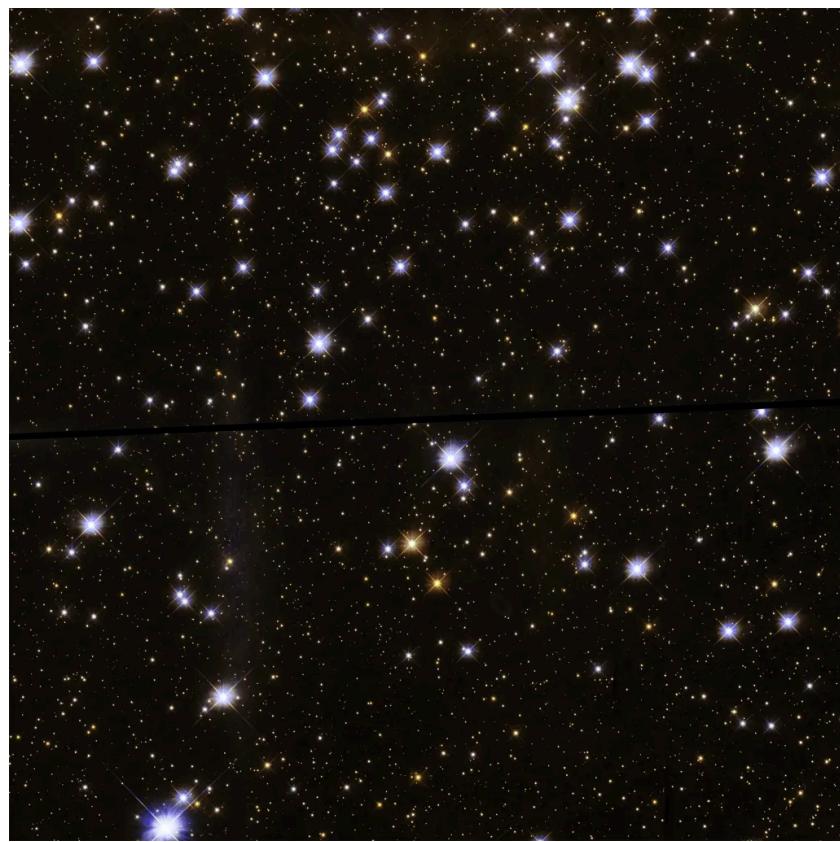


Gambar 2.1.1: Gugus Hercules (M13). Kredit: NASA, ESA, and the Hubble Heritage Team

Gugus bola memiliki bentuk yang hampir simetri radial dengan puluhan ribu hingga jutaan bintang yang memiliki ikatan gravitasi sangat kuat. Kebanyakan bintang pada gugus ini berumur tua dan miskin logam (bintang populasi II) yang ditandai dengan warna kemerahan. Hal ini menunjukkan bahwa evolusi bintangnya telah melewati deret utama. Gugus ini terletak pada bagian halo dan *central bulge* Galaksi Bimasakti. Contoh dari gugus ini adalah gugus Hercules (M13) di rasi Hercules.

2.1.2 Gugus Terbuka

Gugus terbuka adalah kelompok bintang yang terdiri dari puluhan hingga ribuan bintang dengan ikatan gravitasi yang tidak terlalu kuat. Gugus terbuka berisi bintang muda dan panas yang lebih biru dibandingkan pada gugus bola. Struktur gugus yang terbuka dan menyebar ini menyebabkan gugus ini tidak terlalu stabil sehingga memungkinkan bintang-bintang penyusunnya akan tersebar setelah beberapa juta tahun. Gugus ini terletak pada bagian piringan dan lengan spiral Galaksi Bimasakti. Salah satu contoh gugus terbuka adalah *Wild Duck Cluster* (M11) di rasi Scutum.



Gambar 2.1.2: *Wild Duck Cluster* (M11). Kredit: NASA, ESA, STScI and P. Dobbie.

2.1.3 Asosiasi

Asosiasi merupakan kelompok bintang yang ikatan gravitasi antar bintangnya sangat lemah. Gugus ini terdiri dari ratusan hingga puluhan ribu bintang yang merupakan bintang berusia sangat muda, biasanya tipe O dan B yang tersebar dengan diameter 100 hingga 500 tahun cahaya. Umumnya ditemukan di dekat awan molekul tempat mereka terbentuk atau daerah yang kaya akan gas dan debu antar bintang. Salah satu contoh asosiasi bintang adalah Perseus OB-1.



Gambar 2.1.3: Perseus OB-1. Kredit: Stellar Journey.

2.2 NGC 7790

Gugus bintang NGC 7790 adalah gugus terbuka di rasi Cassiopeia dengan RA $23^{\text{h}} 58^{\text{m}} 24^{\text{s}}$ dan deklinasi $+61^{\circ} 12' 30''$. Gugus ini penting dalam astrofisika karena memiliki tiga bintang variabel Cepheid di dalamnya, yaitu CEa Cas, CEb Cas and CF Cas (Sandage 1958). Dengan adanya Cepheid dalam gugus ini dapat digunakan untuk kalibrasi relasi periode-luminositas dan verifikasi jarak dari data fotometri. Bintang ini diklasifikasikan sebagai gugus Trumpler kelas II2m oleh Lyngå (1987), yang berarti gugus bintang ini terpisah dan cukup terkonsentrasi ke arah pusat, rentang kecerlangan sedang, dan anggotanya cukup banyak. Jarak kluster ini sekitar $3,3 \pm 0,23$ Kpc dan berumur sekitar 120 ± 20 Myr menurut Gupta, A. C., et al. (2000). Gugus ini didominasi oleh bintang deret utama bertipe spektral B dan A, dengan sebagian bintang F awal pada magnitudo yang lebih redup. Berdasarkan katalog gugus terbuka terbaru yang memanfaatkan data Gaia EDR3/DR3, jumlah anggota gugus ini berada pada orde ratusan bintang, dengan estimasi sekitar 200–300 anggota bergantung pada ambang probabilitas keanggotaan yang digunakan (Castro-Ginard et al. 2022).

2.3 *Unsupervised Machine Learning* untuk Clustering

Unsupervised machine learning adalah cabang dari kecerdasan buatan yang bertujuan menemukan pola atau struktur intrinsik dalam data tanpa adanya label atau target yang telah ditentukan sebelumnya. Salah satu tugas utamanya adalah *clustering* (pengelompokan), yaitu proses mempartisi set data menjadi beberapa kelompok (klaster) sedemikian rupa sehingga

objek dalam satu klaster memiliki kemiripan yang tinggi satu sama lain dan berbeda dengan objek di klaster lain.

2.2.1 Metode K-Means

Algoritma ini merupakan salah satu metode *clustering* yang paling populer dan berbasis sentroid. K-Means bekerja dengan mengelompokkan data dengan cara meminimalkan jarak Euklides antara setiap titik data ke pusat gugus (*centroid*) terdekatnya. Keunggulan utama K-Means terletak pada efisiensi komputasinya, terutama pada set data berukuran besar, karena kompleksitas waktunya cenderung linear terhadap jumlah data.

2.2.2 Metode HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) adalah algoritma *clustering* canggih yang berbasis kepadatan. Berbeda dengan K-Means, HDBSCAN tidak memerlukan penentuan jumlah gugus di awal. Algoritma ini mampu membentuk gugus dengan berbagai bentuk dan kepadatan secara otomatis, serta secara eksplisit mengidentifikasi titik-titik data yang tidak termasuk dalam klaster mana pun sebagai *noise* atau *outlier*. Fleksibilitas ini menjadikan HDBSCAN sangat cocok untuk data astronomi yang kompleks dan sering kali mengandung anomali.

2.4 Supervised Machine Learning untuk Clustering

Supervised machine learning adalah bagian dari kecerdasan buatan yang dapat digunakan untuk menemukan pola tertentu dari data latih (training set) yang berlabel. Misalnya, terdapat data bintang yang telah ditentukan merupakan anggota gugus atau bukan. Kemudian model dilatih untuk belajar mengenali pola dari fitur-fitur data latihan tersebut. Setelah itu barulah model dapat digunakan untuk memprediksi keanggotaan bintang lain yang belum diketahui labelnya.

2.3.1 Metode Random Forest (RF)

Algoritma ini merupakan salah satu metode *machine learning* yang berbasis *decision tree*, dimana pohon keputusan ini dilatih dengan data dan fitur-fiturnya pada subset acak. Random Forest dapat menghasilkan "matriks proksimitas" atau kedekatan antar sampel data. Prediksi akhir ditentukan dari tingkat kemiripan antar fitur data tersebut. Metode ini mampu menangani data berdimensi tinggi dan memberikan estimasi *feature importance*, yang membantu mengidentifikasi fitur paling berpengaruh untuk menentukan perbedaan antar bintang. Algoritma ini unggul dalam menangani data besar dan berdimensi tinggi serta menghasilkan akurasi yang baik dalam mengidentifikasi jenis objek langit yang berbeda.

2.3.2 Metode Support Vector Machine (SVM)

SVM merupakan metode yang bekerja dengan mencari garis batas (hyper plane) terbaik yang memisahkan dua kelas data dengan margin maksimum. Dalam konteks gugus bintang, SVM dapat digunakan untuk *clustering* berdasarkan data fitur astrometri dan fotometri bintang.

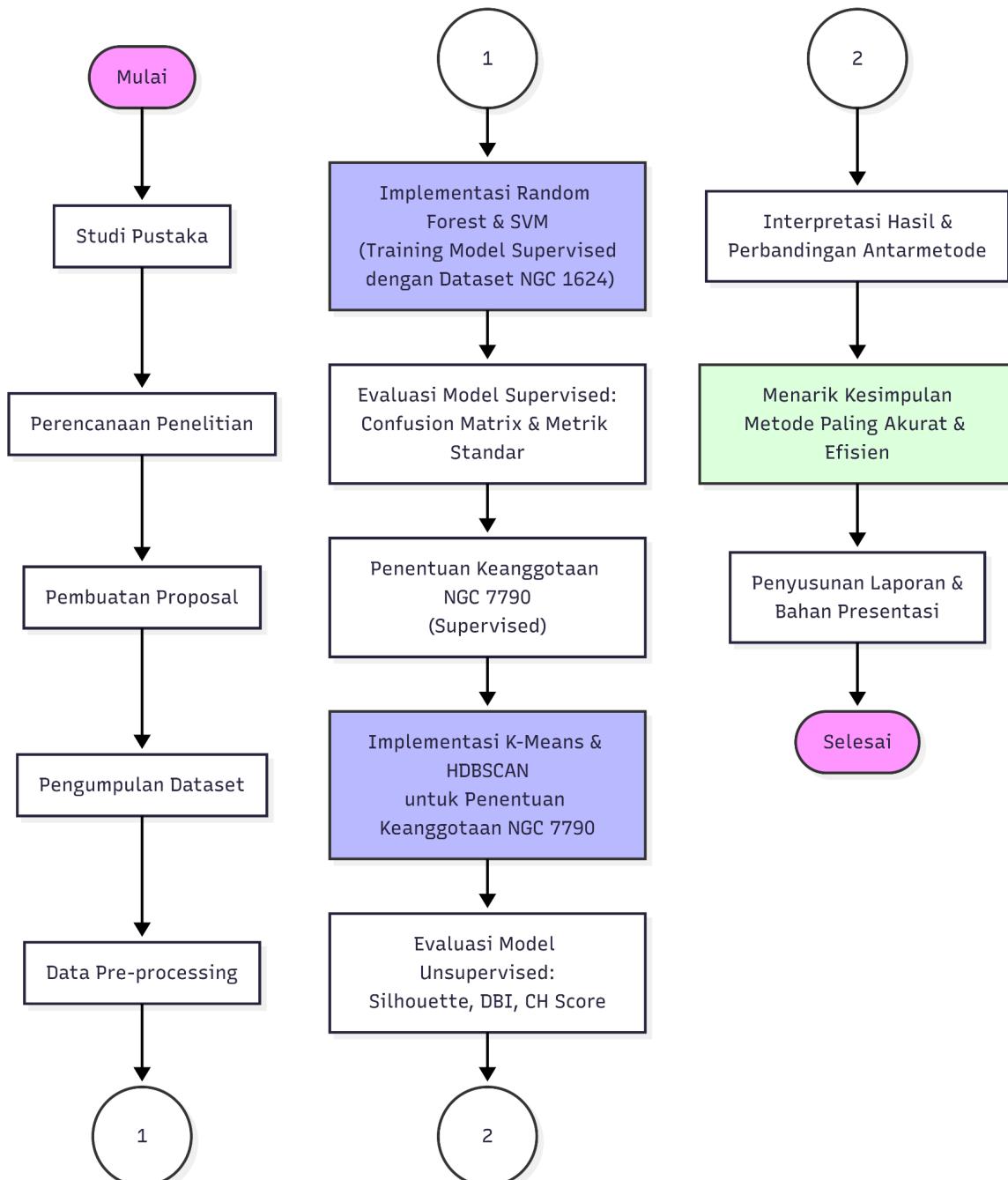
Untuk dua kelas data yang tidak dapat dipisahkan secara linear, SVM menggunakan fungsi kernel untuk memetakan data ke ruang berdimensi lebih tinggi untuk dapat membuat pemisahan linear. Output dari metode ini berupa nilai keputusan atau probabilitas keanggotaan. Namun, metode ini tidak efisien untuk dataset yang sangat besar atau memiliki banyak noise.

BAB III

METODOLOGI DAN RANCANGAN PENELITIAN

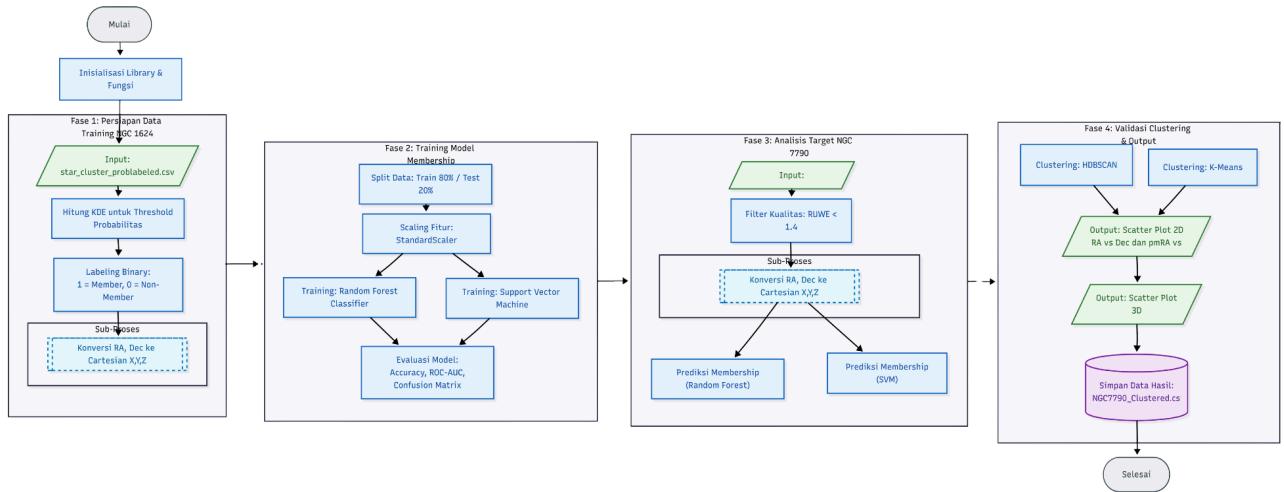
3.1 Rancangan Penelitian

Penelitian ini dilaksanakan mengikuti alur kerja yang terstruktur dan sistematis, sebagaimana digambarkan dalam diagram alur berikut.



Gambar 3.1.1. *Flowchart* garis besar penelitian

3.2 Flowchart Program



Gambar 3.2.1. *Flowchart* dari algoritma yang digunakan

3.3 Pengumpulan dan *Data Pre-processing*

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari survei Gaia DR3 (Data Release 3), sebuah misi dari European Space Agency (ESA) yang menyediakan pengukuran astrometri dan fotometri dengan presisi tinggi untuk lebih dari satu miliar bintang di Bima Sakti. Gaia memindai seluruh langit, mengumpulkan data mentah berskala seluruh langit yang mencakup baik bintang-bintang anggota gugus maupun bintang latar belakang yang tidak terkait. Data tersebut diambil menggunakan perangkat lunak TOPCAT.

2. Deskripsi Data dan Fitur Penting

Data mentah yang diperoleh dapat diakses [di sini](#). Dari sekian banyak parameter yang tersedia dalam dataset Gaia, hanya beberapa yang dianggap relevan dan digunakan untuk analisis antara lain, posisi, paralaks, dan gerak diri (proper motion). Parameter-parameter ini menggambarkan bagaimana bintang-bintang tersebar secara spasial dan bagaimana pergerakannya, yang membantu menentukan apakah mereka terikat secara gravitasi sebagai bagian dari gugus yang sama.

3. *Data Pre-processing*

Sebelum dianalisis, data mentah diproses melalui beberapa tahap untuk memastikan kualitas dan kesiapannya. Langkah-langkah ini meliputi:

- **Data Cleaning:** Membersihkan set data dari nilai yang hilang (*missing values*), data duplikat, atau *outlier* yang jelas dapat mengganggu analisis.

Dataset kemudian disaring lebih lanjut menggunakan parameter **RUWE**, dengan hanya mempertahankan sumber yang memiliki **RUWE < 1.4**, karena nilai yang lebih tinggi sering menunjukkan kualitas astrometri yang buruk. Untuk menghindari ketidakkontinuan sudut di sekitar **0° dan 360°**, nilai **right ascension** dan **declination** dikonversi ke dalam **koordinat Kartesius**, sehingga memungkinkan perhitungan jarak spasial yang lebih akurat selama proses klasterisasi.

- **Standardization:** Mengubah skala fitur-fitur data numerik ke skala yang seragam, seperti menggunakan *z-score*. Langkah ini krusial untuk K-Means, yang kinerjanya sangat sensitif terhadap skala data karena mengandalkan metrik jarak Euklides. Meskipun kurang kritis untuk HDBSCAN yang berbasis kepadatan, standardisasi tetap diterapkan untuk memastikan komparabilitas fitur yang adil di kedua model.
- **Importance Feature:** Mengekstrak fitur-fitur penting yang berpengaruh pada proses penentuan anggota gugus bintang. Fitur-fitur ini tidak lain adalah seperti yang sudah dijelaskan pada bagian **Deskripsi Data**.

3.3 Implementasi Model Komputasi

Implementasi model komputasi dibangun menggunakan bahasa pemrograman Python dengan memanfaatkan pustaka ilmiah seperti scikit-learn, pandas, dan numpy. Struktur program dirancang dengan mengimplementasikan konsep-konsep pemrograman fundamental, seperti *conditioning*, *looping*, rekursif, dan *function*. Selain untuk menuangkan seluruh materi yang telah dipelajari di kuliah, hal ini bertujuan untuk memastikan modularitas dan efisiensi program.

1. **Struktur Program Berbasis Fungsi:** Kode diorganisir ke dalam beberapa **fungsi (function)** yang modular. Setiap fungsi memiliki satu tanggung jawab spesifik, seperti `load_data()`, `preprocess_data()`, `run_randomforest()`, `run_svm()`, `run_kmeans()`, `run_hdbSCAN()`, dan `evaluate_result()`. Pendekatan ini membuat kode lebih mudah dibaca, diuji, dikelola, dan dikembangkan lebih lanjut.
2. **Logika Algoritma dan Implementasi:**
 - Implementasi algoritma Random Forest pada dataset ini menggunakan 100-300 estimators atau pohon keputusan di dalam ensemblenya. Random Forest sangat efektif untuk menentukan keanggotaan bintang karena pendekatan ensemble learning-nya (menggunakan banyak pohon keputusan untuk membuat keputusan akhir), sehingga cukup tahan terhadap overfitting. Algoritma ini sangat cocok untuk dataset astronomi, di mana sampel pelatihan mungkin terbatas atau mengandung ketidakpastian bawaan.
 - Implementasi algoritma SVM pada dataset dengan menggunakan kernel RBF, karena kernel default “linear” tidak dapat memisahkan data tersebut. SVM dipilih karena kinerjanya yang unggul dalam ruang berdimensi tinggi, kemampuannya membentuk batas keputusan non-linear melalui transformasi kernel, serta ketahanannya terhadap outlier berkat penggunaan support vector,

yang membuatnya sangat cocok untuk data astronomi. Dalam data astronomi, ruang parameter sering kali mencakup berbagai dimensi astrometri dan fotometri, di mana populasi bintang gugus dan bintang medan tidak selalu dapat dipisahkan secara linear. Data tersebut juga sering mengandung kesalahan pengukuran serta kontaminasi dari bintang medan. Kemampuan algoritma ini untuk menemukan hiperbidang pemisah optimal bahkan di tengah keberadaan noise menjadi sangat berharga saat menganalisis anggota gugus redup yang berada di batas deteksi.

- Implementasi algoritma K-Means memanfaatkan konsep **iterasi (iteration)**. Proses inti algoritma ini secara berulang (iteratif) menghitung ulang posisi sentroid dan menetapkan kembali setiap titik data ke gugus terdekat hingga posisi sentroid tidak lagi berubah secara signifikan (konvergen).
- Logika **kondisional (conditional)**, seperti struktur if-else digunakan untuk mengelola alur program dan pengambilan keputusan. Contohnya, untuk memeriksa kriteria konvergensi algoritma, menangani data *outlier* yang diidentifikasi oleh HDBSCAN (misalnya, dengan melabelinya sebagai "noise"), atau memilih parameter optimal berdasarkan hasil evaluasi.
- Secara konseptual, proses hierarkis HDBSCAN dalam mengidentifikasi struktur kepadatan pada berbagai skala dapat dianalogikan dengan logika **rekursif (recursion)**, di mana masalah klasterisasi dipecah menjadi sub-masalah yang lebih kecil dan padat.

3.4 Evaluasi Model

Kinerja dan efektivitas keempat model *clustering* dievaluasi dan dibandingkan secara kuantitatif menggunakan *confussion matrix*. Di mana dalam hal ini kami memanfaatkan dataset berlabel dalam perhitungan akurasi untuk mengevaluasi model.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Data Pre-processing

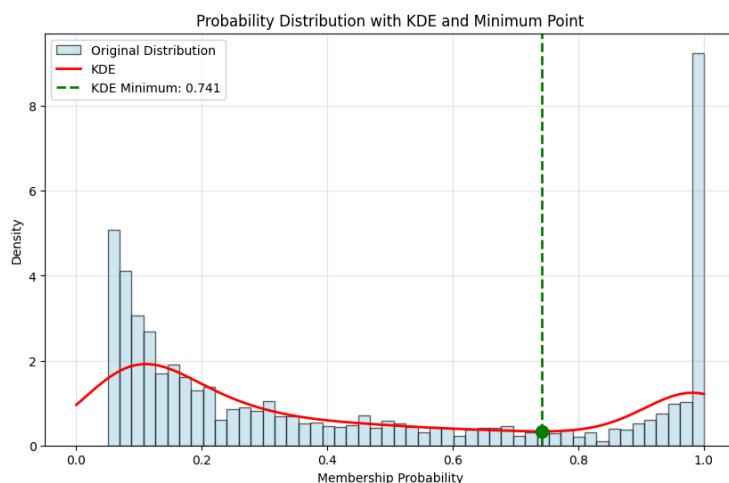
Sebelum pengimplementasian algoritma penentuan keanggotaan gugus bintang, dilakukan serangkaian tahapan pemrosesan awal (*preprocessing*). Baris data yang mengandung *missing values* pada parameter kunci dieliminasi dari dataset. Selanjutnya, dataset difilter lebih lanjut menggunakan parameter RUWE, dengan hanya mempertahankan sumber yang memiliki nilai RUWE < 1,4; hal ini dikarenakan nilai yang lebih tinggi sering kali merepresentasikan kualitas astrometri yang rendah. Guna menghindari diskontinuitas sudut pada rentang 0° hingga 360°, koordinat asensiorekta (RA) dan deklinasi dikonversi ke dalam sistem koordinat Kartesian. Transformasi ini bertujuan untuk menjamin akurasi yang lebih tinggi dalam perhitungan jarak spasial selama proses pengelompokan berlangsung.

4.2 Supervised Learning

4.2.1 Training Model

Kami memanfaatkan dataset NGC 1624 yang memiliki informasi nilai probabilitas *member* dari setiap bintangnya sebagai data *training* pada model *supervised learning*. Alasan kami menggunakan dataset objek ini dibandingkan langsung melatih model dengan dataset objek tujuan kami—NGC 7790 adalah karena tidak ditemukannya dataset dari NGC 7790 yang sudah berlabel atau setidaknya ada nilai probabilitas *member* seperti dataset NGC 1624 ini. Jadi, kami melakukan *transfer learning* dengan catatan telah memastikan karakteristik kedua objek ini mirip, sehingga model yang nantinya dihasilkan dari *training* NGC 1624 relevan untuk diterapkan dalam penentuan keanggotaan pada gugus NGC 7790.

Kami menggunakan *Kernel Density Estimation* (KDE) *minimum* dari persebaran data nilai probabilitas member NGC 1624 untuk mencari ambang batas penentuan yang menjadi member dan bukan.

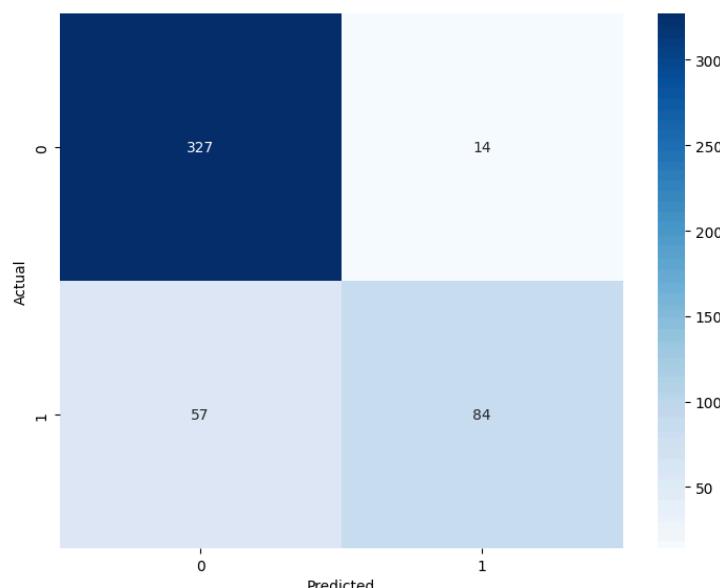


Gambar 4.2.1. Grafik Distribusi Probabilitas dengan Minimum Point KDE

Pada grafik di atas dapat dilihat bahwa nilai ambang batas yang didapatkan adalah 0,741. Artinya, bintang-bintang yang memiliki nilai probabilitas member di atas 0,741 akan dianggap sebagai *member*, dan sisanya bukan. Kemudian, kami melabeli seluruh data bintang di NGC 1624 dengan “1” untuk *member* dan “0” untuk *non-member* berdasarkan identifikasi tersebut.

Setelah dataset NGC 1624 sudah memiliki label, kami mulai melatih model Random Forest dan Support Vector Machine (SVM) dengan fitur-fitur penting yang digunakan adalah yang telah disebutkan pada Subbab 3.3.

Pada Random Forest, kami menerapkan 300 estimator saat melatih model dari dataset NGC 1624. Berikut adalah evaluasi dari model Random Forest yang dibangun.



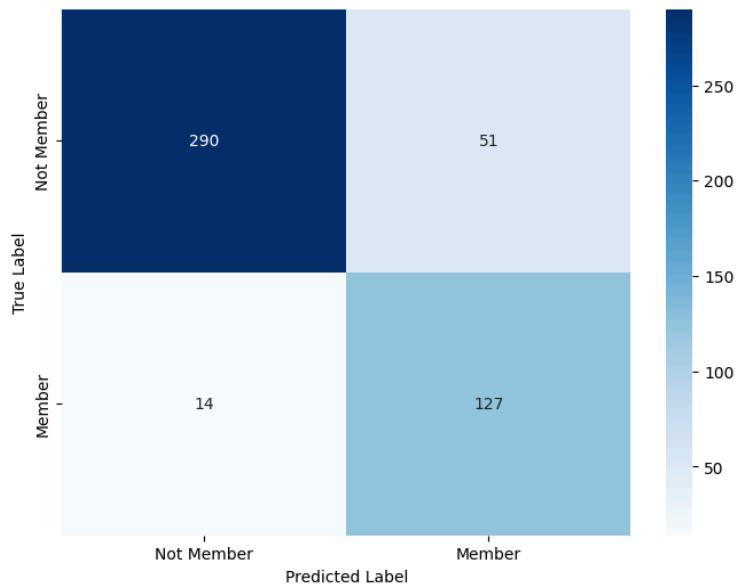
Gambar 4.2.2. *Confusion matrix* model Random Forest

Evaluasi performa melalui *confusion matrix* memberikan gambaran kuantitatif yang lebih mendetail. Model ini berhasil mengidentifikasi 84 anggota gugus (*True Positives*) dan 327 *field stars* (*True Negatives*) dengan laporan klasifikasi sebagai berikut.

Classification Report:				
	precision	recall	f1-score	support
0	0.85	0.96	0.90	341
1	0.86	0.60	0.70	141
accuracy			0.85	482
macro avg			0.85	482
weighted avg			0.85	482

Gambar 4.2.3. Laporan klasifikasi model Random Forest

Pada SVM, kami menggunakan kernel RBF karena kernel *default*–linear tidak dapat memisahkan data tersebut secara linear. Berikut adalah evaluasi dari model SVM yang dibangun.



Gambar 4.2.4. *Confusion matrix* model SVM

Evaluasi performa melalui *confusion matrix* memberikan gambaran kuantitatif yang lebih mendetail. Model ini berhasil mengidentifikasi 127 anggota gugus (*True Positives*) dan 290 *field stars* (*True Negatives*) dengan laporan klasifikasi sebagai berikut.

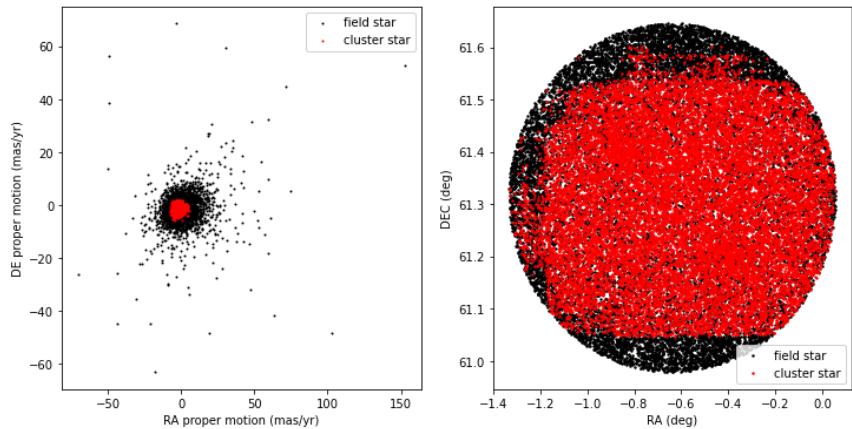
Classification Report:					
	precision	recall	f1-score	support	
0	0.95	0.85	0.90	341	
1	0.71	0.90	0.80	141	
accuracy			0.87	482	
macro avg	0.83	0.88	0.85	482	
weighted avg	0.88	0.87	0.87	482	

Gambar 4.2.5. Laporan klasifikasi model SVM

4.2.2 Penentuan Keanggotaan Gugus NGC 7790

Setelah kami membangun model dari tahap sebelumnya, model tersebut kami terapkan pada dataset NGC 7790 untuk menentukan keanggotaan gugusnya. Berikut adalah hasilnya:

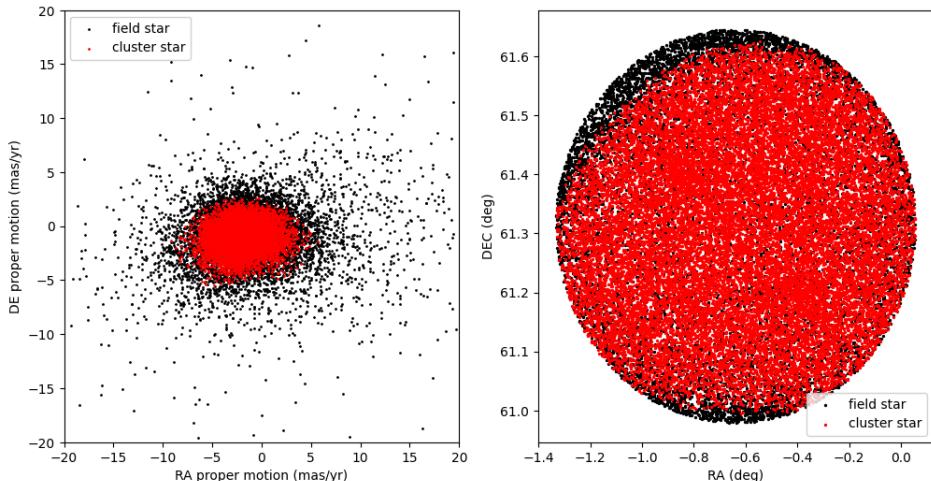
1. Random Forest



Gambar 4.2.6. Hasil penentuan keanggotaan gugus NGC 7790 dengan model Random Forest

Berdasarkan hasil tersebut, anggota gugus yang teridentifikasi (titik merah) menunjukkan konsentrasi yang sangat rapat di sekitar pusat diagram sebagai gugus bintang. Distribusi spasialnya (gambar kanan) juga menegaskan bahwa anggota gugus yang teridentifikasi terkonsentrasi di bagian tengah bidang pandang.

2. Support Vector Machine (SVM)



Gambar 4.2.7. Hasil penentuan keanggotaan gugus NGC 7790 dengan model SVM

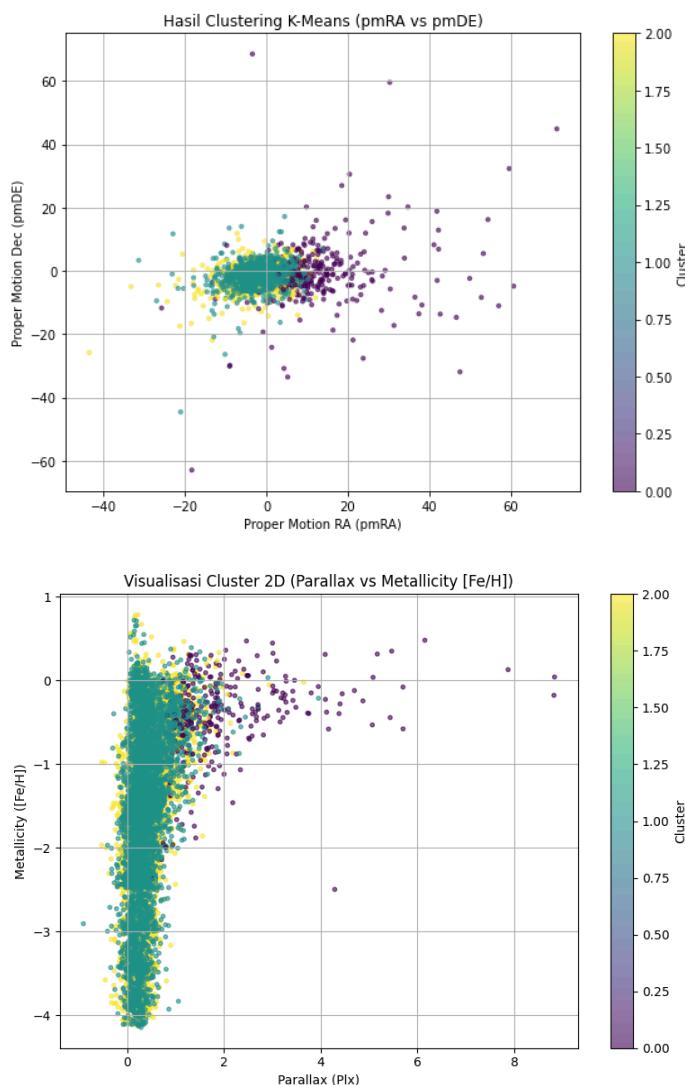
Berdasarkan hasil tersebut, anggota gugus yang teridentifikasi (titik merah) menunjukkan konsentrasi di sekitar pusat diagram sebagai gugus bintang. Distribusi spasialnya (gambar kanan) juga mengonfirmasi bahwa anggota gugus yang teridentifikasi terkonsentrasi di bagian tengah bidang pandang.

4.3 Unsupervised Learning

Pada *unsupervised learning*, kami menerapkan algoritma K-Means dan HDBSCAN pada dataset NGC 7790 secara langsung. Hal ini dikarenakan *unsupervised learning* tidak perlu data berlabel seperti supervised learning, sehingga kami dapat langsung menggunakan dataset objek yang ingin di-*clustering*. Berikut adalah hasil penentuan keanggotaan gugus NGC 7790 dengan algoritma K-Means dan HDBSCAN.

1. K-Means

Penggunaan K-Means clustering untuk menganalisis gugus bintang NGC 7790 menghasilkan temuan yang menarik. Algoritma ini tidak hanya mendeteksi gugus utama, tetapi juga mengidentifikasi gugus lain yang berdekatan, yang sesuai dengan NGC 7788 sebagaimana ditunjukkan pada perangkat lunak peta bintang Stellarium. Kami menerapkan K-Means clustering dengan tiga klaster ($k = 3$), yang merepresentasikan NGC 7790, NGC 7788, dan bintang latar belakang. Metode ini berhasil mengelompokkan bintang-bintang tersebut dan memvisualisasikannya dalam plot 2D. Hal ini menunjukkan bagaimana pembelajaran mesin dapat mengotomatisasi penentuan keanggotaan gugus dan mengungkap keberadaan beberapa gugus dalam satu wilayah langit.



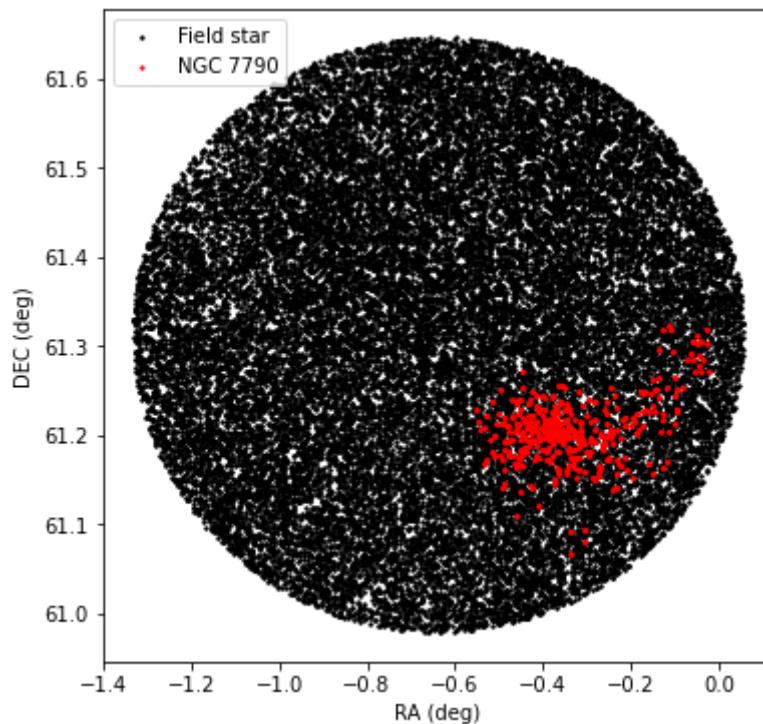
Gambar 4.2.8. Hasil penentuan keanggotaan gugus NGC 7790 dengan algoritma K-Means

Kemudian untuk mengevaluasi hasilnya, kami memanfaatkan *Silhouette Score*, *Davies-Bouldin Index*, *Calinski-Harabasz Score*. Berikut adalah laporan evaluasinya.

```
K-Means Silhouette Score: 0.249  
K-Means Davies-Bouldin Index: 1.414  
K-Means Calinski-Harabasz Score: 2579.740
```

2. HDBSCAN

Kami menerapkan algoritma HDBSCAN untuk mengidentifikasi kelebihan kerapatan bintang (*stellar overdensities*) dalam dataset. Parameter `min_cluster_size` diatur sebesar 50, yang mencerminkan batas bawah perkiraan jumlah bintang anggota yang umum diamati pada gugus terbuka. Hasil pengelompokan dapat dilihat di bawah ini.



Gambar 4.2.9. Hasil penentuan keanggotaan gugus NGC 7790 dengan algoritma HDBSCAN

Hasil tersebut menunjukkan bahwa HDBSCAN berhasil mengidentifikasi wilayah pusat gugus bintang yang padat. Lokasi kelompok yang teridentifikasi ini sangat sesuai dengan posisi NGC 7790 yang telah diketahui dalam katalog astronomi.

Kemudian untuk mengevaluasi hasilnya, kami memanfaatkan *Silhouette Score*, *Davies-Bouldin Index*, *Calinski-Harabasz Score*. Berikut adalah laporan evaluasinya.

```
HDBSCAN Silhouette Score: 0.441
HDBSCAN Davies-Bouldin Index: 0.881
HDBSCAN Calinski-Harabasz Score: 546.323
```

4.4 Analisis Hasil dari Keseluruhan Algoritma

Berdasarkan hasil implementasi dan evaluasi keempat algoritma machine learning yang digunakan, yaitu Random Forest, Support Vector Machine (SVM), K-Means, dan HDBSCAN, dapat dilakukan analisis komparatif terhadap kemampuan masing-masing metode dalam menentukan keanggotaan gugus bintang NGC 7790. Analisis ini mencakup aspek akurasi klasifikasi, kemampuan mengenali struktur gugus, ketahanan terhadap bintang latar (*field stars*), serta kesesuaian metode terhadap karakteristik data astronomi.

Pada pendekatan *supervised learning*, Random Forest dan SVM menunjukkan kemampuan yang baik dalam mengidentifikasi anggota gugus ketika diterapkan pada dataset NGC 7790, meskipun model dilatih menggunakan dataset lain (NGC 1624). Random Forest menghasilkan distribusi anggota gugus yang sangat terpusat di bagian tengah diagram dan bidang pandang, menunjukkan bahwa model ini cukup stabil dan mampu menangkap pola umum keanggotaan gugus berdasarkan fitur astrometri. Keunggulan Random Forest terletak pada sifat ensemble-nya yang relatif tahan terhadap *noise* dan *overfitting*, sehingga hasil klasifikasinya terlihat konsisten secara spasial.

Sementara itu, SVM juga berhasil mengidentifikasi konsentrasi anggota gugus di sekitar pusat, meskipun distribusinya sedikit lebih menyebar dibandingkan Random Forest. Hal ini mengindikasikan bahwa SVM lebih sensitif terhadap batas keputusan, terutama ketika data target memiliki perbedaan distribusi dengan data latih. Meskipun demikian, hasil SVM tetap menunjukkan bahwa pendekatan *supervised learning* dapat digunakan untuk penentuan keanggotaan gugus, dengan catatan ketersediaan data latih yang representatif sangat berpengaruh terhadap performa model.

Pada pendekatan *unsupervised learning*, K-Means dan HDBSCAN menunjukkan karakteristik yang berbeda namun saling melengkapi. K-Means dengan jumlah klaster $k = 3$ tidak hanya berhasil mengidentifikasi gugus utama NGC 7790, tetapi juga mendeteksi gugus lain yang berdekatan, yaitu NGC 7788, serta memisahkannya dari bintang latar. Hasil ini menegaskan bahwa K-Means efektif dalam mengelompokkan struktur besar dan terpisah secara spasial. Namun, keterbatasan K-Means terletak pada keharusan menentukan jumlah klaster di awal serta asumsi bentuk klaster yang cenderung sferis, yang tidak selalu sesuai dengan distribusi alami data astronomi.

HDBSCAN menunjukkan performa yang paling unggul dalam konteks data NGC 7790. Algoritma ini berhasil mengidentifikasi wilayah pusat gugus yang padat tanpa perlu menentukan jumlah klaster terlebih dahulu, sekaligus mengklasifikasikan bintang-bintang lain sebagai *noise* atau *field stars*. Kemampuan HDBSCAN dalam mendeteksi klaster berbasis kepadatan menjadikannya sangat sesuai untuk data astronomi yang kompleks dan

terkontaminasi oleh bintang latar. Lokasi klaster yang teridentifikasi juga sangat konsisten dengan posisi NGC 7790 yang tercatat dalam katalog astronomi, menunjukkan validitas hasil clustering yang tinggi.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa *unsupervised learning*, khususnya HDBSCAN, lebih unggul dalam menentukan keanggotaan gugus bintang tanpa ketergantungan pada data berlabel. Sementara itu, metode *supervised learning* tetap memberikan hasil yang baik, tetapi performanya sangat bergantung pada kesesuaian data latih dengan objek yang dianalisis. Oleh karena itu, untuk kasus penentuan keanggotaan gugus bintang dengan data yang kompleks dan minim label seperti NGC 7790, HDBSCAN merupakan metode yang paling optimal dibandingkan algoritma lainnya.

BAB V

SIMPULAN DAN SARAN

5.1 SIMPULAN

Berikut adalah beberapa simpulan dari penelitian ini:

1. Sebuah program berbasis *machine learning* yang mampu melakukan klusterisasi gugus bintang secara otomatis telah berhasil dikembangkan. Dokumentasi program dapat diakses pada tautan [RBL Alpro.ipynb](#)
2. Hasil penerapan empat metode *machine learning*, yaitu Random Forest, SVM, K-Means, dan HDBSCAN pada dataset untuk *clustering* dapat dilihat pada Bab IV.
3. Perbandingan performa masing-masing metode berdasarkan metrik evaluasi dapat dilihat pada Bab IV.
4. Metode yang paling optimal dalam mengidentifikasi keanggotaan gugus bintang seperti NGC 7790 yang memiliki data latih terbatas adalah HDBSCAN.

5.1 SARAN

Berikut adalah beberapa saran untuk penelitian ini:

1. Gunakan semakin banyak data latih agar program bisa semakin akurat dalam menentukan keanggotaan bintang dari sebuah gugus.
2. Eksplor semakin banyak metode *machine learning* untuk *clustering* dan bandingkan performanya, agar bisa menemukan metode paling efisien.
3. Jika memungkinkan, lakukan kolaborasi metode untuk menghasilkan akurasi yang semakin tinggi.
4. Sebaiknya dalam menentukan keanggotaan gugus bintang memakai unsupervised learning atau jika ingin tetap menggunakan supervised learning gunakan data training dengan objek yang sama.

DAFTAR PUSTAKA

Britannica Editors. (n.d.). Stellar association. In Encyclopaedia Britannica. Retrieved November 11, 2025, from <https://www.britannica.com/science/stellar-association>

Carroll, B. W., & Ostlie, D. A. (2017). An Introduction to Modern Astrophysics (2nd ed.). Pearson Education Limited.

Castro-Ginard, A., Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder, *Astronomy and Astrophysics*, vol. 661, Art. no. A118, EDP, 2022. doi:10.1051/0004-6361/202142568

European Space Agency/ESA-Hubble Outreach Team. (n.d.). Open cluster. Retrieved November 11, 2025, from <https://esahubble.org/wordbank/open-cluster/>

Fadiyah, S. Z. (2021). *Deteksi Aliran Bintang di Halo Galaksi dengan Metode Unsupervised Learning*. Tesis Magister, Institut Teknologi Bandung.

Gupta, A. C., Subramaniam, A., Sagar, R., and Griffiths, W. K., “A complete photometric study of the open cluster NGC 7790 containing Cepheid variables”, *Astronomy and Astrophysics Supplement Series*, vol. 145, EDP, pp. 365–375, 2000. doi:10.1051/aas:2000247.

Lynga G., 1987, Catalogue of open cluster data, 5th edition, 1/1 S7041, Centre de Donnees Stellaires, Strasbourg

Nasution, A. U., Ambarita, E. P., Nurhidayanti, Fadlia, H. E., & Victor Asido Elyakim P. (2025). Penerapan Algoritma Gaussian Mixture Models (GMM) untuk Identifikasi Pola Kesalahan pada Kompilasi. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, 4(2), 88–98.

National Aeronautics and Space Administration. (2025, March 27). Hubble’s Messier Catalog. Retrieved November 11, 2025, from <https://science.nasa.gov/mission/hubble/science/explore-the-night-sky/hubble-messier-catalog/>

Sandage, A., Cepheids in Galactic Clusters. I. CF Cass in NGC 7790., *The Astrophysical Journal*, vol. 128, IOP, p. 150, 1958. doi:10.1086/146532.

Syawly, A. M. (2025). *Klasterisasi Gugus Bintang Ganda dari Kondisi Awal Distribusi Fraktal Menggunakan K-Means dan HDBSCAN*. Tesis Magister, Institut Teknologi Bandung.

LAMPIRAN

Dokumentasi kode program pada penelitian ini dapat diakses melalui tautan berikut

 [RBL Alpro.ipynb](#)