BUKU TUGAS AKHIR:
KLASIFIKASI SUBKELAS KATAI PUTIH DENGAN METODE RANDOM FOREST

TUGAS AKHIR

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Sarjana dari
Institut Teknologi Bandung

Oleh
ANNEKE DIAN ISLAMIATI
NIM 10319037

PROGRAM STUDI SARJANA ASTRONOMI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI BANDUNG

2023

PENGESAHAN

Buku Tugas Akhir: *Klasifikasi Subkelas Katai Putih dengan Metode Random Forest*

Oleh
Anneke Dian Islamiati
NIM 10319037

Program Studi Sarjana Astronomi
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Teknologi Bandung

Bandung, tt bbbb TTTT
Menyetujui Dosen Pembimbing,

Dr. rer. nat. Mochamad Ikbal Arifyanto
NIP 197405192006041015

Tim Penguji:
1. **Dr. Aprilia, S.Si., M.Si.**
2. **Lucky Puspitarini, S.Si., M.Sc.,**

PEDOMAN PENGGUNAAN BUKU TUGAS AKHIR

Buku Tugas Akhir Sarjana ini tidak dipublikasikan, namun terdaftar dan tersedia di Perpustakaan Institut Teknologi Bandung. Buku ini dapat diakses umum, dengan ketentuan bahwa penulis memiliki hak cipta dengan mengikuti aturan HaKI yang berlaku di Institut Teknologi Bandung. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau peringkasan hanya dapat dilakukan seizin penulis, dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Memperbanyak atau menerbitkan sebagian atau seluruh buku Tugas Akhir harus atas izin Program Studi Sarjana Astronomi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung.

…

On the darkest nights of our souls, the Qur'an is a faithful companion with embracing arms. For every feeling we are experiencing, the Qur'an has a shooting verse, and for every pain we carry it has a timeless remedy.

…

If you are tired of pain, stop getting attached to things that pass away.

…

Aru Barzak

KATA PENGANTAR

Penulis dengan tulus bersyukur dan mengucapkan puji kepada Allah SWT, karena tanpa limpahan rahmat dan kasih sayang-Nya, Tugas Akhir dengan judul "Klasifikasi Subkelas Katai Putih dengan Metode Random Forest" tidak akan dapat diselesaikan.

Alam semesta selalu memberikan kejutan dalam setiap penemuan fakta di dalamnya. Sebagai ilmu yang mempelajari alam semesta, Astronomi sering menemukan pola keteraturan alam semesta yang unik. Setiap detail keteraturan dan pola yang diberikan selalu berhasil membuat kagum kepada Tuhan atas segala bentuk ciptaan-Nya. Oleh karena itu, penulis mencoba menggambarkan kekaguman tersebut dalam pengerjaan Tugas Akhir ini.

Tugas akhir ini akan sedikit membahas mengenai salah satu objek dalam alam semesta yang telah mengalami evolusi. Tidak hanya manusia dan makhluk hidup lain di Bumi yang mengalami evolusi, tetapi bintang juga mengalami evolusi dalam kehidupannya. Jika manusia diberikan berbagai kelebihan dan kelemahan pada saat kelahirannya oleh Tuhan, begitu pula bintang pada saat kelahirannya diberikan massa awal yang berbeda sebagai bekal untuk perjalanannya dalam hidup. Setelah mengalami evolusi, bintang biasanya mengakhiri hidupnya dan berubah menjadi berbagai bentuk objek tergantung pada massa awal dan proses yang dialaminya. Salah satu bentuk bintang yang telah mengalami evolusi adalah katai putih (*white dwarf*).

Katai putih sendiri merupakan bentuk evolusi dari bintang dengan massa menengah. Nama "katai" menggambarkan ukuran kecil atau mendekati ukuran planet. "Putih" pada istilah ini bukan mengacu pada warna, melainkan temperatur yang sangat tinggi. Karena katai putih tersebar di berbagai tempat, diperlukan suatu proses untuk memisahkannya dari bintang-bintang lain.

Dikarenakan jumlah data yang sangat besar, diperlukan suatu algoritma yang mampu mengklasifikasikan bintang katai putih secara otomatis. Oleh sebab itu, teknik pembelajaran mesin (*machine learning*) dianggap memiliki kemampuan yang baik dalam menghadapi jumlah data yang besar dengan tingkat performa yang memadai. Salah satu metode

klasifikasi dalam teknik pembelajaran mesin yang digunakan dalam klasifikasi ini adalah Random Forest.

Telah selesainya pengerjaan Tugas Akhir ini merupakan wujud dari kasih sayang Tuhan yang diberikan melalui keluarga dan lingkungan yang senantiasa memberikan dukungan dan bantuan kepada saya. Saya mengucapkan terima kasih yang sebesar-besarnya kepada keluarga saya yang memberikan dukungan moral selama pengerjaan Tugas Akhir ini. Saya juga ingin menyampaikan rasa terima kasih yang mendalam kepada:

1. Kedua orang tua, adik terkasih dan keluarga lainnya yang selalu memberikan dukungan moral dan dedikasinya terhadap saya dalam menyelesaikan perkuliahan di Institut Teknologi Bandung.
2. Dr. rer. Nat. Mochamad Ikbal Arifyanto sebagai dosen pembimbing yang telah meluangkan waktu dan memberikan ilmu, semangat, kesempatan, dan pengetahuan kepada saya selama pengerjaan Tugas Akhir ini.
3. Dr. Lucky Puspitarini dan Dr. Aprilia yang bersedia menjadi Tim Penguji dan memberikan masukan terkait Tugas Akhir ini.
4. Dr. Chatief Kunjaya, M.Sc. selaku dosen wali saya selama berkuliah di Program Studi Sarjana Astronomi.
5. Teman-teman Astronomi ITB 2019, terutama Fathia, Akniz, Khariza, Luthfi, Aulia, Tasya, Nynda, Raihan, Dida, Hafiz, Hiskia dan Gishel atas bantuan dan dukungan yang kalian berikan selama perkuliahan di Astronomi ITB.
6. Staf dosen Program Studi Astronomi yang telah berbagi ilmu selama masa perkuliahan di Program Studi Sarjana Astronomi.
7. Staf Tata Usaha Program Studi Astronomi yang telah membantu saya dalam berbagai hal selama perkuliahan.

Terima kasih yang sebesar-besarnya atas semua bantuan dan dukungan yang telah diberikan

Bandung, 27 Juni 2023

Anneke Dian Islamiati
NIM 10319037

ABSTRAK

Bintang katai putih merupakan fase terakhir dalam evolusi sebagian besar bintang. Diperkirakan sekitar 97% dari seluruh bintang akan mengakhiri siklus hidupnya secara pasif dengan melepaskan lapisan luar dan berubah menjadi katai putih. Studi mengenai katai putih akan memberikan informasi mengenai evolusi bintang dari awal hingga akhir. selain itu, dengan mempelajari katai putih maka akan diperoleh informasi mengenai evolusi kimia di galaksi kita. Tidak hanya itu, proses evolusi katai putih juga akan memberikan kita pengetahuan mengenai sifat materi dalam keadaan terdegenerasi.

Pada pengerjaan Tugas Akhir ini, akan dilakukan klasifikasi untuk menentukan subkelas katai putih dan regresi parameter fisis dari katai putih berupa temperatur efektif dan gravitasi permukaan. Metode yang digunakan yaitu Random Forest. Random Forest merupakan metode machine learning yang dibangun dari beberapa *decision tree* untuk melakukan tugas klasifikasi dan regresi. Untuk mendapatkan nilai parameter fisis serta klasifikasi subkelas katai putih, pengerjaan Tugas Akhir ini akan memanfaatkan spektrum dari katai putih dengan panjang gelombang sebagai fiturnya. Spektrum katai putih yang digunakan pada Tugas Akhir ini diperoleh dari data Gaia DR3 dan LAMOST DR8. Dalam klasifikasi, pengerjaan Tugas Akhir ini akan mengklasifikasikan dua subkelas katai putih, yaitu subkelas DA dan DAZ. Hasil dari klasifikasi ini kemudian akan dibandingkan dengan paper Echeverry, D., dkk. (2022). yang memisahkan objek antara katai putih, bintang deret utama kelas M dan bintang ganda deret utama-katai putih. Sedangkan hasil dari regresi akan dibandingkan dengan error yang diperoleh dari intrumen pengamtan dan tersedia pada database. Didapatkan akurasi untuk proses klasifikasi mencapai nilai 90% sedangkan error pada parameter fisis lebih baik dibandingkan dengan error pada instrumennya.

Kata kunci: Katai putih, Spektrum, Random Forest.

ABSTRACT

White dwarf stars are the final stage in the evolution of most stars. It is estimated that approximately 97% of all stars will passively end their life cycles by shedding their outer layers and transforming into white dwarfs. Studies on white dwarfs provide information about the evolution of stars from beginning to end. Furthermore, by studying white dwarfs, we can gather information about the chemical evolution of our galaxy. Not only that, but the process of white dwarf evolution also offers insights into the properties of degenerate matter.

In this Final Project, classification will be conducted to determine the subclasses of white dwarfs and the regression of their physical parameters, namely effective temperature and surface gravity. The method used is Random Forest, which is a machine learning method built from multiple decision trees to perform its task. To obtain the values of the physical parameters and classify the subclasses of white dwarfs, this Final Project will utilize the spectra of white dwarfs with wavelength as the feature. The spectra of white dwarfs used in this Final Project were obtained from Gaia DR3 and LAMOST DR8 data. In the classification phase, this Final Project will classify two subclasses of white dwarfs, namely the DA and DAZ subclasses. The results of this classification will then be compared with the paper Echeverry, D., dkk. (2022), which distinguishes objects between white dwarfs, main sequence stars of class M, and main sequence-white dwarf binaries. Meanwhile, the results of the regression will be compared with the error values obtained from the observation instrument and available in the database. The accuracy for the classification process reaches 90%, while the error values for the physical parameters are better compared to the errors in the instrument.

Key words: Whitedwrafs, Spectra, Random Forest.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

I.1 Latar Belakang

Katai putih merupakan tahap akhir evolusi bintang yang merupakan inti dari bintang bermassa rendah hingga menengah yang mengalami pembakaran hidrogen. Bintang katai putih tidak memiliki sumber energi nuklir, sehingga seiring berjalannya waktu, bintang ini akan mendingin dan hanya akan berkontribusi sebanyak ≤ 10% terhadap materi gelap dalam Galaksi. Sekitar 97% dari total bintang akan mengakhiri hidupnya sebagai katai putih dengan menghilangkan lapisan luar mereka, sehingga katai putih ini menyimpan informasi mengenai evolusi bintang individu dari awal hingga akhir. Pelepasan lapisan luar ini juga berpengaruh pada evolusi kimia galaksi. Dalam survei yang semakin berkembang, banyak katai putih ditemukan dengan parameter fisis di dalamnya yang berada dalam kondisi terdegenerasi. Oleh karena itu, studi mengenai katai putih juga memberikan informasi tentang sifat materi pada kerapatan dan densitas tinggi.

Pada tahun 2022, David Echeverry, Santiago Torres, Alberto Rebassa-Mansergas, dan Aina Ferrer-Burjachs melakukan penelitian mengenai katai putih. Dalam penelitian tersebut, dilakukan pengelompokan spektrum antara katai putih, bintang deret utama, dan bintang ganda katai putih-deret utama. Metode yang digunakan adalah Random Forest. Random Forest dipilih karena terbukti sebagai teknik yang paling menjanjikan dalam kinerja dan keandalannya dalam mengelola data dalam jumlah besar. Berkembangnya basis data astronomi yang besar telah melampaui kemampuan manusia untuk melakukan analisis secara langsung. Sejak adanya misi otomatis seperti Hipparcos, SDSS, LAMOST, dan lainnya, informasi yang tersedia menjadi sangat melimpah dan jumlahnya belum pernah terjadi sebelumnya. Sebagai akibatnya, basis data astronomi terus tumbuh dengan cepat baik dari segi volume maupun kompleksitasnya. Fenomena ini biasa disebut sebagai astronomi berbasis data (*data-driven astronomy*). Untuk menghadapi tantangan ini, diperlukan penggunaan teknik kecerdasan buatan, terutama dalam hal data mining dan metode pembelajaran mesin, guna mengelola data dalam jumlah yang besar. Pada Tugas Akhir ini, akan dilakukan pengelompokan spektrum subkelas katai

putih dengan menggunakan Random Forest.

I.2 Rumusan dan Batasan Masalah

Beberapa perumusan masalah yang dibahas dalam Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana proses dan hasil algoritma Random Forest dalam melakukan klasifikasi subkelas katai putih dan regresi parameter fisis berdasarkan spektrum yang diamati?
2. Apakah hasil klasifikasi subkelas katai putih data LAMOST DR8 dan Gaia DR3 melalui algoritma Random Forest efektif dan akurat?
3. Apakah hasil regresi parameter fisis data LAMOST DR8 dan Gaia DR3 melalui algoritma Random Forest efektif dan akurat?

I.3 Tujuan

Berdasarkan rumusan masalah yang telah dirumuskan, tujuan dari Tugas Akhir ini adalah sebagai berikut:

1. Mempelajari algoritma Random Forest sebagai metode untuk menentukan subkelas dan parameter fisis katai putih.
2. Menentukan akurasi klasifikasi Random Forest yang diterapkan pada data LAMOST DR8 dan Gaia DR3.
3. Menentukan akurasi regresi Random Forest yang diterapkan pada data LAMOST DR8 dan Gaia DR3.

I.4 Metodologi

Pada Tugas Akhir ini, dilakukan beberapa metode, antara lain:

1. Studi Pustaka

   Studi pustaka dilakukan dengan mempelajari berbagai publikasi mengenai Katai putih, termasuk klasifikasi dan proses pembentukannya. Selain itu, dilakukan studi pustaka untuk memahami *decision tree* dan Random Forest melalui artikel ilmiah dan paper yang relevan.
2. Pengumpulan Data

   Selanjutnya, data dikumpulkan dari Gaia Data Release 3 dengan label subkelas katai putih. Spektrum katai putih dari Gaia Data Release 3 diperoleh melalui layanan Gaia@AIP Services [1].

3. Pengolahan Data
   Pengolahan data dilakukan dengan menggunakan algoritma Random Forest.

I.5 Sistematika Penulisan

Tugas Akhir ini terdiri dari lima bab yang meliputi Pendahuluan, Konsep Dasar dan Studi Pustaka, Machine Learning, Data dan Perangkat Penelitian, serta Hasil dan Analisis, dan Simpulan.

Pada Bab I, Pendahuluan, telah dijelaskan secara rinci mengenai Tugas Akhir ini. Bab II akan membahas topik-topik seperti Katai Putih, Spektroskopi, dan Machine Learning. Selanjutnya, Bab III akan mengulas tentang sumber data dan memberikan gambaran umum mengenai pengolahan data. Bab IV akan fokus pada hasil dan analisis dari klasifikasi subkelas Katai putih. Terakhir, Bab V akan menyajikan simpulan dan saran yang diambil dari Tugas Akhir ini.

BAB II

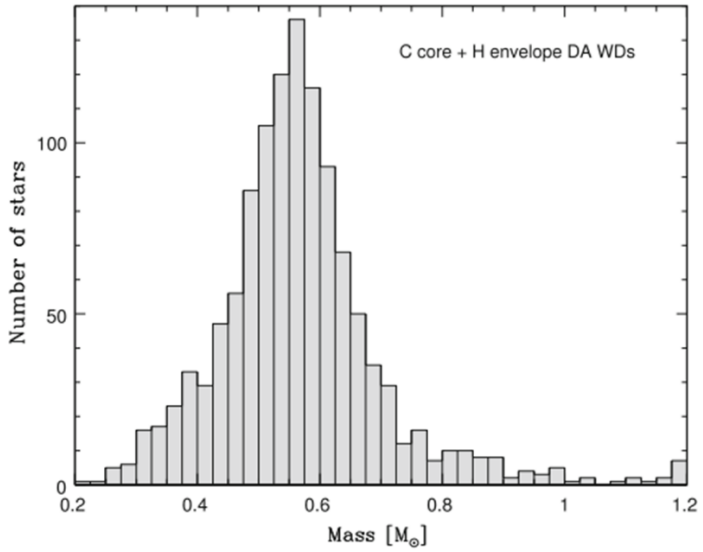
KONSEP DASAR DAN STUDI PUSTAKA

II.1 Katai Putih

Bintang katai putih merupakan tahap akhir dalam evolusi sebagian besar bintang. Diperkirakan sekitar 97% dari seluruh bintang deret utama akan mengakhiri siklus hidupnya secara pasif dengan melepaskan lapisan luar dan berubah menjadi katai putih. Oleh karena itu, populasi katai putih saat ini mengandung informasi berharga mengenai evolusi individual bintang dari awal hingga akhir, serta menyimpan informasi tentang laju pembentukan bintang sepanjang sejarah Galaksi Bimasakti. Sisa-sisa bintang ini merupakan inti dari bintang dengan massa rendah hingga menengah yang telah menghabiskan kandungan hidrogen didalamnya dan tidak memiliki sumber energi nuklir. Seiring berjalannya waktu, katai putih akan mengalami peredupan dan mendingin seiring dengan pelepasan energi panas yang tersimpan di dalamnya (Althaus, L. G., dkk, 2010). Parameter fisis dari katai putih dapat dilihat pada Tabel II.1.

Tabel II. 1. Parameter fisis katai putih

Massa	$1M_{\odot} - 1.3M_{\odot}$
Gravitasi permukaan (log g)	~ 8
Temperatur efektif (T_{eff})	$4000 - 150000\ K$
Massa bintang awal	$1 - 10M_{\odot}$
Luminositas	$10^{3} - 10^{-5}L_{\odot}$

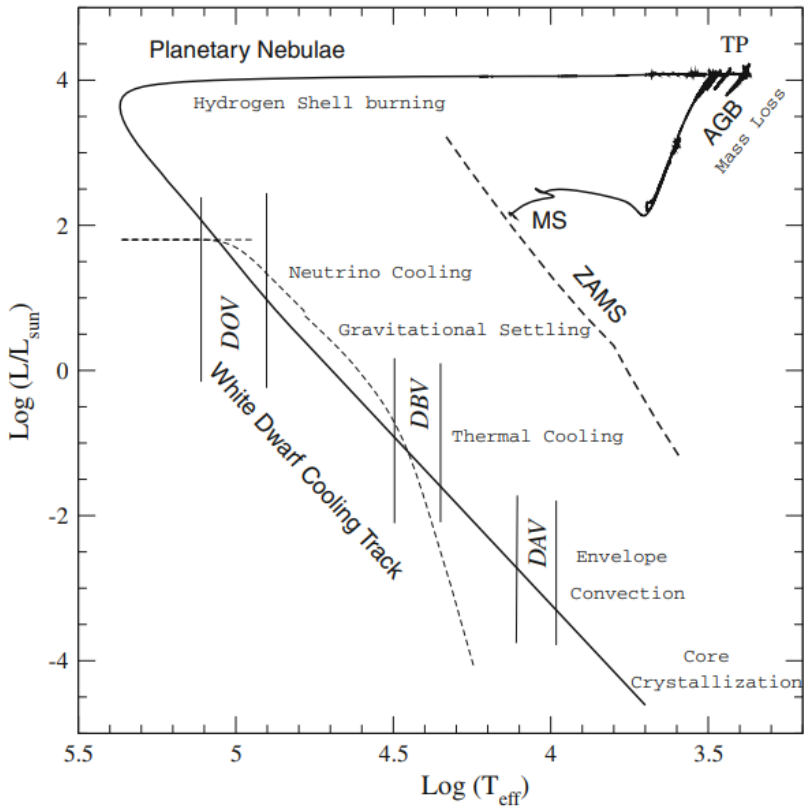


Gambar II. 1. Distribusi masa katai putih (Althaus, L. G., dkk, 2010).

II.1.1 Pembentukan Katai Putih

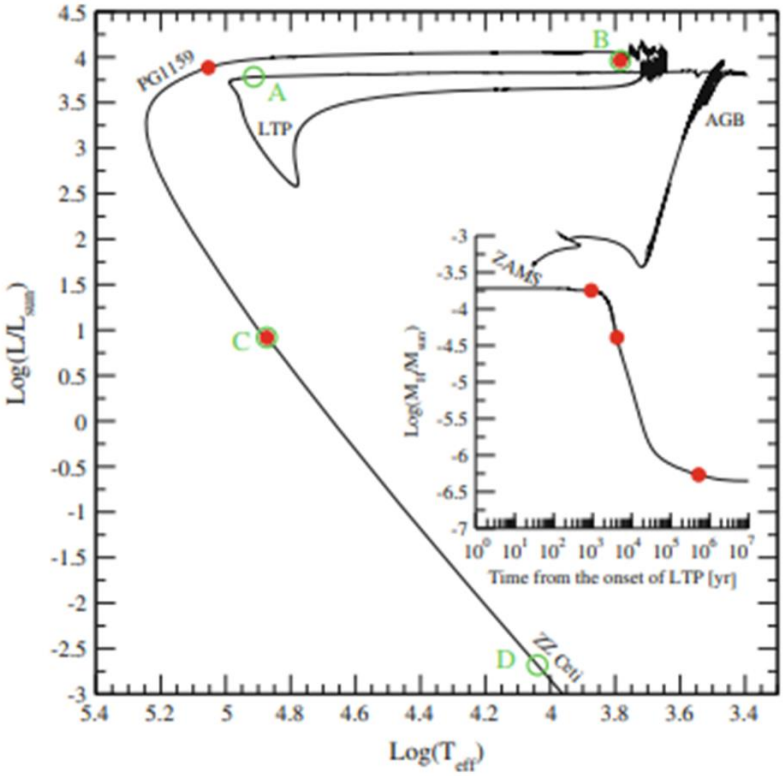
Fase utama dalam kehidupan progenitor katai putih yang khas dapat divisualisasikan dalam diagram Hertzsprung-Russell Gambar II.2. Setelah melewati tahap deret utama di mana hidrogen terbakar di intinya, bintang progenitor katai putih mengalami evolusi menjadi raksasa merah untuk membakar helium di dalam intinya. Pada tahap ini, komposisi inti yang terdiri dari karbon-oksigen akan mencirikan sisa-sisa katai putih yang terbentuk. Setelah helium di inti terbakar habis, evolusi berlanjut menjadi tahap AGB (Asymptotic Giant Branch). Pada tahap ini, selubung yang membakar helium menjadi tidak stabil, dan bintang mengalami periode ketidakstabilan termal berulang yang disebut "thermal pulse". Selama tahap AGB, massa inti karbon juga meningkat secara signifikan, dan selubung pembakaran helium bergerak ke luar. Pada tahap ini, sebagian besar selubung yang kaya akan hidrogen dilepaskan melalui proses kehilangan massa yang sangat kuat. Ketika fraksi massa selubung yang tersisa berkurang menjadi $10^{-3}\ M_{\odot}$, bintang yang tersisa bergerak ke arah kiri dalam diagram Hertzsprung-Russell menuju daerah nebula planet. Jika keluarnya dari tahap AGB terjadi pada tahap lanjut dari siklus *flash shell* helium, bintang pasca-AGB akan kehilangan seluruh hidrogennya dan menjadi katai putih dengan kelimpahan hidrogen di permukaan yang tipis. Ketika sisa-sisa selubung hidrogen berkurang menjadi $10^{-4}\ M_{\odot}$, pembangkitan energi nuklir di inti hampir sepenuhnya hilang. Akibatnya, luminositas permukaan bintang akan secara cepat berkurang, dan bintang memasuki fase akhir kehidupannya sebagai katai putih (Althaus, L. G., dkk., 2010). Skema proses pembentukan katai putih ini dijelaskan dalam diagram HR Gaia yang ditunjukkan

pada Gambar II.2.



Gambar II. 2. Diagram HR untuk evolusi binrang 3.5 $M_{\odot}$ *dari daerah ZAMS hingga katai putih (Althaus, L. G., dkk, 2010).*

Secara umum, katai putih yang terbentuk akan memiliki kandungan hidrogen di lapisan luar yang melimpah. Namun terdapat beberapa katai putih yang memiliki lapisan luar dengan kandungan hidrogen yang tipis. Mekanisme yang umum diterima untuk pembentukan sebagian besar katai putih dengan lapisan hidrogen di permukaan yang tipis adalah adanya proses VLTP (*Very Late Thermal Pulse*) selama tahap awal evolusi katai putih ketika pembakaran hidrogen hampir berhenti (Althaus dkk., 2005a). Selama VLTP, zona konveksi yang berkembang ke luar terbentuk karena pembakaran helium dan mencapai selubung yang mengandung banyak hidrogen. Hal ini mengakibatkan sebagian besar kandungan hidrogen terbakar dalam zona konveksi yang dipicu oleh *helium-flash* (Miller Bertolami dkk., 2006). Selain VLTP, terdapat juga proses LTP (*Late Thermal Pulse*) yang dapat menyebabkan pembentukan katai putih dengan sedikit kandungan hidrogen dalam selubungnya. LTP terjadi sebelum bintang yang tersisa mencapai tahap katai putih, dan selubung pembakaran hidrogen masih aktif sehingga kandungan hidrogen tidak habis terbakar (Althaus, L. G., dkk., 2005b).

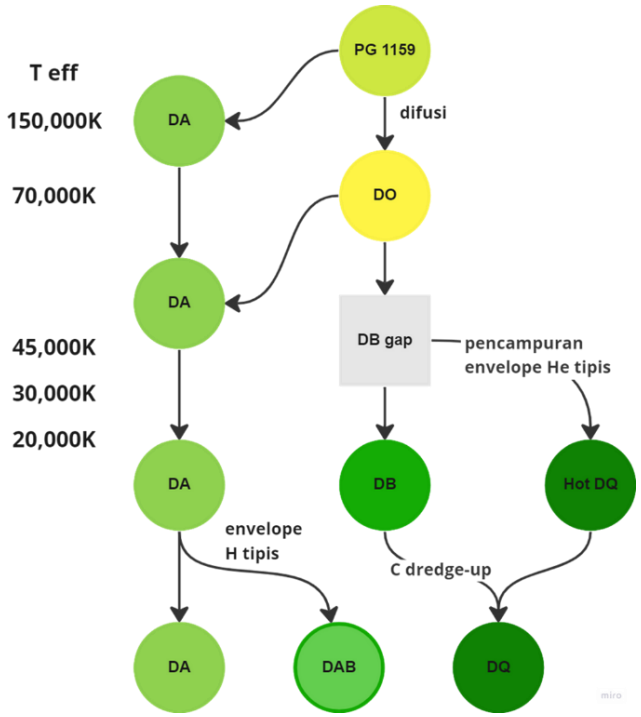


Gambar II. 3. Evolusi bintang 2.7 $M_{\odot}$ *menjadi katai putih dengan envelope hidrogen tipis. Inset pada gambar diatas menunjukkan evolusi pembakaran hidrogen pada bintang (Althaus, L. G., dkk, 2010).*

Contoh skema alur pembentukan katai putih dengan proses LTP dapat dilihat pada Gambar II.3. Setelah bintang melewati tahap ZAMS dan AGB, bintang akan memasuki daerah nebula planet. Di daerah ini, terjadi proses LTP pada bintang yang ditunjukkan oleh titik A. Tak lama setelah terjadinya LTP, sisa bintang akan kembali ke wilayah bintang raksasa. Di wilayah tersebut, hidrogen akan diencerkan oleh konveksi permukaan dan tercampur dengan wilayah yang didominasi oleh helium, karbon, dan oksigen. Proses ini ditunjukkan oleh titik B. Setelah itu, evolusi bintang akan berlanjut ke wilayah nebula planet dengan suhu efektif yang tinggi untuk menjadi bintang PG 1159. Pada tahap ini, hidrogen akan terbakar kembali hingga mencapai suhu efektif maksimum yang menandakan bahwa bintang telah memasuki wilayah katai putih. Proses ini ditunjukkan oleh titik C. Kandungan hidrogen yang tersisa pada tahap ini mencapai $8 \times 10^{-7}\ M_{\odot}$ yang menandai katai putih dengan lapisan hidrogen yang tipis. Evolusi temporal kandungan hidrogen diilustrasikan dalam inset pada Gambar 8. Dari inset Gambar 8, terlihat bahwa sebagian besar kandungan hidrogen yang tersisa akan terbakar selama 100.000 tahun pada tahap PG 1159. Kemudian, hidrogen akan berdifusi keluar pada jalur pendinginan dan mengubah katai putih menjadi salah satu tipe DA dengan selubung hidrogen yang tipis, mencapai $10^{-7}\ M_{\odot}$.

II.1.2 Evolusi Katai Putih

Meskipun katai putih merupakan tahapan akhir dari sebuah siklus hidup bintang, pada tahap katai putih sendiri, bintang akan mengalami berbagai proses evolusi kelas spektral sebelum bintang tersebut mencapai akhir dari hidupnya. Evolusi kelas spektral katai putih dapat diamati melalui perubahan komposisi permukaan yang mungkin terjadi akibat proses seperti konveksi, transfer massa, akresi, radiative levitation, dan gravitational settling. Laju evolusi katai putih seiring dengan penurunan suhu efektif bintang ditunjukkan dalam Gambar II.4.

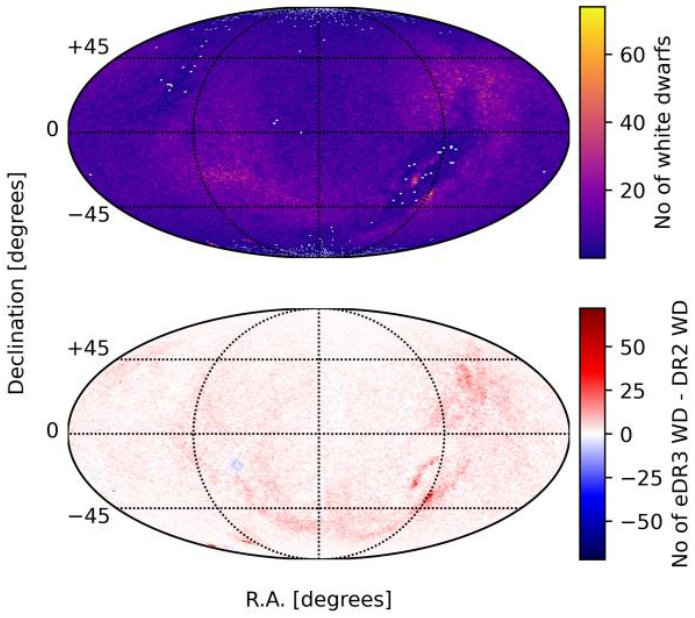


Gambar II. 4. Evolusi katai putih (Althaus, L. G., dkk, 2010).

Evolusi katai putih pada Gambar II.4. berhubungan dengan hasil pengamatan yang menunjukkan perubahan rasio antara katai putih dengan permukaan yang didominasi hidrogen dengan katai putih yang permukaannya didominasi oleh unsur lain seperti helium. Perubahan rasio tersebut terjadi seiring dengan penurunan suhu efektif dari bintang katai putih. Selain itu, dari Gambar II.4. juga dapat dilihat bahwa evolusi spektral katai putih juga dapat disebabkan oleh proses pengendapan gravitasi. Proses ini akan mengubah bintang PG 1159 yang memiliki lapisan hidrogen tipis di permukaannya menjadi katai putih dengan lapisan luar yang dominan dalam hal hidrogen.

Selain katai putih yang memiliki kelimpahan permukaan hidrogen, evolusi spektral katai putih menghasilkan berbagai objek katai putih dengan kelimpahan permukaan luar yang bervariasi. Selain kelas spektral DA (dengan lapisan luar didominasi oleh kandungan hidrogen), terdapat berbagai kelas spektral katai putih dengan kelimpahan yang telah diamati, yaitu:

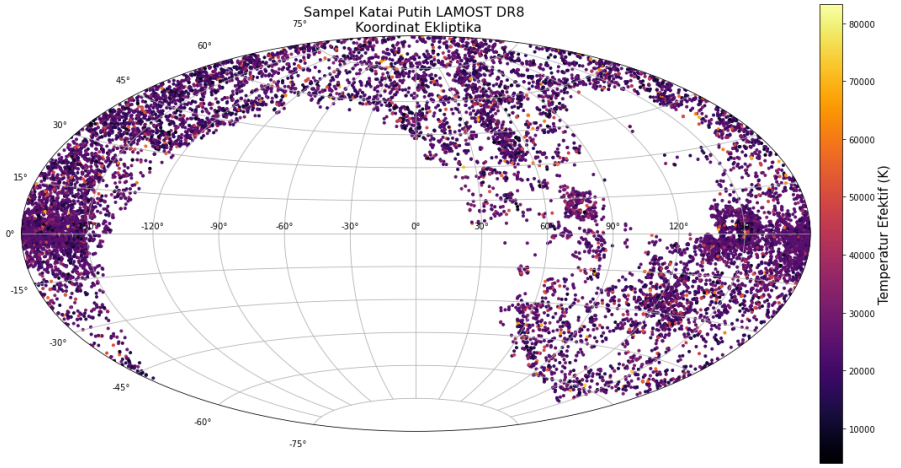
- Kelas spektrum DO (45.000-200.000 K) dengan garis yang relatif kuat dari helium terionisasi (He II).
- Kelas spektrum DB (11.000 – 30.000 K) yang menunjukkan garis helium netral (He I) yang kuat.
- Kelas spektrum DC, DQ, dan DZ dengan temperature kurang dari 11.000 K dengan spektrum yang menunjukkan jejak karbon dan logam.



Gambar II. 5. Atas : Peta sebaran katai putih Gaia eDR3. Bawah : Perbedaan sebaran katai putih Gaia DR3 dan Gaia eDR3. (Fusillo, G., dkk., 2019).

Dari pengamatan fotometri dan spektroskopi bintang katai putih, diperoleh rentang parameter fisis seperti temperatur efektif dan luminositas. Temperatur efektif katai putih yang tercatat berkisar antara 150.000 K hingga 4.000 K, sedangkan luminositasnya berkisar dari 10^3 hingga $10^{-5}\ L_{\odot}$. Karena parameter fisis ini, katai putih memiliki tingkat kecerahan yang rendah, sehingga pengamatannya terbatas pada daerah sekitar Matahari karena kemampuan deteksi yang terbatas. Namun, dengan adanya survei yang lebih luas dan teknologi yang

lebih canggih, pengamatan dapat melibatkan wilayah yang lebih dalam dan mengungkap populasi katai putih yang terletak lebih jauh, seperti dalam gugus terbuka dan gugus bola yang jauh. Contoh survei tersebut termasuk penggunaan teleskop luar angkasa Hubble, LAMOST, dan Gaia. Berikut adalah peta distribusi katai putih yang telah ditemukan dalam survei LAMOST[1] (Gambar II.6.) dan Gaia EDR3[2] (Gambar II.5.).



Gambar II. 6. Peta sebaran katai putih LAMOST DR8

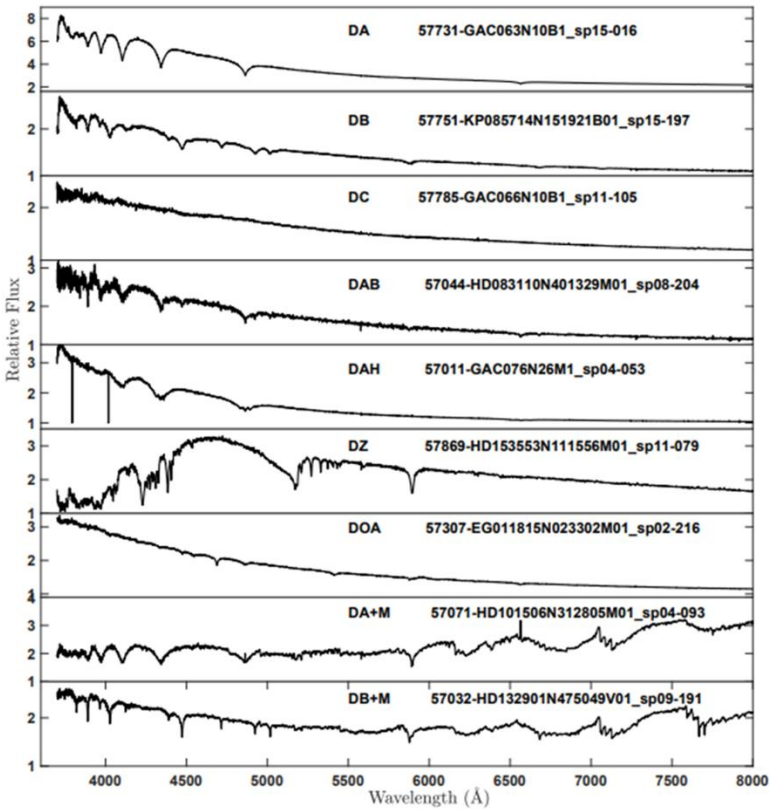
II.1.3 Subkelas Katai Putih Secara Spektroskopi

Secara tradisional, katai putih telah dikelompokkan menjadi dua jenis berdasarkan komposisi utama permukaannya. Pengamatan spektroskopi menunjukkan bahwa sebagian besar katai putih terdiri hampir seluruhnya dari hidrogen (tipe DA) dan mereka menyumbang sekitar 85% dari total katai putih. Ada juga jenis katai putih lain yang kekurangan hidrogen dengan atmosfer kaya helium (tipe non-DA), yang membentuk sekitar 15% dari populasi keseluruhan. Katai putih yang kekurangan hidrogen ini dianggap sebagai hasil dari kilatan termal akhir yang dialami oleh bintang progenitor setelah keluar dari tahap AGB atau melalui episode penggabungan. Katai putih non-DA umumnya dapat dibagi menjadi beberapa subkelas yang berbeda. Misalnya, tipe spektral DO menunjukkan garis yang kuat dari helium terionisasi tunggal (HeII) dengan rentang temperatur efektif antara 45.000 hingga 200.000 K. Tipe DB memiliki garis helium netral (HeI) yang kuat dengan rentang temperatur efektif antara 11.000 hingga 30.000 K. Sedangkan tipe DC, DQ, dan DZ menunjukkan adanya jejak karbon dan logam dalam spektrumnya dengan temperatur efektif kurang dari 11.000 K. Selain itu, terdapat juga katai putih

[1] http://www.lamost.org/dr8/v2.0/

[2] https://www.cosmos.esa.int/web/gaia/earlydr3

dengan atmosfer campuran atau kelimpahan aneh, seperti penemuan dua katai putih dengan atmosfer kaya oksigen (Gansicke, B. T., dkk., 2010), serta ditemukan jenis katai putih baru dengan atmosfer yang didominasi oleh karbon (katai putih Hot-DQ) dan temperatur efektif sekitar 20.000 K (Dufour, P., dkk., 2008a).



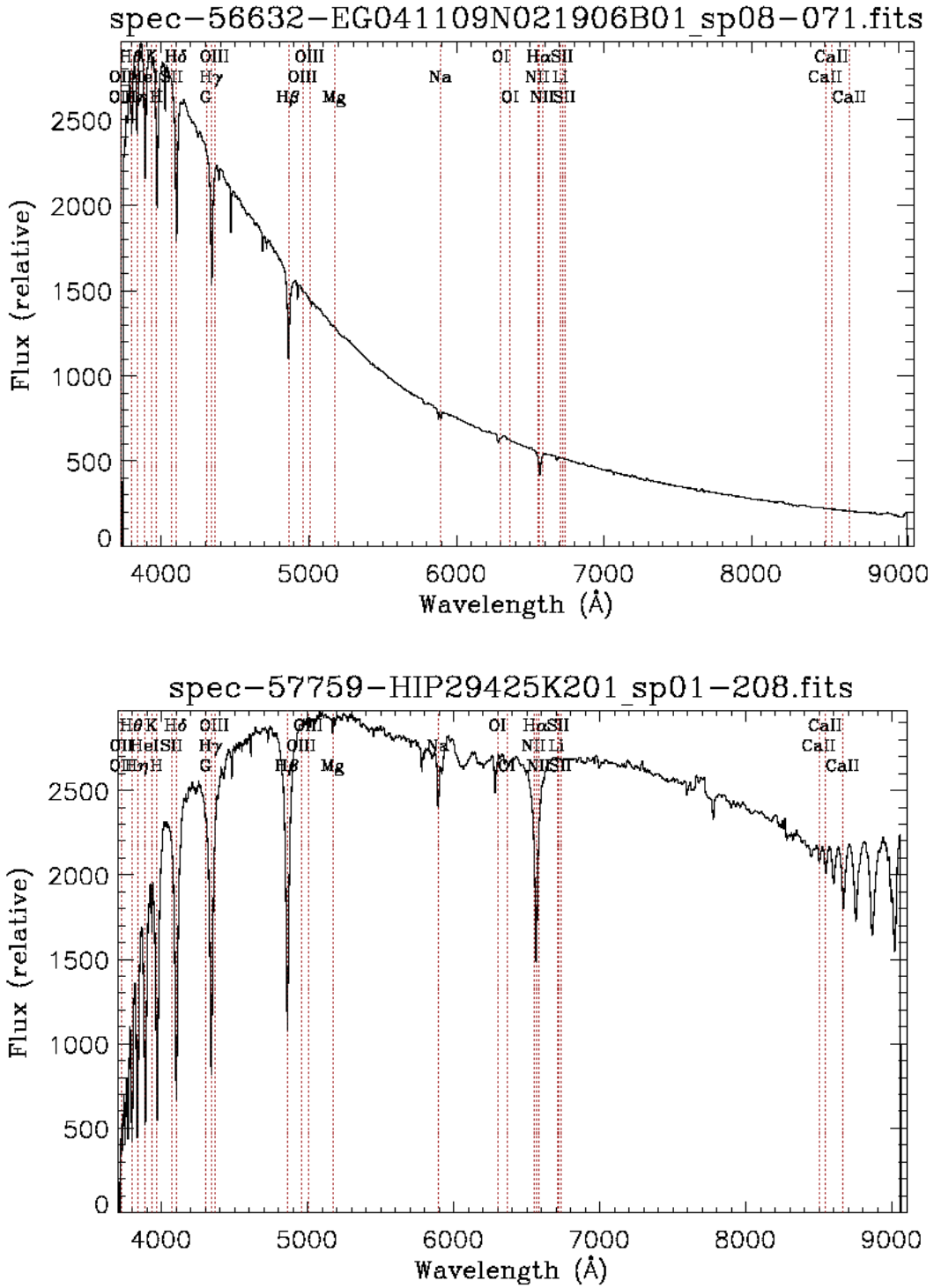
Gambar II. 7. Spektrum subkelas katai putih sampel LAMOST DR5 (Guo, J., dkk., 2022)

Perbedaan kelimpahan permukaan katai putih dapat diamati melalui spektrumnya yang terlihat dalam Gambar II.7., dengan garis-garis kelimpahan yang teramati dalam spektrum subkelas katai putih, yaitu:

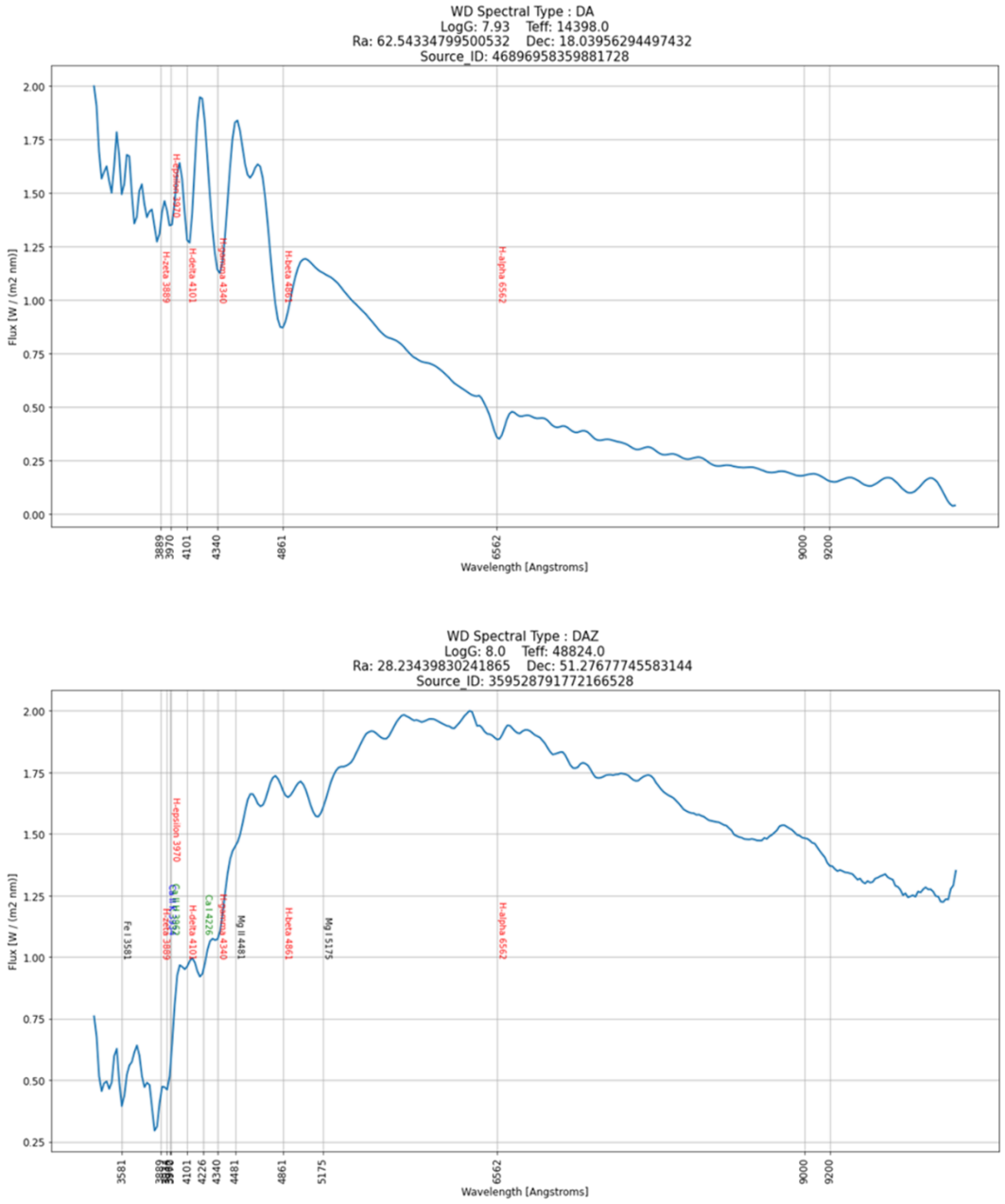
- Garis Balmer [DA, DAB, DBA, DZA, subdwarfs]
- Fitur yang lebih sedikit teramati [DC]
- OI 6158, 7774, 8448 A [DS, oxy-dominated]
- Garis CI atau C2 [DQ]
- CII 4367A [Hot DQ]
- CaII H & K [DZ, DAZ, DBZ]
- Zeeman splitting [magnetic WD]
- He II 4686 [DO, PG1159, sdO]
- He I 4471 A [DB, sdB]

II.1.4 Katai Putih dengan Kelimpahan Hidrogen

Sebagian besar katai putih yang ditemukan memiliki lapisan permukaan yang mengandung hidrogen. Katai putih dengan lapisan permukaan yang didominasi oleh hidrogen dapat dikelompokkan ke dalam dua kelas spektral yang berbeda, yaitu DA dan DAZ. Katai putih tipe DA adalah katai putih dengan kelimpahan hidrogen dan spektrum yang didominasi oleh garis Balmer. Sementara itu, katai putih tipe DAZ adalah katai putih dengan kelimpahan lapisan luar berupa hidrogen dan beberapa garis elemen berat seperti CaII H dan K. Keberadaan berbagai garis elemen berat tersebut diduga merupakan hasil dari proses akresi dan pendinginan. Proses akresi dan difusi logam pada katai putih menyebabkan elemen berat seperti kalsium dan kalium mengendap dan terkumpul di atmosfer katai putih. Selain itu, pendinginan cepat pada katai putih juga menyebabkan anomali termal yang memengaruhi penyebaran elemen berat pada katai putih. Selain proses internal, terdapat juga proses eksternal yang menyebabkan munculnya garis logam pada spektrum katai putih, seperti akresi materi asteroid (Werner, K., et al., 2009; Zuckerman, B., et al., 2003) dan interaksi dengan awan antarbintang berkerapatan sedang (Lacombe, P., et al., 1938). Garis elemen berat tersebut dapat dilihat pada dua contoh spektrum sampel katai putih dari LAMOST (Gambar II.8.) dan Gaia (Gambar II.9.).



Gambar II. 8. Spektrum katai putih DA (atas) dan DAZ (bawah) sampel LAMOST DR8. Sumber : http://www.lamost.org/dr8/v2.0/.



Gambar II. 9. Gambar II. 8. Spektrum katai putih DA (atas) dan DAZ (bawah) sampel Gaia DR3

II.2 Spektroskopi

II.2.1 Spektrum dan Spektroskopi

Kajian spektroskopi dimulai pada tahun 1621 dengan penemuan Willebrord Snel tentang hukum pembiasan Snell. Penelitian ini kemudian diteruskan oleh Sir Isaac Newton melalui eksperimen yang dijelaskan dalam karya Voltaire dan direproduksi dalam teori terkenal Condon dan Shortly mengenai Spektra Atom pada tahun 1935. Pada saat itu, Newton menjelaskan bahwa sinar matahari yang melewati prisma akan menghasilkan pita warna, dan jika sinar tersebut melewati prisma kedua dengan orientasi yang berlawanan, maka sinar tersebut akan kembali menjadi cahaya putih. Selanjutnya, pada tahun 1800, Sir William Herschel menemukan spektrum inframerah dengan menempatkan termometer di luar ujung merah spektrum tampak. Sementara itu, spektrum ultraviolet ditemukan oleh Johann Ritter pada tahun yang berikutnya.

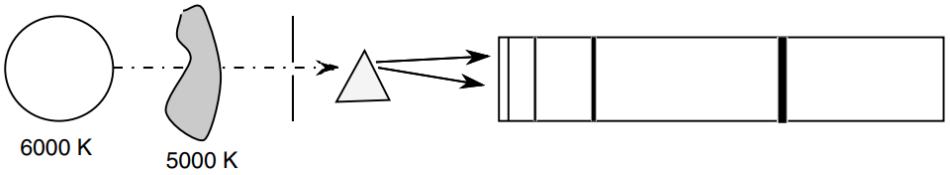
Pada tahun 1802, William Wollaston menemukan pola garis-garis gelap dalam spektrum Matahari, namun hingga saat itu penemuan ini masih belum dapat dijelaskan dengan pasti. Dari garis-garis gelap yang terlihat dalam spektrum Matahari tersebut, Joseph von Fraunhofer melakukan penelitian dan membuat peta yang mencakup 700 garis yang dikenal sebagai 'garis Fraunhofer' pada tahun 1814. Tidak hanya pada Matahari, Fraunhofer juga mengamati spektrum dari bintang pertama dan mencatat bahwa spektrum bintang tersebut mirip dengan spektrum Matahari. Selanjutnya, Fraunhofer menciptakan sebuah kisi difraksi dan menemukan garis pada panjang gelombang 588,7 nm yang kemudian diidentifikasi sebagai garis Natrium (Na I). Kajian mengenai pola garis-garis gelap tersebut kemudian dikenal dengan istilah spektroskopi.

Kirchhoff dan Bunsen melakukan pemeriksaan lanjutan terhadap spektrum beberapa unsur. Temuan mereka mampu menjelaskan asal garis-garis Fraunhofer di Matahari serta mengkaji komposisi kimia atmosfer Matahari. Hukum yang menguraikan penemuan tersebut dikenal sebagai Hukum Kirchhoff yang diumumkan pada tahun 1859.

II.2.2 Hukum Kirchoff

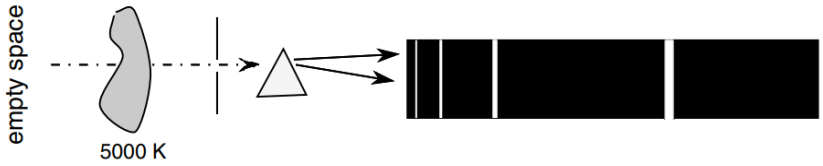
Pada tahun 1860, Kirchhoff dan Bunsen mengembangkan ide dalam karya klasik yang berjudul "Chemical Analysis by Spectral Observations", bahwa setiap unsur memiliki pola garis spektral yang khas dan berbeda dari unsur lainnya. Kirchhoff kemudian menggambarkan pembentukan garis spektral ini dalam tiga prinsip.

1. Apabila sebuah benda padat atau gas panas dipanaskan, maka cahaya akan dipancarkan dalam berbagai panjang gelombang yang menghasilkan spektrum kontinu tanpa adanya garis spektral. Spektrum kontinu ini disebabkan oleh radiasi benda hitam yang dipancarkan pada temperatur di atas nol mutlak. Fenomena ini dapat dijelaskan melalui fungsi Planck $B_\lambda(T)$ dan $B_\nu(T)$.
2. Apabila gas dipanaskan dan memiliki kepadatan rendah, maka akan menghasilkan garis-garis spektral terang yang sesuai dengan sifat-sifat kimia gas tersebut. Garis-garis tersebut dikenal sebagai garis emisi dan muncul ketika elektron dalam gas tersebut berpindah ke orbit yang lebih rendah.



Gambar II. 10. Spektrum garis emisi (Strobel, 2007).

3. Apabila gas berada dalam keadaan dingin dan renggang, ketika terpapar oleh sumber cahaya kontinu, gas tersebut akan menyerap sebagian spektrum pada panjang gelombang tertentu dan menghasilkan garis-garis gelap pada spektrum tersebut. Garis-garis ini dikenal sebagai garis absorpsi dan terjadi ketika elektron dalam gas menyerap foton dengan energi yang sesuai dengan perbedaan energi antara orbit elektron, sehingga elektron berpindah ke orbit yang lebih tinggi.



Gambar II. 11. Spektrum garis absorpsi (Strobel, 2007).

II.3 Machine Learning

II.3.1 *Data Mining* dan *Machine Learning*

Secara umum, data mining adalah proses mengubah data yang diperoleh dari pengamatan menjadi informasi yang bermanfaat. Informasi tersebut dapat diinterpretasikan melalui hipotesis atau teori, dan digunakan untuk membuat prediksi yang lebih lanjut. Selama beberapa dekade terakhir, terjadi peningkatan pesat dalam kemampuan komputasi yang tersedia, yang juga mengakibatkan peningkatan jumlah data yang tersedia dalam bentuk digital. Peningkatan yang signifikan dalam jumlah data yang tersedia menciptakan dunia digital yang dapat menggali informasi baru dan bermanfaat dari data yang telah dikumpulkan. Proses penemuan pengetahuan dalam basis data (KDD, knowledge discovery in databases) ini dikenal sebagai data mining.

Astronomi merupakan salah satu disiplin ilmu yang pertama mengalami peningkatan jumlah data pengamatan. Kemunculan data mining dalam bidang astronomi sendiri merupakan paradigma keempat. Paradigma awal astronomi terkait dengan teori dan observasi atau pengamatan. Paradigma ketiga membahas mengenai simulasi komputer atau pemodelan data. Dengan volume data yang sangat besar, diperlukan pendekatan paradigmatik baru yang sebagian besar bersifat otomatis. Pendekatan ini diperlukan untuk mengatasi tingkat pertumbuhan data sejak munculnya misi otomatis seperti Gaia, SDSS, Hipparchos, dan lainnya. Dalam istilah yang lebih formal, pendekatan ini melibatkan pemanfaatan komputasi mesin untuk menemukan pola dalam data digital dan menerjemahkan pola tersebut menjadi informasi yang bermanfaat yang dikenal sebagai pembelajaran mesin (*machine learning*). Pembelajaran ini diharapkan dapat memberikan manfaat bagi peneliti manusia dalam bentuk pembelajaran manusia yang berarti.

Algoritma pembelajaran mesin secara umum dapat dibagi menjadi pembelajaran terpandu (*supervised learning*) dan pembelajaran tak terpandu

(*unsupervised learning*), yang juga dikenal sebagai pembelajaran prediktif dan pembelajaran deskriptif. Metode *supervised learning* mengandalkan kumpulan objek pelatihan (data latihan) di mana properti targetnya diketahui dengan pasti, seperti dalam klasifikasi. Metode ini dilatih menggunakan kumpulan objek tersebut, dan klasifikasi yang dihasilkan diterapkan pada objek berikutnya di mana properti targetnya tidak tersedia. Objek tambahan ini disebut sebagai kumpulan pengujian (data uji). Kumpulan pelatihan harus representatif, artinya ruang parameter yang dijangkau oleh atribut input harus mencakup area yang digunakan dalam algoritma.

Beberapa penerapan teknik tersebut dapat ditemukan dalam berbagai bidang, seperti klasifikasi antara bintang dan galaksi, klasifikasi galaksi berdasarkan morfologi, klasifikasi populasi deret utama dalam katalog Hipparcos, dan klasifikasi katai putih dalam komponen Galaktiknya. Sejak saat itu, banyak pendekatan yang telah diusulkan, terutama berdasarkan jaringan saraf tiruan (ANN, *artificial neural network*), pohon keputusan (*decision tree*), analisis diskriminan, dan metode lainnya. Beberapa karakteristik metode yang sering digunakan dalam astronomi ditunjukkan oleh Tabel II.1.

Tabel II. 2. Karakteristik metode yang sering digunakan dalam Astronomi Ball, N. M., & Brunner, R. J. (2010).

Algoritma	A_1	A_2
Artificial Neural Network	- Aproksimasi yang baik dari fungsi nonlinier, - kemampuan prediksi baik, - banyak digunakan dalam astronomi, - tahan terhadap atribut yang tidak relevan.	- *Black-box model*, - banyak parameter yang dapat disesuaikan, - dipengaruhi oleh data yang noisy, - dapat overfit, - waktu pelatihan model yang lama, - tidak ada *missing values*.
Decision Tree	- Algoritma *data mining* yang populer, - dapat memasukkan dan menghasilkan variabel numerik atau kategori, - model yang dapat ditafsirkan (*white box model*), - tahan terhadap *outliers*, atribut yang *noisy* atau redundan.	- Dapat menghasilkan pohon kompleks yang membutuhkan pemangkasan (*pruning*), - daya prediksi tidak lebih baik daripada JST, SVM atau kNN, - dapat overfit, - banyak parameter yang dapat disesuaikan.
Support Vector Machine	- Dapat mengatasi data yang *noisy*, - memberikan tingkat kesalahan yang diharapkan, - algoritma populer dalam astronomi, - dapat mengaproksimasi fungsi nonlinier, - skalabilitas yang baik dengan beberapa jumlah atribut, - solusi unik (tidak ada minimal lokal).	- Lebih sulit untuk mengklasifikasikan lebih dari dua kelas, - waktu pelatihan model yang lama, - interpretabilitas yang kurang baik, - kurang baik dalam menangani atribut yang tidak relevan, - dapat melakukan overfit, - beberapa parameter yang dapat disesuaikan.
Nearest Neighbor	- Menggunakan semua informasi yang tersedia, - tidak memerlukan pelatihan, - sedikit atau tidak ada parameter yang dapat disesuaikan, - daya prediksi yang baik.	- Komputasi yang lebih rumit, - tidak ada model yang dibuat, - dapat dipengaruhi oleh *noise* dan atribut yang tidak relevan.
Expectation Maximization	- Memberikan jumlah klaster dalam data, - konvergensi yang cepat, mengatasi data yang hilang, - dapat menghasilkan label kelas untuk *semi-supervised learning*.	- Dapat bias terhadap Gaussian, - minima lokal.

II.3.2 *decision tree*

Salah satu metode umum yang digunakan dalam penambangan data (data mining) adalah pohon keputusan (*decision tree*). *decision tree* merupakan salah satu metode pembelajaran terawasi (*supervised learning*) yang melakukan klasifikasi berdasarkan fitur-fitur yang diberikan pada data. Dalam *supervised learning*, data yang akan diproses telah memiliki kolom target, baik dalam bentuk variabel diskrit maupun kontinu.

	Fitur 1	Fitur 2	Fitur 3	Target
Objek 1	45	21	99	X
Objek 2	30	37	50	X
Objek 3	19	5	44	Y

Gambar II. 12. Contoh tabel data pada *decision tree*

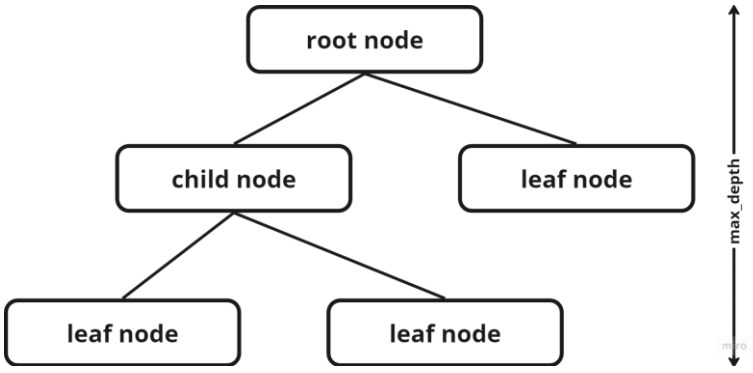
Pohon Keputusan memiliki beberapa algoritma yang digunakan untuk membangun modelnya. Terdapat tiga algoritma yang sering digunakan dalam membangun pohon dalam Pohon Keputusan, yaitu ID3 (Iterative Dichotomiser 3), C4.5, dan CART.

- ID3 akan membangun pohon tanpa adanya pemotongan (pruning) dan memilih fitur kategorikal pada setiap simpulnya yang memiliki gain informasi maksimum untuk target kategorikal. Algoritma ID3 digunakan untuk klasifikasi dengan menggunakan Information Gain untuk menentukan fitur yang akan digunakan dalam membangun pohon Pohon Keputusan.
- C4.5 merupakan algoritma yang digunakan untuk kasus klasifikasi dengan kriteria pemilihan fitur menggunakan metode Gain Ratio. Pohon yang dibangun menggunakan algoritma C4.5 dapat dipotong untuk mengurangi *overfitting* dengan menghentikan pemisahan simpul dan menggantikannya dengan simpul daun. Selain itu, data yang hilang pada algoritma C4.5 dapat langsung ditangani dengan melakukan penggantian nilai menjadi nilai rata-rata.
- CART merupakan algoritma yang digunakan untuk kasus klasifikasi dan regresi. Pemotongan pohon dalam algoritma

CART dilakukan ketika nilai simpul anak yang dihasilkan tidak memberikan peningkatan signifikan dalam akurasi. Pemotongan atau kriteria pemilihan fitur dalam algoritma CART bergantung pada masalah yang ingin ditangani. Dalam kasus klasifikasi, pemotongan biasanya menggunakan Indeks Gini atau Entropi. Sedangkan dalam kasus regresi, pemotongan dilakukan dengan menggunakan metode MAE (Mean Absolute Error) atau MSE (Mean Squared Error).

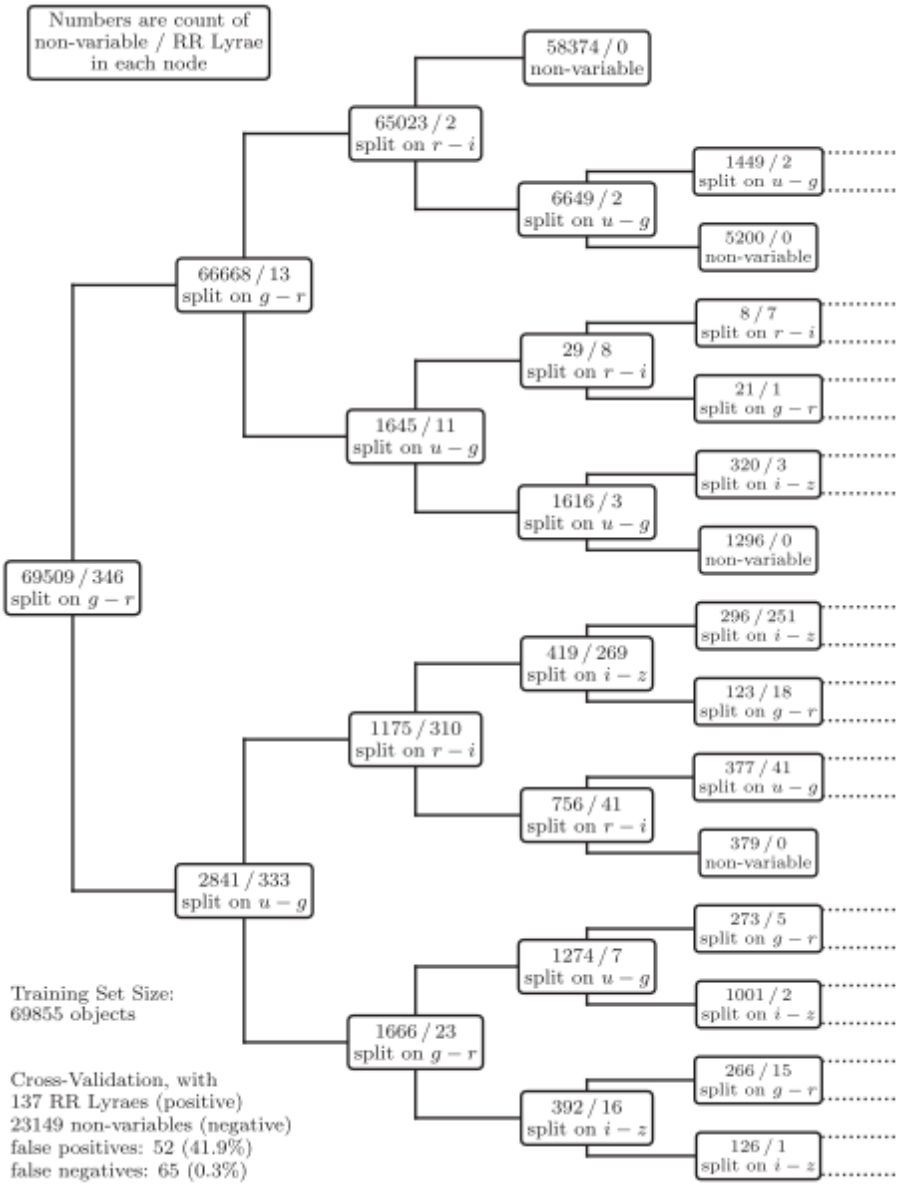
Pohon Keputusan yang dibangun dalam modul scikit-learn menggunakan algoritma CART. Dalam kasus klasifikasi, target kelas yang ditangani oleh modul scikit-learn tidak berbentuk variabel kategorikal. Target kelas untuk klasifikasi yang disediakan oleh modul scikit-learn berupa probabilitas suatu objek diklasifikasikan sebagai kelas tertentu.

Pertama, algoritma *decision tree* akan mengevaluasi seluruh fitur data untuk menentukan simpul akar (*root node*). Penentuan fitur yang digunakan sebagai simpul akar tersebut akan mengikuti suatu kriteria tertentu, seperti Entropi atau Indeks Gini, yang menghitung tingkat kesalahan atau akurasi dari hasil pemisahan yang dilakukan. Fitur yang dipilih adalah yang memiliki tingkat kesalahan paling rendah. Setiap objek atau titik data kemudian akan diklasifikasikan berdasarkan simpul akar tersebut. Proses selanjutnya melibatkan pemisahan data berdasarkan fitur yang memberikan tingkat kesalahan minimum setelahnya. Fitur-fitur yang digunakan untuk pemisahan setelah simpul akar disebut simpul anak (*child node*).



Gambar II. 13. Istilah simpul dalam decision tree

Proses pemisahan dengan simpul anak dilakukan secara berulang hingga tingkat kesalahan pemisahan tidak mengalami perbaikan atau mencapai tingkat kesalahan minimum. Terdapat juga proses lain yang dapat menghentikan pemisahan, seperti kriteria jumlah minimum populasi objek dalam sebuah simpul, kriteria jumlah maksimum simpul antara simpul sebelumnya dan simpul akar (misalnya, kriteria kedalaman maksimum pohon atau `max_depth`), atau kriteria lain seperti kriteria nilai tingkat kesalahan minimum untuk melanjutkan proses pemisahan. Contoh skema pemisahan *decision tree* dalam bidang Astronomi dapat dilihat pada Gambar II.14.



Gambar II. 14. decision tree untuk klasifikasi RR Lyrae. Angka dalam setiap simpul merupakan statistic dari sampel pelatihan ~70000 objek. (Ivezić, Ž., dkk. 2020).

Untuk membangun *decision tree*, perlu dilakukan pemilihan fitur dan nilai yang digunakan untuk membagi data. Secara umum, kriteria pemisahan tersebut didasarkan pada perolehan informasi atau entropi dari data yang tersedia. Persamaan yang digunakan untuk menghitung entropi dalam suatu dataset adalah sebagai berikut:

$$E(Q_m) = -\sum_k p_{mk} \log(p_{mk}) \qquad (II.1)$$

Dengan p_{mk} merupakan probabilitas suatu objek masuk ke dalam kelas k pada simpul m. sedangkan perolehan informasi (Information Gain) dapat didefinisikan sebagai pengurangan dalam entropi yang diakibatkan oleh pemisahan pada data, yaitu selisih antara entropi simpul anak dan simpul sebelumnya. Untuk klasifikasi biner, dengan $m = 0$ merepresentasikan titik data dibawah threshold pemisahan dan $i = 1$ untuk titik data diatas pemisahan, perolehan informasi ($IG(Q)$) didefinisikan sebagai berikut:

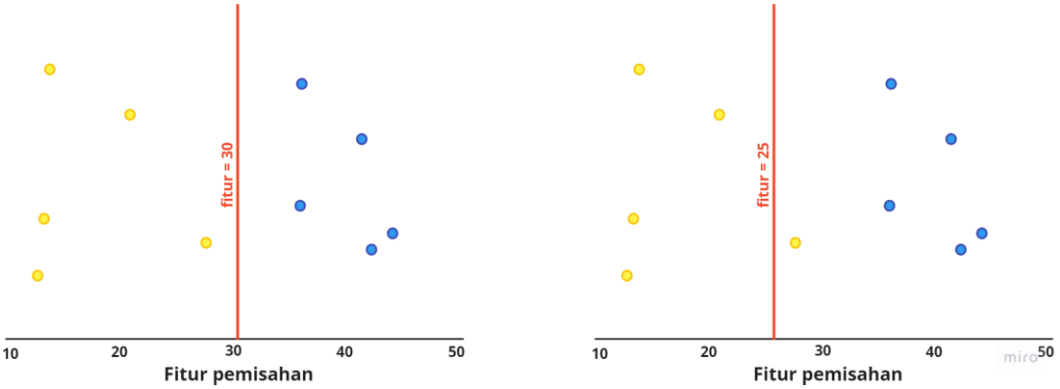
$$IG(Q|Q_m) = E(Q_m) - \sum_{i=0}^{1} \frac{N_m}{N} E(Q_m) \qquad (II.2)$$

Dengan N_m merupakan jumlah titik data untuk kelas m dan $E(Q_m)$ merupakan entropi pada kelas tersebut. Selain entropi, terdapat juga *loss function* lain yang biasa digunakan dalam *decision tree*, yaitu koefisien Gini. Indeks Gini atau koefisien Gini akan menghitung probabilitas sebuah titik data akan salah diklasifikasikan jika dipilih secara acak dari suatu dataset dan label dipilih secara acak berdasarkan distribusi klasifikasi dalam dataset. Persamaan yang digunakan dalam menghitung indeks Gini adalah sebagai berikut:

$$G = \sum_k p_{mk}(1 - p_{mk}) \qquad (II.3)$$

Dalam klasifikasi biner, untuk menentukan fitur yang akan digunakan untuk pemisahan dataset, proses yang dilakukan adalah dengan menghitung Indeks Gini untuk seluruh dataset menggunakan Persamaan II.3. terlebih dahulu. Kemudian, dihitung Gini impuritas pada

suatu fitur pemisahan tertentu melalui Persamaan II.3. untuk kedua kelas.



Gambar II. 15. Distribusi sampel dengan pemisahan pada dua fitur berbeda.

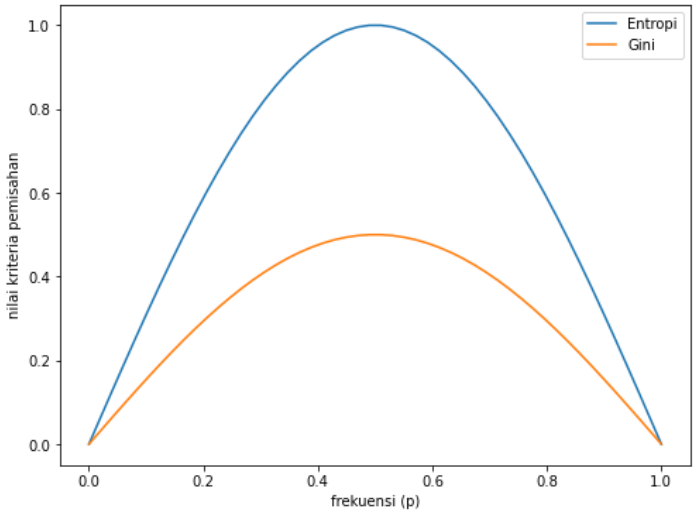
Setelah Gini impuritas diperoleh, dilakukan pembobotan terhadap nilai sampel yang diklasifikasikan dengan benar atau salah. Hal ini kemudian didefinisikan sebagai Indeks Gini dengan persamaan sebagai berikut:

$$IG(f) = \sum_{k} \left(\frac{N_k}{N} G_k \right) \qquad (II.4)$$

Dengan N_k merupakan jumlah data yang masuk ke dalam kelas k, N merupakan jumlah data, dan G_k merupakan impuritas Gini pada kelas k. Setelah indeks Gini diperoleh, kemudian dihitung Gini keseluruhan yang didefinisikan sebagai Gini Gain, menggunakan persamaan sebagai berikut:

$$Gini\ Gain = G(Q) - IG(f) \qquad (II.5)$$

Dengan $G(Q)$ merupakan indeks Gini untuk seluruh dataset dan $IG(f)$ merupakan indeks Gini pada suatu fitur tertentu. Semakin besar Gini Gain, maka semakin baik klasifikasi yang dilakukan oleh fitur tersebut. Fitur dengan nilai Gain tertinggi akan digunakan sebagai kriteria pemisahan simpul akar (*root node*) , sementara fitur dengan Gain selanjutnya akan digunakan sebagai kriteria pemisahan untuk menghasilkan simpul anak (*child node*).



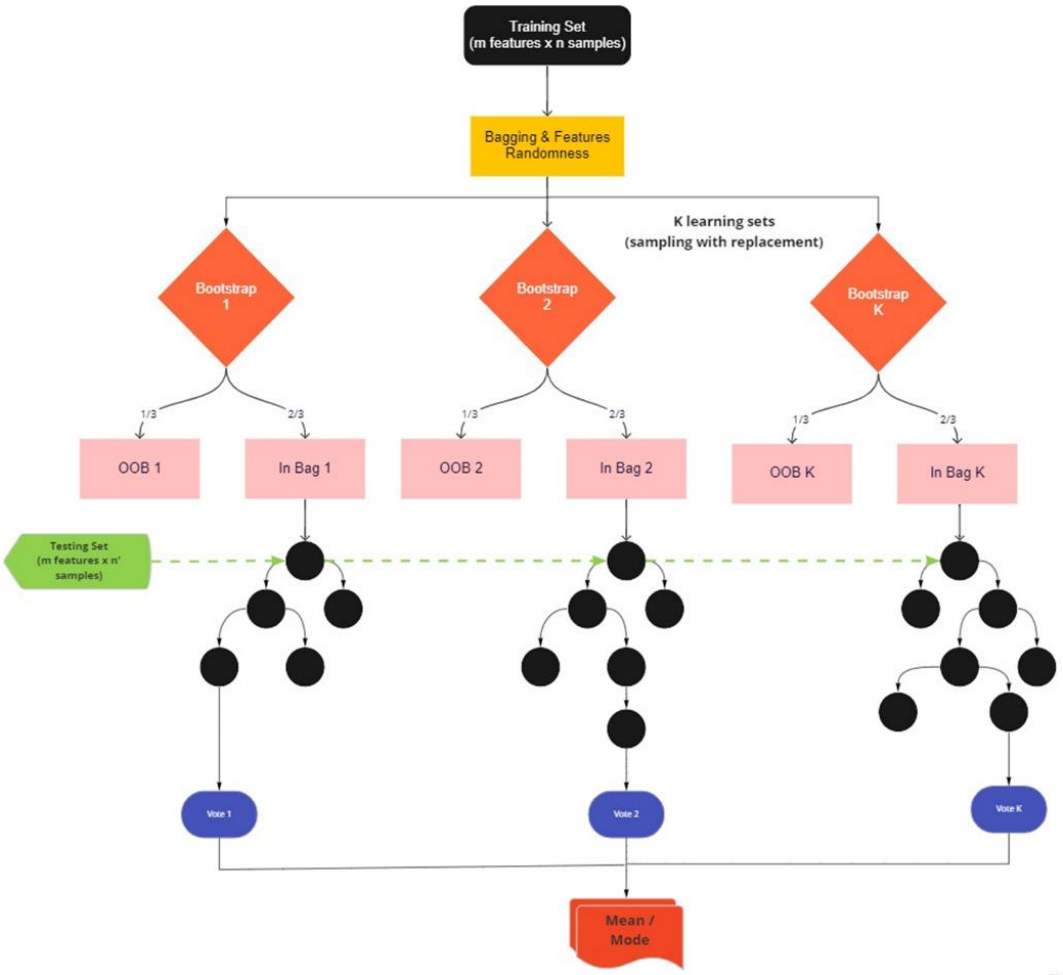
Gambar II. 16. Perbedaan nilai Entropi dan Gini indeks pada klasifikasi biner.

Perbedaan kriteria pemisahan entropi dan indeks Gini untuk klasifikasi biner dapat dilihat pada Gambar II.16. Dari gambar tersebut, terlihat bahwa nilai maksimum untuk entropi adalah 1, sedangkan nilai maksimum untuk indeks Gini adalah 0.5. Nilai maksimum dari kedua kriteria ini diperoleh ketika frekuensi mencapai nilai 0.5, yang berarti kelas dapat diklasifikasikan dengan benar untuk kedua kelas. Indeks Gini merupakan kriteria yang sering digunakan dalam *decision tree* ketika kolom target berbentuk data kategorikal. Hal ini dikarenakan kriteria pemisahan dengan indeks Gini memberikan komputasi yang lebih efisien serta lebih mudah dipahami.

II.3.3 Random Forest

Dasar algoritma dari Random Forest merupakan gabungan dari beberapa pohon keputusan (*decision tree*). Salah satu keuntungan dari menggunakan Random Forest adalah mengurangi risiko *overfitting* yang dapat terjadi pada *decision tree*. Metode Random Forest menggunakan beberapa fitur untuk mengurangi *overfitting*, yaitu bagging (bootstrap aggregating) dan fitur randomness. Dalam Random Forest, kedua fitur tersebut akan secara acak mengambil data dan fitur pemisahan dari dataset, yang kemudian digunakan untuk membentuk pohon keputusan. Fitur bagging juga akan mengambil sub-sampel dari data pelatihan dengan penggantian, sehingga memungkinkan penggunaan kembali data yang telah terpilih untuk membangun pohon-pohon lainnya. Setelah

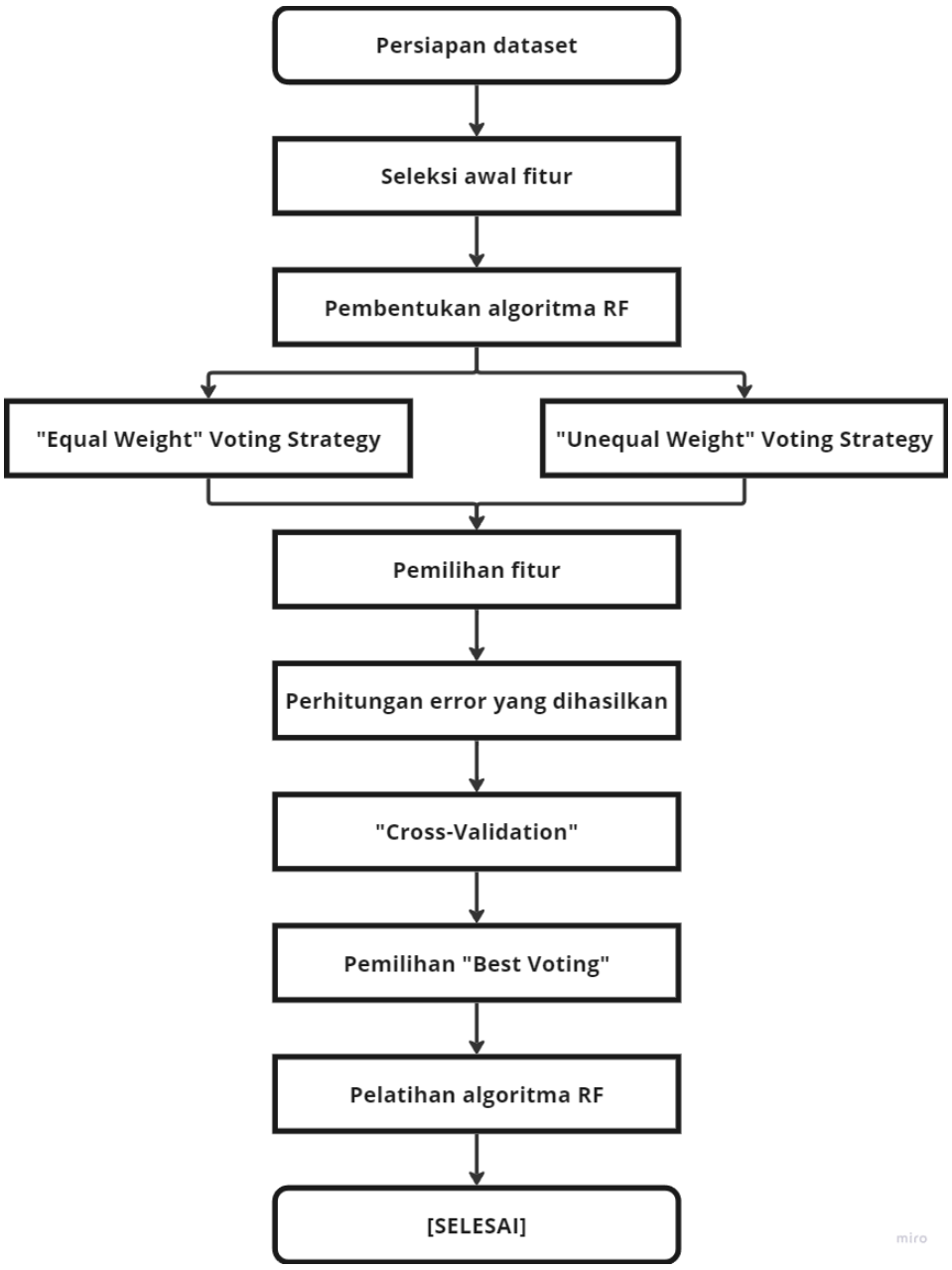
terbentuk beberapa pohon keputusan, Random Forest akan memilih satu keputusan dari setiap pohon, dan kemudian menggunakan rerata atau modus untuk menentukan variabel target akhir. Gambaran umum algoritma Random Forest dapat dilihat pada Gambar II.17.



Gambar II. 17. Contoh ensemble Random Forest.

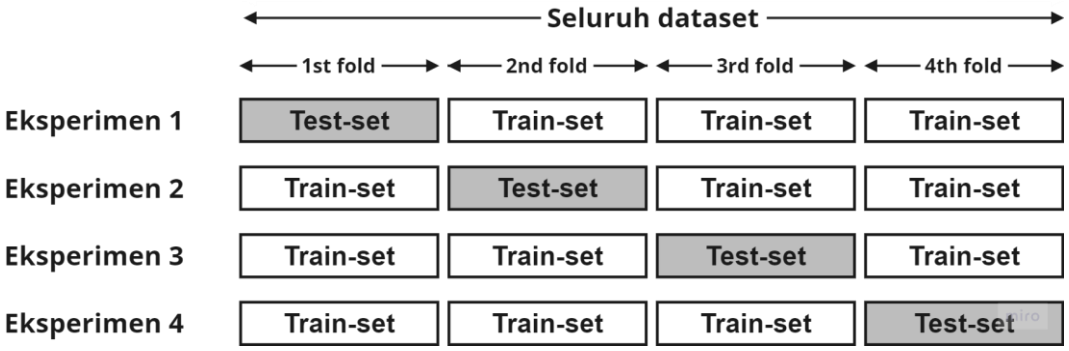
Secara umum, langkah-langkah dalam skema alur Random Forest terlihat pada Gambar II.18. Tahap awal adalah menyiapkan dataset pelatihan (train-set). Selanjutnya, dilakukan pengambilan sampel secara acak dengan metode bagging (pengambilan sampel acak dengan pengembalian) untuk membangun setiap pohon dalam ensemble Random Forest. Tahap selanjutnya adalah memilih fitur sebagai kriteria pemisahan pada setiap *decision tree*. Fitur yang dipilih tersebut kemudian digunakan untuk membangun setiap pohon dalam ensemble Random Forest sehingga menghasilkan target klasifikasi. Setelah itu, dilakukan perhitungan kesalahan dari seluruh pohon dalam ensemble

Random Forest dengan mempertimbangkan dua strategi pemilihan. Strategi pemilihan yang pertama adalah "Equal Weight" Voting Strategy, di mana setiap pohon diberi nilai bobot yang sama. Strategi pemilihan yang kedua adalah "Unequal Weight" Voting Strategy, di mana setiap pohon memberikan nilai bobot yang berbeda. Kemudian, tingkat kesalahan dari kedua strategi pemilihan tersebut dievaluasi, dan strategi pemilihan yang memberikan nilai minimum dipilih sebagai strategi pemilihan dalam ensemble Random Forest.



Gambar II. 18. Algoritma Random Forest (Mohapatra, N., dkk. 2020).

Untuk mengurangi *overfitting*, dapat digunakan fitur Cross-validation dengan membagi dan mengevaluasi data sampel menjadi beberapa bagian. Penggunaan Cross-validation sangat efektif pada dataset kecil, namun pada dataset yang besar, cukup dilakukan satu kali validasi karena membutuhkan memori komputasi yang besar. Pada tahap akhir, setelah ensemble Random Forest terbentuk, algoritma akan diujikan pada data pengujian (test-set) dan dihasilkan akurasi model berdasarkan data pengujian tersebut. Jika menggunakan fitur Cross-validation, tingkat kesalahan dari algoritma Random Forest dapat dievaluasi dengan menghitung rerata dari kesalahan tiap dataset.



Gambar II. 19. Cross-validation dengan 4 folds.

Random Forest dapat dibagi menjadi dua jenis berdasarkan variabel target yang ditangani, yaitu klasifikasi Random Forest (Random Forest Classification) untuk data target berupa kategorikal, dan regresi Random Forest (Random Forest Regression) untuk data target berupa data kontinu. Parameter input yang memiliki persamaan sifat pada kedua jenis tersebut adalah `n_estimators`. Parameter input `n_estimators` menentukan jumlah pohon maksimum yang akan dibentuk dalam ensemble Random Forest. Persamaan sifat pada kedua jenis Random Forest tersebut adalah semakin banyak pohon yang terbentuk, model yang dibangun akan semakin kompleks, sehingga kemungkinan nilai akurasi juga akan meningkat. Namun, perlu diperhatikan bahwa jumlah pohon yang terlalu banyak dapat menyebabkan *overfitting*, yang mengakibatkan penurunan akurasi dari model yang terbentuk. Perbedaan dari kedua jenis Random Forest tersebut dapat dilihat pada Tabel II.2.

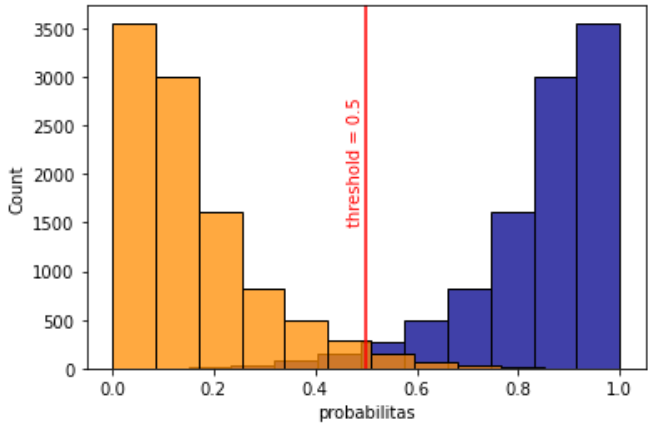
Tabel II. 3. Perbedaan klasifikasi dan regresi Random Forest

Perbedaan	Klasifikasi Random Forest	Regresi Random Forest
Variabel Target	Variabel targetnya berupa variabel kategorikal yang menggambarkan kelas.	Variabel targetnya berupa variabel kontinu yang merupakan nilai numerik atau kuantitatif.
Pemotongan pohon (*pruning*)	Parameter input : `max_features`. Dalam klasifikasi, variasi fitur yang digunakan akan membantu dalam mempengaruhi kompleksitas, keragaman model dan mencegah *overfitting*.	Parameter input : `max_depth`. Dalam regresi, tujuan target merupakan variabel kontinu sehingga dengan mengatur kedalaman maksimum pohon maka akan mempengaruhi kompleksitas dan kemampuan generalisasi model terhadap data baru.
Kriteria pemisahan (criterion)	Dalam konteks klasifikasi, kriteria pemisahan umumnya menggunakan ‘gini’ (Gini impurity) atau “entropy” (Information Gain) untuk mengukur keberagaman kelas dalam setiap setiap simpul pemisahan.	Dalam konteks regresi, kriteria pemisahan umumnya menggunakan "mse" (Mean Squared Error) atau “mae” (Mean Absolute Error) untuk mengukur kesalahan prediksi dan meminimalkan varians hasil prediksi.

Strategi voting	"Unequal weight" voting strategy. Dalam konteks klasifikasi, biasanya digunakan strategi voting dengan bobot yang tidak sama, di mana setiap pohon memiliki kontribusi yang berbeda dalam proses voting.	"Equal weight" voting strategy. Dalam konteks regresi, biasanya digunakan strategi voting dengan bobot yang sama, di mana setiap pohon memiliki bobot yang setara dan hasil prediksi diperoleh dengan mengambil rata-rata dari prediksi semua pohon.

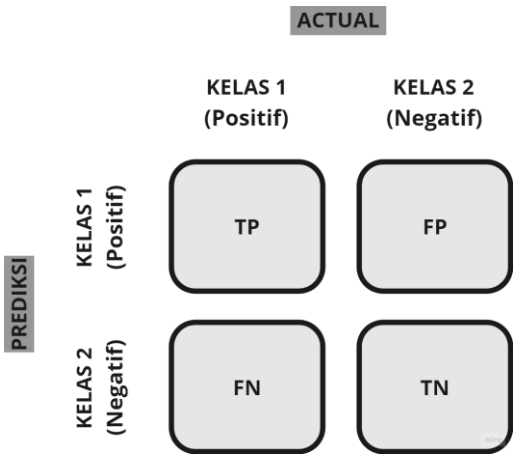
Evaluasi Model Random Forest

Dalam klasifikasi, nilai target yang dihasilkan merupakan probabilitas untuk setiap kelas. Pada kasus biner, nilai 1 digunakan untuk merepresentasikan kelas pertama, sementara nilai 0 digunakan untuk merepresentasikan kelas lainnya. Oleh karena itu, diperlukan penentuan threshold pada hasil klasifikasi. Nilai probabilitas di atas threshold akan diklasifikasikan sebagai kelas 1, sementara nilai di bawah threshold akan diklasifikasikan sebagai kelas 0.



Gambar II. 20. Distribusi probabilitas hasil prediksi dengan threshold 0.5 pada klasifikasi biner.

Setelah threshold ditentukan dan semua objek diklasifikasikan, kita dapat membangun Confusion Matrix untuk melihat jumlah kelas yang diklasifikasikan dengan benar dan yang salah. Contoh Confusion Matrix untuk kelas biner ditunjukkan pada Gambar II.21.

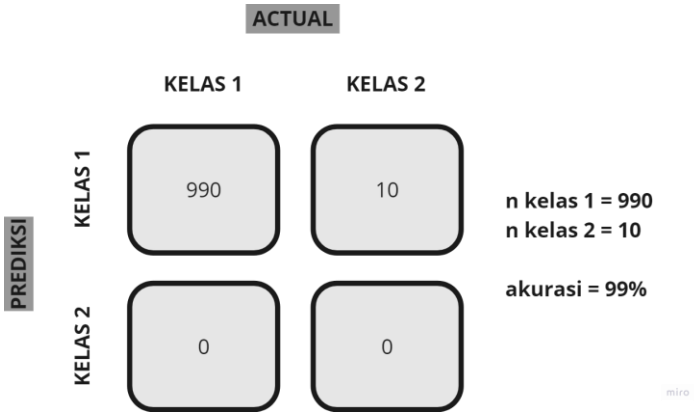


Gambar II. 21. Confusion Matrix klasifikasi biner.

Salah satu metrik umum yang digunakan untuk mengevaluasi model yang dibangun adalah akurasi. Persamaan yang digunakan untuk menghitung akurasi adalah:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \qquad (II.6)$$

Namun, akurasi sebuah model terbatas pada jenis data yang diamati. Ketika salah satu kelas memiliki jumlah yang dominan dalam target kelasnya (imbalanced dataset), akurasi dapat menghasilkan nilai yang tinggi meskipun semua data dikategorikan sebagai kelas yang dominan tersebut. Contoh kasus limitasi akurasi serta perhitungannya dapat dilihat pada Gambar II.22.



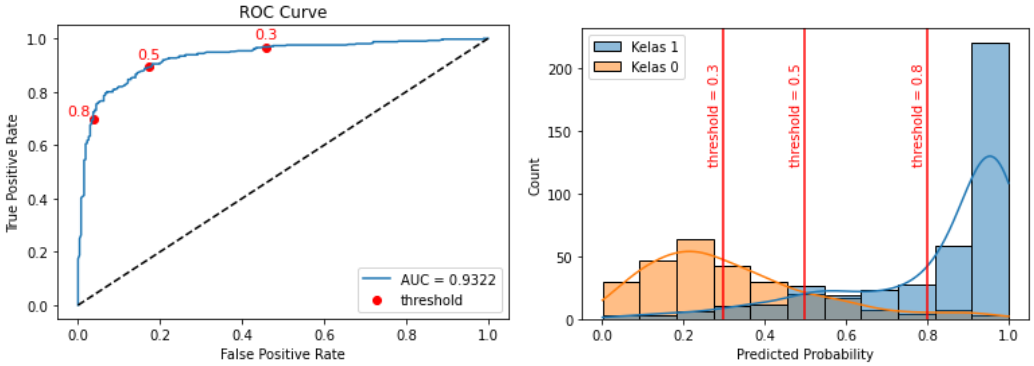
Gambar II. 22. Contoh limitasi dari metrik akurasi. Kasus ketika semua titik data diklasifikasikan sebagai kelas 1 namun menghasilkan akurasi 99%.

Oleh karena itu, diperlukan evaluasi model yang lebih kompleks selain akurasi. Salah satu model yang umum digunakan dengan tingkat kompleksitas yang lebih tinggi adalah kurva Receiver Operating Characteristic (ROC). Kurva ROC adalah fungsi dari True Positive Rate (TPR) pada sumbu y dan False Positive Rate (FPR) pada sumbu x. Persamaan untuk keduanya diberikan oleh persamaan berikut:

$$FPR = \frac{FP}{FP + TN} \qquad (II.7)$$

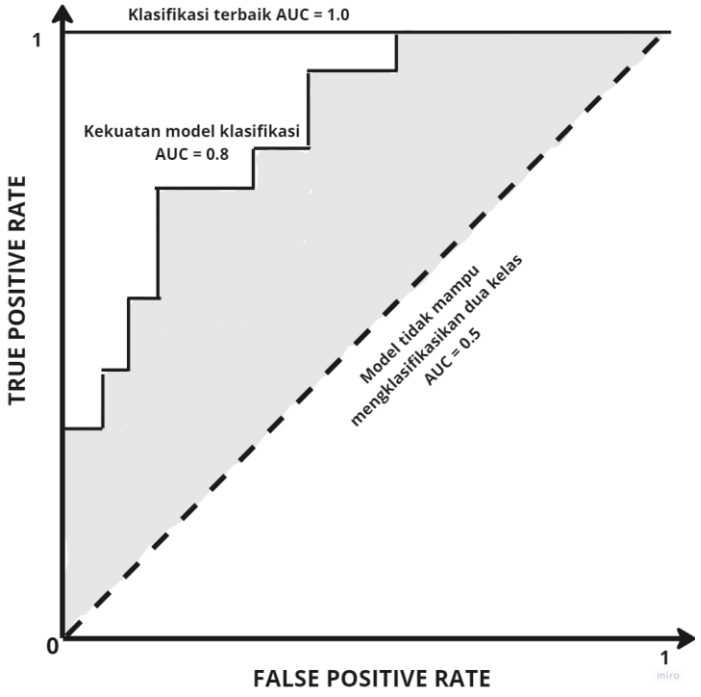
$$TPR = \frac{TP}{TP + FN} \qquad (II.8)$$

Dengan mengubah nilai threshold pada algoritma Random Forest, lokasi titik data atau objek dalam distribusi sampel atau Confusion Matrix akan berubah. Sehingga performa dan akurasi klasifikasi direpresentasikan sebagai titik pada kurva ROC. Berikut contoh kurva ROC dan distribusi probabilitas prediksi dengan variasi nilai threshold yang ditunjukkan pada Gambar II.23.



Gambar II. 23. Kurva ROC dan distribusi probabilitas hasil prediksi pada berbagai threshold.

TPR 0 dan FPR 0 menunjukkan bahwa semua titik data diklasifikasikan sebagai kelas 0. TPR 1 dan FPR 1 menunjukkan bahwa semua titik data diklasifikasikan sebagai kelas 1. Kurva ROC yang ideal akan memiliki nilai TPR 1 dan FPR 0. Nilai akurasi dari berbagai variasi threshold tersebut kemudian diadopsi sebagai akurasi model secara umum. Istilah yang umum digunakan untuk menghitung nilai tersebut adalah Area Under Curve (AUC) yang merupakan luas area di bawah kurva ROC. Model dengan akurasi terbaik akan memberikan nilai AUC 1, sedangkan nilai AUC 0.5 menunjukkan bahwa titik data diklasifikasikan secara acak dan model tidak mampu mengklasifikasikan kedua kelas tersebut.



Gambar II. 24. Contoh kurva ROC.

Untuk mengevaluasi akurasi model Random Forest pada data target kontinu, digunakan dua metrik evaluasi yang umum, yaitu Mean Absolute Error (MAE) dan Mean Squared Error (MSE). MAE merupakan nilai rerata absolut dari selisih antara hasil prediksi dan nilai sebenarnya. Semakin kecil nilai MAE, semakin baik performa dan akurasi model yang dihasilkan. Sedangkan MSE merupakan rata-rata kuadrat selisih antara hasil prediksi dan nilai sebenarnya. Semakin kecil MSE, semakin baik akurasi dan performa model yang dihasilkan. Persamaan untuk MAE dan MSE diberikan oleh persamaan seperti berikut:

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \qquad (II.8)$$

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \qquad (II.9)$$

Dengan $\hat{y}_i$ merupakan nilai prediksi dari sampel ke $-i$ dan y_i merupakan nilai yang sebenarnya. Selain metrik-metrik tersebut, hasil prediksi dapat divisualisasikan dalam bentuk scatter plot yang membandingkan hasil prediksi dengan titik data sebenarnya. Dengan visualisasi ini, dapat dilihat apakah hasil regresi yang dihasilkan oleh model fit dengan titik data yang sebenarnya. Jika kedua garis hampir sama, maka dapat dikatakan bahwa model Random Forest memiliki performa dan akurasi yang baik.

BAB III

DATA DAN PERANGKAT PENELITIAN

III.1 LAMOST Data Release 8

LAMOST (*The Large Sky Area Multi-Object Fiber Spectroscopic Telescope*) merupakan fasilitas penelitian ilmiah nasional China yang dioperasikan oleh NAO (*National Astronomical Observatories*), Chinese Academy of Sciences. Survei LAMOST terdiri dari dua bagian, yaitu LEGAS (LAMOST ExtraGAlactic Survey) dan LEGUE (LAMOST Experiment for Galactic Understanding and Exploration), yang bertujuan untuk mempelajari struktur bintang di dalam Galaksi Bima Sakti. Pengamatan dilakukan menggunakan teleskop reflektor Schmidt yang dilengkapi dengan 4000 serat optik, dengan medan pandang langit (FoV) sebesar 20 derajat persegi. Desain yang unik ini memungkinkan LAMOST untuk mengambil sekitar 4000 spektrum dalam satu eksposur, bahkan pada magnitudo kecerlangan yang sangat redup sekitar r = 19, dengan resolusi R = 1800.

Publikasi Survei LAMOST Data Release 8 dilakukan pada tanggal 30 September 2022. Survei LAMOST, berdasarkan resolusi spektrumnya, terbagi menjadi dua bagian, yaitu LSR (*Low-Resolution Spectroscopic Survey*) dan MSR (*Medium-Resolution Spectroscopic Survey*). Pada data release delapan dari LAMOST, terdapat sekitar 10.633.515 spektrum resolusi rendah yang telah dikalibrasi terhadap panjang gelombang. Spektrum tersebut mencakup panjang gelombang dalam rentang 3700 Å hingga 9000Å dengan resolusi sekitar 1800 pada panjang gelombang 5000Å.

III.2 KATALOG Katai Putih LSR LAMOST DR8[3]

Katalog katai putih yang tersedia dalam survei resolusi rendah LAMOST DR8 berisi 15.601 spektrum katai putih. Katalog tersebut juga mencakup beberapa parameter penting, seperti subkelas katai putih (`wd_subclass`), temperatur efektif (`teff`), error dari temperatur efektif (`teff_err`), gravitasi permukaan (`logg`), dan *error* dari gravitasi permukaan (`logg_err`). Subkelas katai putih didapatkan dengan menerapkan metode machine learning "LASSO+SVM", sedangkan empat parameter lainnya didapatkan melalui algoritma fitting least-square. Penentuan subkelas katai putih dilakukan dengan menggunakan model dan parameter fisis yang disediakan oleh J.K. Zhao dan D. Koester untuk subkelas DA dan DB. Parameter fisis bintang katai putih DA dan DB yang diadopsi memiliki rentang temperatur efektif antara 5.000 K hingga 80.000 K dan gravitasi permukaan antara 7.0 hingga 9.5. Katalog katai putih LAMOST LSR dapat diunduh dalam dua format file, yaitu FITS dan CSV.

III.3 GAIA Data Release 3

Satelit Gaia dilengkapi dengan tiga instrumen utama dalam misinya, yaitu instrumen astrometri yang mengumpulkan citra dalam rentang pita G Gaia (330 – 1050 nm), prisma fotometri BP (330 – 680 nm) dan RP (640 – 1050 nm) untuk spektrum resolusi rendah, serta RVS. Resolusi spektrumnya bervariasi antara 30 – 100 untuk BP dan 70 – 100 untuk RP dalam satuan $\lambda/\Delta\lambda$ (tergantung pada posisi dalam spektrum dan CCD). Salah satu produk dari kalibrasi BP-RP adalah xp_mean_sampled_spectrum. Kalibrasi spektrum Gaia dilakukan secara eksternal dengan menggunakan bintang standar spektroskopi Gaia (SPSS) untuk mendapatkan panjang gelombang dan fluks absolut. Semua spektrum pada kalibrasi eksternal memiliki panjang gelombang absolut yang sama, yaitu terdiri dari 343 nilai mulai dari 336 nm hingga 1020 nm dengan interval 2 nm per langkah (Montegriffo, P., dkk. 2022).

[3]http://www.lamost.org/dr8/v2.0/catalogue

III.4 Gaia TAP+ (astroquery.Gaia)

Gaia TAP+ adalah sebuah layanan yang disediakan untuk mengakses Arsip Gaia, yang merupakan kumpulan data dari Badan Antariksa Eropa (ESA). Akses ini menggunakan layanan TAP+ REST (*Table Access Protocol*) sebagai dasar teknologinya. Bahasa query yang digunakan dalam TAP adalah *Astronomical Data Query Language* (ADQL), yang merupakan bahasa yang umum digunakan untuk mengambil data dari basis data astronomi. TAP menyediakan dua mode operasi yang dapat digunakan untuk mengakses data, yaitu mode Sinkron dan mode Asinkron. Dalam mode Sinkron, data akan diberikan setelah server menerima permintaan. Sementara itu, dalam mode Asinkron, informasi yang diperlukan untuk memeriksa status permintaan akan diberikan terlebih dahulu sebelum data sebenarnya disampaikan.

III.5 Scikit-learn

Scikit-learn merupakan sebuah modul yang digunakan dalam pembelajaran mesin dan telah dikembangkan sejak tahun 2007 oleh David Cournapeau sebagai proyek dalam Google Summer of Code. Modul ini menyediakan berbagai algoritma pembelajaran mesin yang dapat digunakan dalam bentuk *supervised* maupun *unsupervised*. Dengan adanya modul ini, tersedia beragam pilihan algoritma untuk melakukan klasifikasi, regresi, pengelompokan, dan pengurangan dimensi. Scikit-learn dirancang dengan mengintegrasikan beberapa modul lain seperti NumPy, SciPy, Pandas, dan Seaborn.

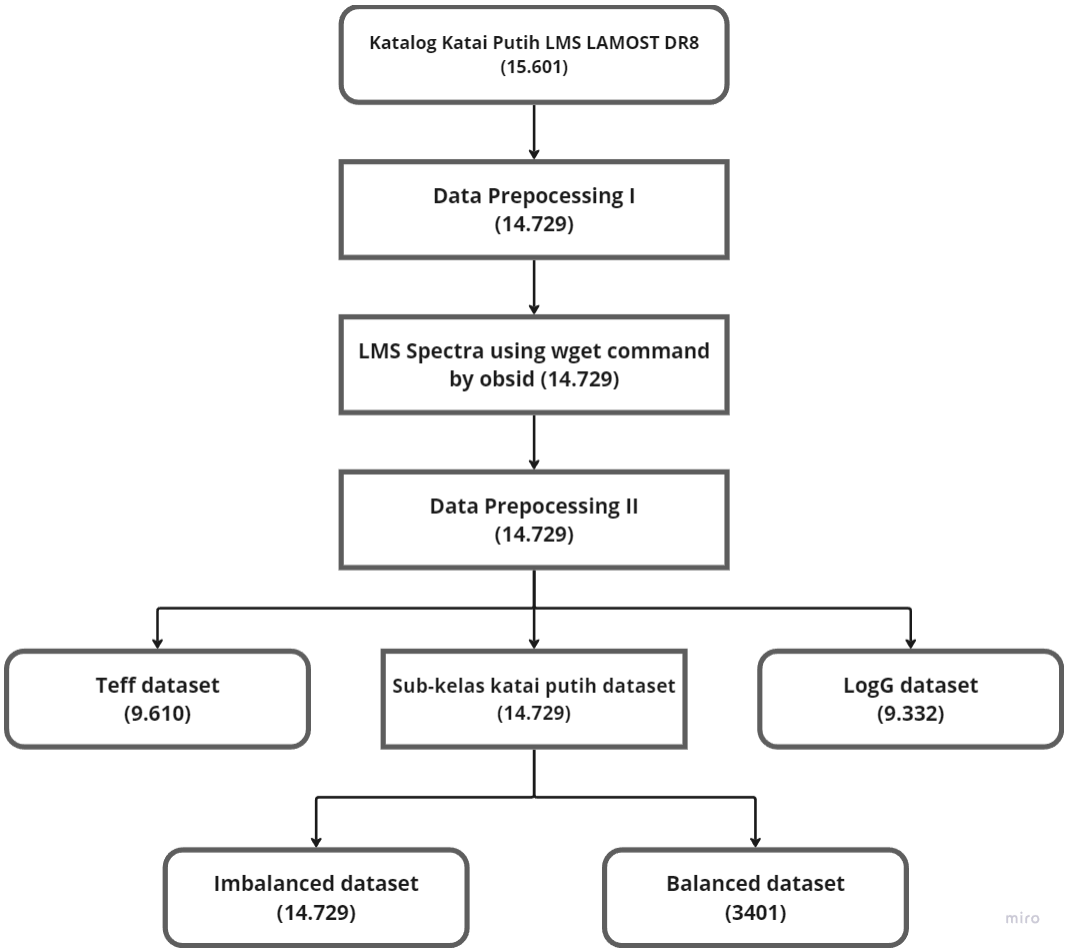
III.6 Data Utama

Data subkelas katai putih yang tersedia dalam katalog LAMOST DR8 akan dibagi menjadi dua bagian, yakni data pertama dan data kedua. Data pertama melibatkan beragam parameter yang diperoleh dari LAMOST, seperti spektrum, temperatur efektif, gravitasi permukaan, dan kelimpahan untuk setiap objek katai putih. Di sisi lain, data kedua menggunakan parameter fisis berupa fluks dan panjang gelombang yang

diperoleh dari Gaia.

III.6.1 Data Pertama

Data pertama menggunakan semua parameter yang tersedia dari pengamatan LAMOST DR8 dalam katalog katai putih LRS (*Low Resolution Search*) LAMOST DR8. Proses untuk mendapatkan data pertama dijelaskan dalam skema yang terlihat pada Gambar III.1.



Gambar III. 1.Skema mendapatkan data LAMOST DR8

Pada katalog katai putih LSR LAMOST DR8 terdapat 15.601 data katai putih dengan berbagai subkelas. Dari katalog tersebut kemudian dilakukan pra-pemrosesan data I, yaitu seleksi terhadap *Signal-to-Noise Ratio* (SNR) untuk setiap objek. Untuk kasus klasifikasi, dilakukan seleksi objek yang termasuk dalam subkelas katai putih DA dan DAZ. Sedangkan untuk kasus regresi, dilakukan seleksi objek berdasarkan

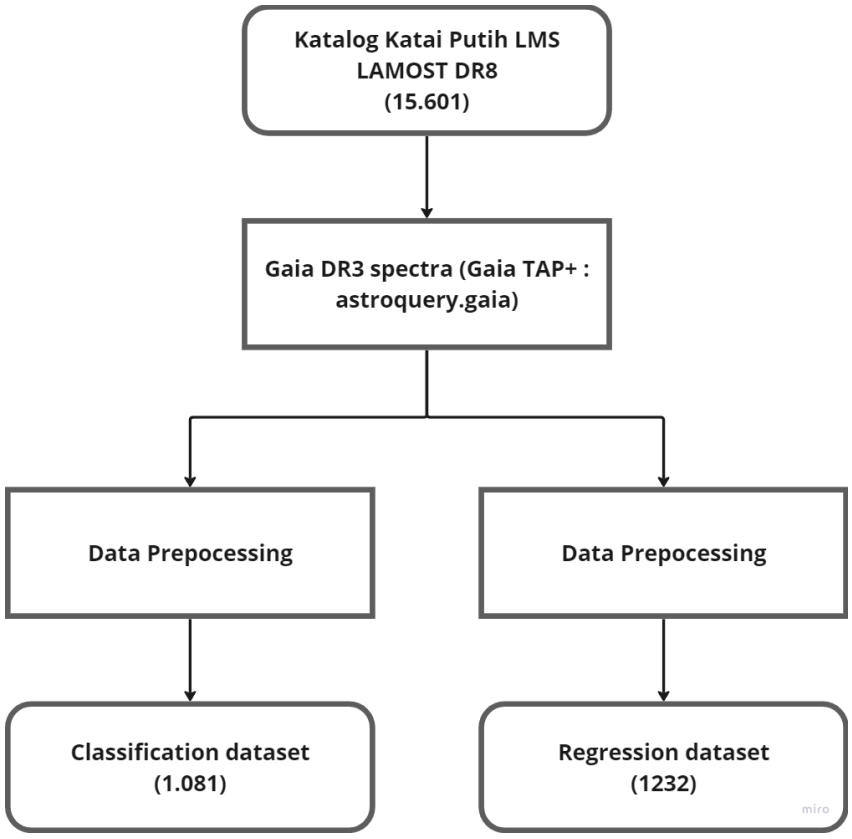
parameter fisis seperti temperatur efektif dan gravitasi permukaan. Spektrum LSR LAMOST DR8 dari setiap objek dalam katalog tersebut diperoleh menggunakan perintah `wget`. Setelah mendapatkan seluruh spektrum objek katai putih, dilakukan pra-pemrosesan data II. Pada tahap pra-pemrosesan data II, dilakukan normalisasi panjang gelombang untuk setiap objek, pemilihan panjang gelombang antara 3700 A hingga 6700 A, serta penggabungan spektrum setiap objek menjadi satu tabel. Output yang dihasilkan berupa tabel dengan fitur berupa panjang gelombang. Setiap baris dalam tabel mewakili objek katai putih dengan nilai flux. Sebagian data yang digunakan tersedia pada Lampiran A. Kolom target dalam tabel tersebut merupakan subkelas katai putih DA dan DAZ. Terdapat 14.729 objek katai putih yang dapat digunakan dalam model klasifikasi Random Forest pada dataset yang tidak seimbang (imbalanced dataset). Jumlah objek DA dan DAZ tidak sama (DA lebih banyak daripada DAZ), sehingga dilakukan penyeimbangan data dengan memotong beberapa objek kelas DA dan hanya memilih objek dengan SNR tinggi. Diperoleh 3.401 objek katai putih yang dapat digunakan dalam model klasifikasi pada dataset yang seimbang (balanced dataset).

Dalam katalog LSR LAMOST DR8, terdapat beberapa parameter fisis seperti temperatur efektif dan gravitasi permukaan. Dalam penelitian tugas akhir ini, dilakukan pula pengujian model Random Forest untuk data kontinu seperti temperatur efektif dan gravitasi permukaan pada katai putih. Setelah melalui proses pengolahan data, ditemukan 9.601 objek katai putih dengan nilai temperatur efektif dan 9.332 objek katai putih dengan nilai gravitasi permukaan. Sebagian data yang digunakan tersedia pada Lampiran A.

III.6.2 Data Kedua

Data kedua merupakan data yang berisi spektrum objek katai putih yang terdapat dalam Gaia DR3. Skema untuk memperoleh data kedua dapat dilihat pada Gambar III.2. Dilakukan pencocokan data antara katalog katai putih LRS LAMOST DR8 dengan Gaia DR3. Dalam proses ini, Gaia TAP+ digunakan untuk mendapatkan spektrum yang terkandung

dalam katalog katai putih LRS LAMOST DR8. Setelah memperoleh spektrum dari Gaia DR3, dilakukan pra-pemrosesan data (*data pre-processing*). Pada kasus klasifikasi, pra-pemrosesan data berupa pemilihan subkelas katai putih DA dan DAZ. Selain itu, dilakukan pra-pemrosesan agar tabel yang dihasilkan sesuai dengan parameter masukan dalam model klasifikasi Random Forest. Sebagai hasilnya, ditemukan 1.081 objek katai putih yang akan digunakan dalam model klasifikasi Random Forest (dataset Klasifikasi).



Gambar III. 2. Skema mendapatkan data Gaia DR3

Skema untuk mendapatkan parameter fisis bintang katai putih seperti temperatur efektif, dan gravitasi permukaan dapat dilihat pada Gambar III.2. Setelah data objek diperoleh, nilai parameter Gaia DR3 diperoleh menggunakan SQL Query oleh Gaia@AIP. Selanjutnya, dilakukan pembersihan dan pra-pemrosesan data parameter fisis yang telah diperoleh agar tabel yang dihasilkan sesuai dengan data masukan

pada model hutan acak. Ditemukan 1232 objek katai putih yang memiliki nilai temperatur efektif, gravitasi permukaan, dan kelimpahan logam sebagai dataset dalam model regresi.

BAB IV

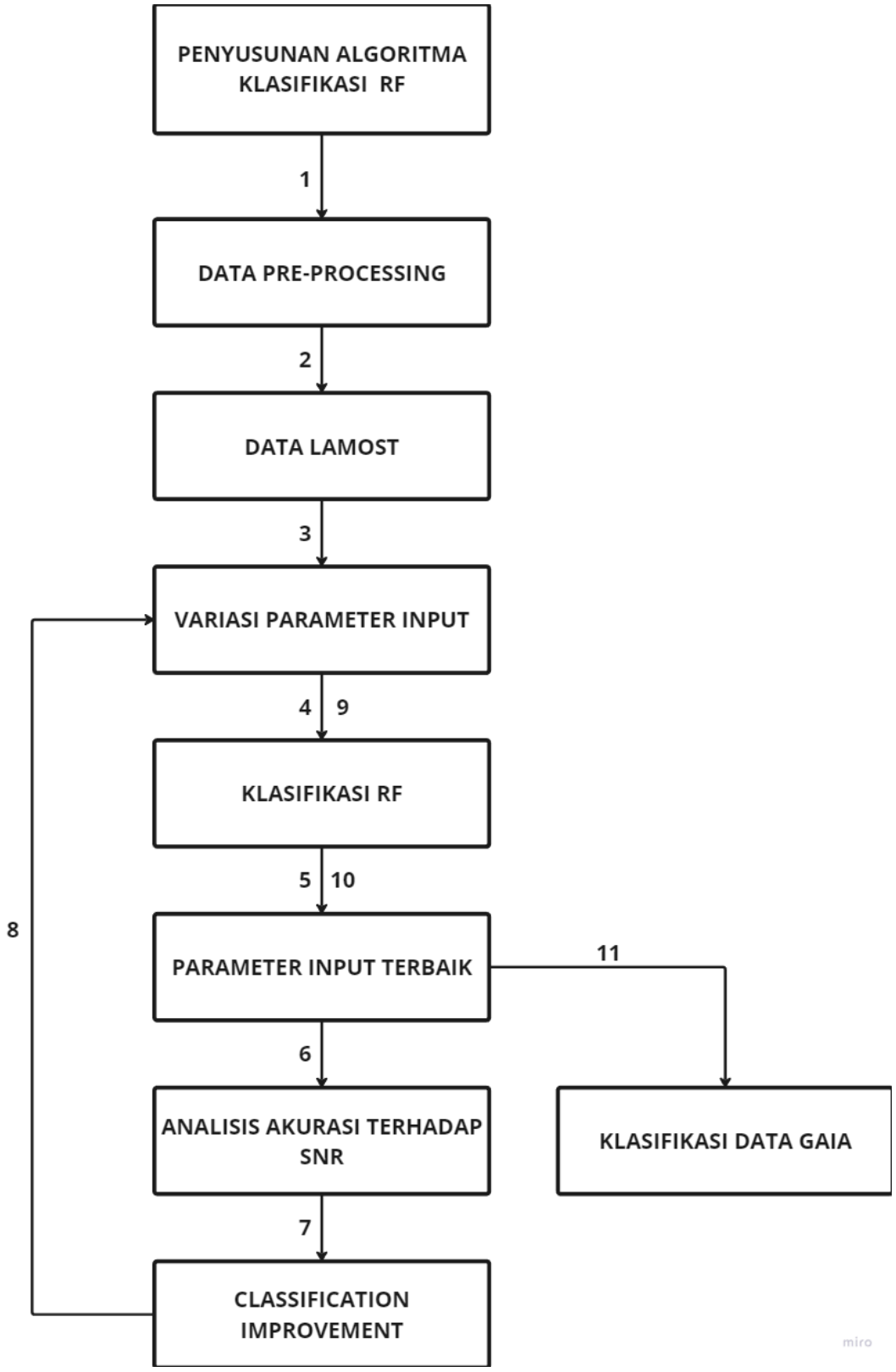
HASIL DAN ANALISIS

IV.1 Random Forest *Classification*

Pengerjaan Tugas Akhir ini akan mengikuti alur pada Gambar IV.1. Hal pertama yang dilakukan adalah membuat model klasifikasi Random Forest dengan modul scikit-learn. Kemudian akan dilakukan pra-pemrosesan data, termasuk pemotongan panjang gelombang. Data yang pertama diuji pada model adalah data LAMOST. Dilakukan variasi parameter input terhadap model untuk mendapatkan akurasi terbaik, dan parameter input tersebut selanjutnya akan digunakan dalam model. Selanjutnya, akan dilakukan analisis mengenai akurasi model terhadap kualitas spektrum (SNR) dengan membagi data menjadi tiga sampel, yaitu SNR rendah, SNR menengah, dan SNR tinggi. Setelah dilakukan analisis, kemudian dilakukan perbaikan klasifikasi (*classification improvement*) dengan menyeimbangkan variabel target (menyeimbangkan distribusi data DA & DAZ). Ketika model menunjukkan peningkatan akurasi, maka model siap diaplikasikan pada data Gaia yang memiliki kualitas lebih rendah (resolusi spektrum & SNR). Hasil akurasi keseluruhan data dapat dilihat pada Tabel IV.1.

Tabel IV. 1. Hasil akurasi model klasifikasi Random Forest

	Data LAMOST					**Data Gaia**
	Imbalanced	SNR < 10	SNR < 50	SNR < 100	Balanced	
Akurasi (AUC)	0.8245	0.7885	0.7892	0.8485	0.9334	0.8224



Gambar IV. 1. Skema alur klasifikasi data LAMOST DR8 dan Gaia DR3

IV.1.1 Akurasi Model Secara Umum

Model klasifikasi Random Forest yang dibangun akan diuji pada data LAMOST terlebih dahulu dengan parameter sampel katai putih yang dapat dilihat pada Tabel IV.2.

Tabel IV. 2. Parameter sampel katai putih LAMOST DR8

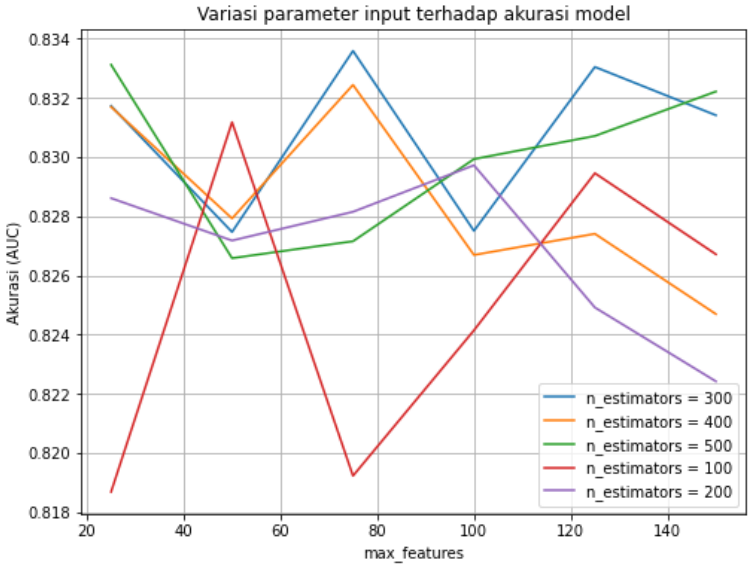
	DA	**DAZ**
#Jumlah	13278	1449
Teff (kK)	4.024 – 83.271	6.121 – 73.701
Log g	6.75 – 9.748	6.751 – 9.749
SNR	0.08 – 99.792	0.46 – 97.322

Untuk mencapai tingkat akurasi yang tinggi, dilakukan variasi parameter awal terhadap model yang digunakan. Sampel yang digunakan merupakan data pertama dari katalog LAMOST DR8. Hasil dari variasi parameter awal tersebut ditunjukkan pada Gambar IV.2., yang menggambarkan akurasi model terhadap parameter `max_features` pada berbagai nilai `n_estimators`. Nilai parameter `max_features` yang digunakan adalah 25, 50, 75, 100, 125, dan 150, sedangkan nilai parameter `n_estimators` yang digunakan adalah 100, 200, 300, 400, dan 500. Hasil variasi parameter terhadap akurasi yang dihasilkan dapat dilihat pada Gambar IV.2. dan Tabel IV.2.

Dalam variasi `n_estimators`, terlihat bahwa akurasi memiliki nilai terendah saat `n_estimators` = 100 (warna merah), kemudian meningkat pada `n_estimators` = 200 (warna ungu), mencapai puncaknya pada `n_estimators` = 300 (warna biru). Namun, akurasi kemudian menurun saat `n_estimators` = 400 (warna oranye) dan semakin rendah pada `n_estimators` = 500 (warna hijau). Pada `n_estimators` = 300 (akurasi tertinggi), terdapat nilai maksimal akurasi saat parameter `max_features` = 75. Oleh karena itu, dapat disimpulkan bahwa model Random Forest mencapai akurasi tertinggi ketika `n_estimators` bernilai 300 dan `max_features` bernilai 75. Oleh karena itu, kedua nilai parameter awal ini akan digunakan untuk

melakukan klasifikasi dan analisis pada sampel data 1 dari LAMOST DR8.

Dalam penilaian performa model tersebut, digunakan nilai AUC sebagai pembanding dibandingkan dengan akurasi. Hal ini disebabkan karena akurasi lebih cocok diterapkan pada dataset yang seimbang (*balanced dataset*) karena sifatnya yang sederhana. Sementara itu, AUC lebih cocok digunakan pada dataset yang tidak seimbang (*imbalanced* dataset) seperti sampel katai putih yang akan digunakan dalam pengerjaan tugas akhir ini, karena mencakup sensitivitas dan spesifisitas dalam akurasi yang dihasilkan.



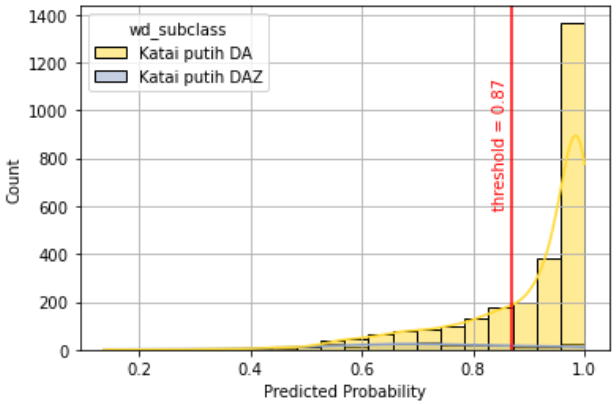
Gambar IV. 2. Variasi parameter input untuk Data LAMOST DR8. Grafik diatas menampilkan hubungan antara akurasi (sumbu y) dan parameter `max_features` (sumbu x) yang divariasikan terhadap parameter `n_estimators`.

Tabel IV. 3. Akurasi model data LAMOST DR8

Max features	`n_estimators`				
	100	200	300	400	500
25	0.818684	0.828611	0.831730	0.831697	0.833123
50	0.831180	0.827179	0.827466	0.827926	0.826583
75	0.819223	0.828151	0.833592	0.832443	0.827154

100	0.824143	0.829725	0.827506	0.826691	0.829933
125	0.829456	0.824915	0.833048	0.827407	0.830714
150	0.826716	0.822420	0.831416	0.824694	0.832217

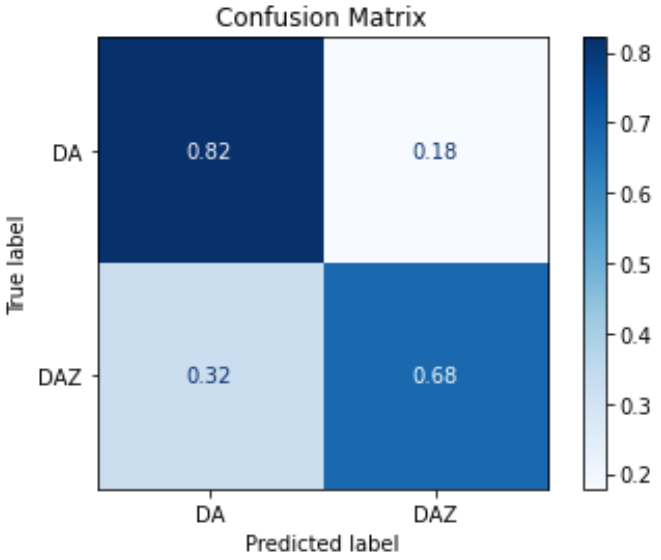
Distribusi probabilitas subkelas bintang katai putih hasil klasifikasi disajikan pada Gambar IV.3. Dari distribusi tersebut, terlihat adanya dua kelas yang berbeda dengan distribusi kelas yang hampir menyatu. Hal ini menunjukkan bahwa model kurang baik dalam mengklasifikasikan kedua kelas tersebut. Dengan memilih ambang batas (*threshold*) 0.87, maka sampel dengan nilai probabilitas di atas ambang batas tersebut akan diklasifikasikan sebagai kelas 1 (subkelas DA), sedangkan sampel dengan probabilitas di bawah ambang batas akan diklasifikasikan sebagai kelas 0 (subkelas DAZ).



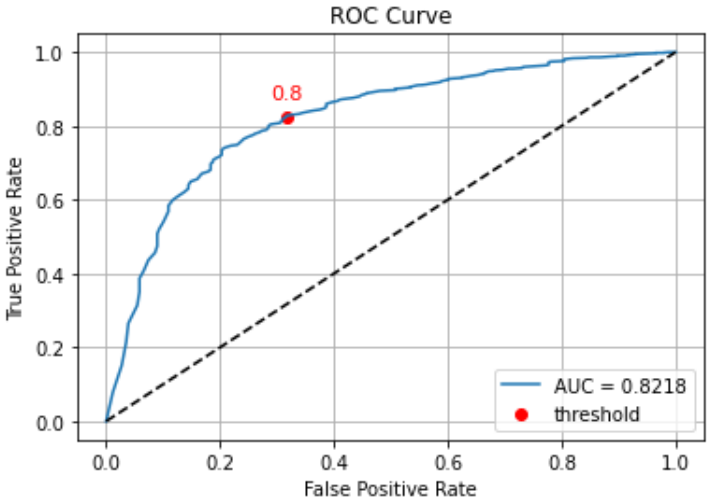
Gambar IV. 3. Ditribusi probabilitas hasil klasifikasi data LAMOST DR8

Akurasi klasifikasi pada sampel data pertama dapat dievaluasi melalui kurva ROC dan *confusion matrix* yang ditampilkan pada Gambar IV.4 dan Gambar IV.5. *Confusion matrix* yang dihasilkan telah dinormalisasi, sehingga hasil yang ditampilkan merupakan rasio pada setiap kelasnya. Batang warna (*colorbar*) di samping menunjukkan tingkat rasio tersebut. Berdasarkan *confusion matrix*, dapat disimpulkan bahwa model yang dibangun berhasil mengklasifikasikan sekitar 82% katai putih subkelas DA dengan benar, sementara hanya sekitar 18% katai putih subkelas DA yang salah diklasifikasikan sebagai DAZ. Untuk

subkelas DAZ, model yang dibangun hanya mampu mengklasifikasikan sekitar 69% katai putih dengan benar, sementara 32% diklasifikasikan sebagai DA. Hasil klasifikasi ini juga didukung oleh kurva ROC yang menunjukkan nilai TPR terhadap FPR pada setiap ambang batas dalam distribusi probabilitas kelas target. Kurva ROC yang dihasilkan memiliki AUC (luas area di bawah kurva ROC) sekitar 0,8218 (82,18%), yang menunjukkan kinerja algoritma secara umum dalam mengklasifikasikan seluruh sampel subkelas katai putih.



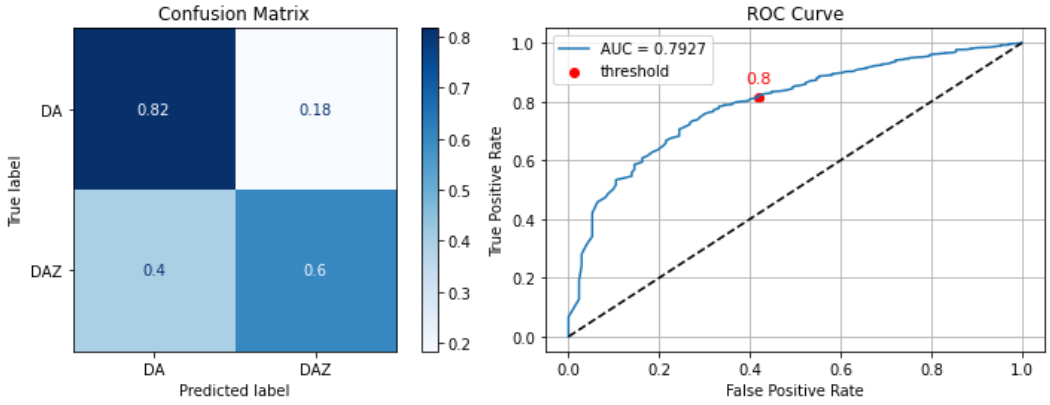
Gambar IV. 4. Confusion Matrix data LAMOST DR8



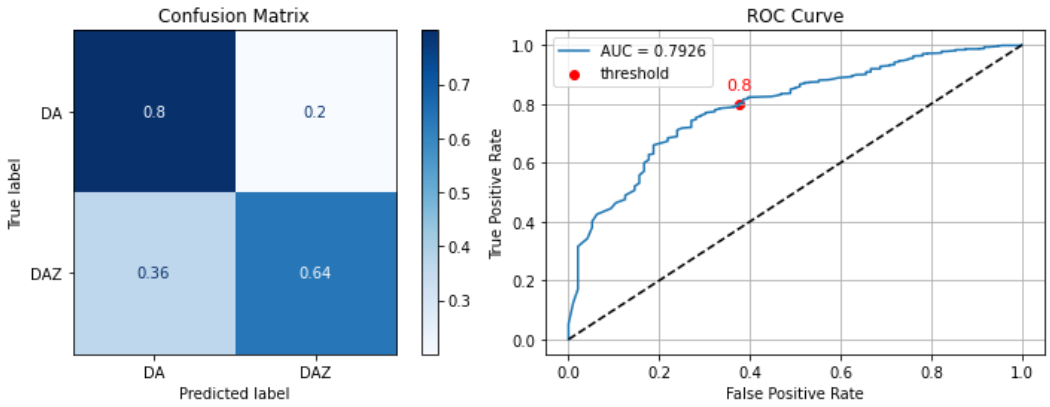
Gambar IV. 5. Kurva ROC data LAMOST DR8

Secara umum, akurasi model yang dibangun terhadap data pertama sudah cukup baik. Namun, terdapat perbedaan signifikan dalam akurasi antara subkelas katai putih DA dan DAZ. Akurasi untuk subkelas katai putih DA jauh lebih baik daripada subkelas katai putih DAZ. Hal ini mungkin disebabkan oleh ketidakseimbangan dalam sampel data katai putih antara subkelas DA dan DAZ, yang dikenal sebagai dataset tidak seimbang (*imbalanced dataset*). Sebagai akibatnya, hasil akurasi akan lebih baik pada kelas yang dominan, yaitu subkelas katai putih DA.

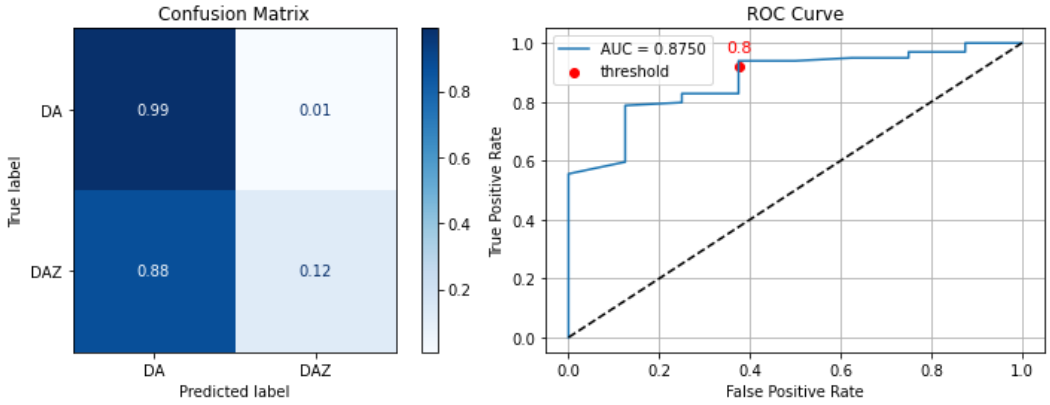
IV.1.2 Akurasi Model Berdasarkan SNR



Gambar IV. 6. Confusion Matrix (kiri) dan ROC Curve (kanan) untuk SNR rendah



Gambar IV. 7. Confusion Matrix (kiri) dan ROC Curve (kanan) untuk SNR menengah

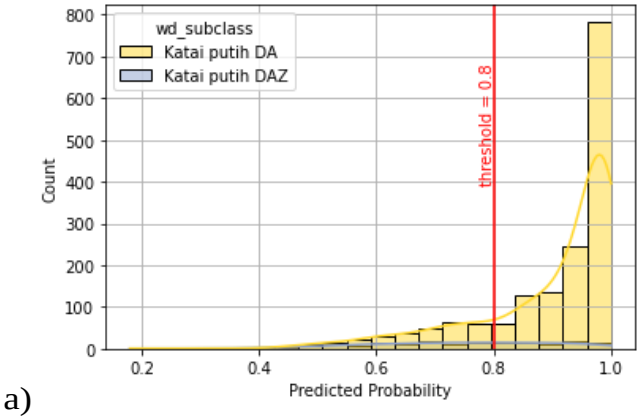


Gambar IV. 8. Confusion Matrix (kiri) dan ROC Curve (kanan) untuk SNR tinggi

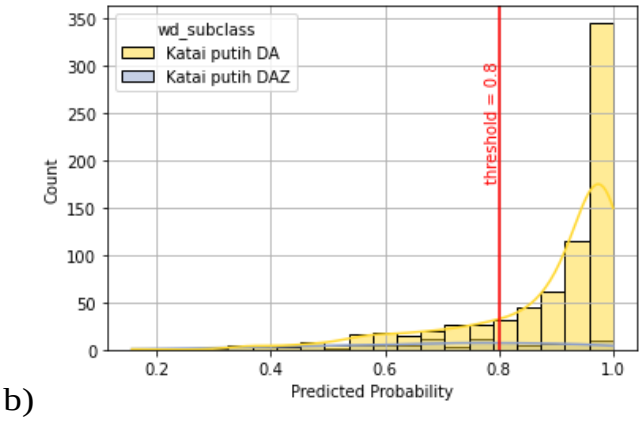
Untuk menguji performa model yang telah dibangun terhadap nilai SNR pada setiap sampel katai putih, data pertama dari katalog LAMOST DR8 akan dibagi menjadi tiga sampel berdasarkan SNR, yaitu:

a) SNR rendah (SNR $\leq$ 10). Terdapat 9.660 katai putih, dengan 8.630 termasuk dalam subkelas katai putih DA dan 1.030 termasuk dalam subkelas katai putih DAZ.
b) SNR sedang (10 < SNR $\leq$ 50). Terdapat 4.436 katai putih, dengan 4.053 termasuk dalam subkelas katai putih DA dan 383 termasuk dalam subkelas katai putih DAZ.
c) SNR tinggi (50 < SNR $\leq$ 100). Terdapat 550 katai putih, dengan 525 termasuk dalam subkelas katai putih DA dan 25 termasuk dalam subkelas katai putih DAZ.

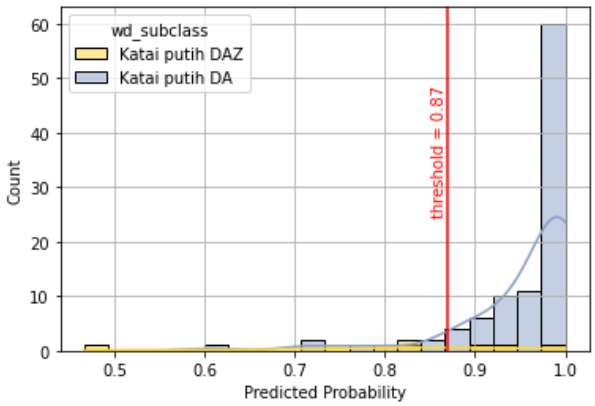
Distribusi probabilitas hasil klasifikasi subkelas katai putih ditampilkan dalam Gambar IV.9. Dari distribusi tersebut, terlihat adanya dua kelas yang berbeda dengan distribusi kelas yang hampir menyatu. Hal ini menunjukkan bahwa model kurang baik dalam mengklasifikasikan kedua kelas tersebut. Dengan memilih ambang batas 0,8 untuk setiap sampel, maka sampel dengan nilai probabilitas di atas ambang batas tersebut akan diklasifikasikan sebagai kelas 1 (subkelas DA), sedangkan sampel dengan probabilitas di bawah ambang batas akan diklasifikasikan sebagai kelas 0 (subkelas DAZ). Secara umum, model yang dihasilkan dari ketiga sampel (SNR rendah, menengah, dan tinggi) menunjukkan bahwa model Random Forest kurang mampu dalam memisahkan kedua kelas katai putih.



a)



b)



c)

Gambar IV. 9. Distribusi probabilitas hasil klasifikasi sampel SNR rendah (a), sedang (b), dan tinggi (c)

Hasil dari ketiga sampel berdasarkan SNR tersebut dievaluasi melalui confusion matrix dan kurva ROC. Kurva ROC dan confusion matrix pertama (Gambar IV.6.) menunjukkan hasil dari model klasifikasi Random Forest pada sampel dengan SNR rendah. Sekitar 82% sampel subkelas katai putih DA diklasifikasikan dengan benar, sementara hanya

18% yang diklasifikasikan sebagai DAZ. Untuk subkelas DAZ, model klasifikasi Random Forest hanya mampu mengklasifikasikan sekitar 60% sampel dengan benar, sedangkan 40% diklasifikasikan sebagai DA. Secara keseluruhan, model klasifikasi Random Forest memiliki akurasi sekitar 78% dalam mengklasifikasikan data pertama pada kasus SNR rendah.

Kurva ROC dan confusion matrix kedua (Gambar IV.7.) menunjukkan performa model pada sampel kedua dengan SNR menengah. Dari confusion matrix tersebut, dapat dilihat bahwa model mampu mengklasifikasikan 80% sampel subkelas katai putih DA dengan benar, sedangkan 20% diklasifikasikan sebagai DAZ. Untuk subkelas DAZ, model klasifikasi Random Forest hanya mampu mengklasifikasikan 64% sampel dengan benar, sedangkan 36% diklasifikasikan sebagai subkelas DA.

Kurva ROC dan confusion matrix terakhir (Gambar IV.8.) menunjukkan performa model yang diterapkan pada sampel dengan SNR yang paling tinggi. Hasilnya menunjukkan bahwa model klasifikasi Random Forest mampu mengklasifikasikan 99% sampel subkelas katai putih DA dengan benar, sementara hanya 1% yang diklasifikasikan sebagai DAZ. Namun, pada subkelas katai putih DAZ, model klasifikasi hanya mampu mengklasifikasikan 12% sampel dengan benar, sedangkan 88% diklasifikasikan sebagai DA.

Secara umum, berdasarkan kurva ROC pada ketiga kasus dengan SNR yang berbeda, dapat dilihat bahwa akurasi model klasifikasi Random Forest meningkat seiring dengan meningkatnya nilai SNR dari sampel katai putih. Selain itu, dari hasil ketiga sampel juga dapat dilihat bahwa semakin besar rasio sampel DA dibanding DAZ, maka klasifikasi untuk kelas DAZ juga akan semakin buruk. Hal ini dapat diamati dari nilai klasifikasi untuk subkelas DAZ pada SNR tinggi yang lebih buruk dibandingkan dengan SNR menengah atau SNR rendah dengan rasio DA dan DAZ yang lebih rendah.

IV.1.3 Classification Improvement

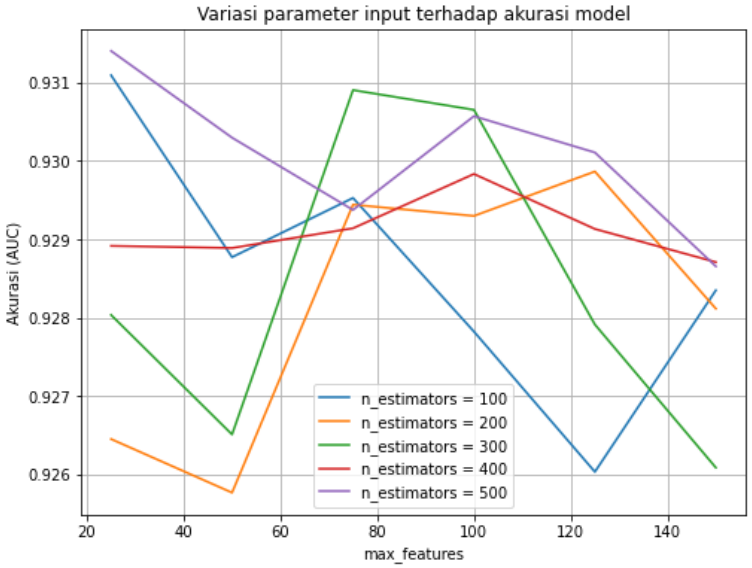
Untuk memperoleh model terbaik, dilakukan penyeimbangan distribusi target subkelas dalam dataset dengan mengambil sampel katai putih subkelas DA yang memiliki SNR tinggi. Hal ini dilakukan untuk menyamakan jumlah rasio sampel katai putih DA dan DAZ, sehingga model yang dihasilkan akan semakin baik. Sampel subkelas katai putih DA yang digunakan adalah sampel dengan nilai SNR di atas 20. Dengan demikian, diperoleh 3388 sampel katai putih, terdiri dari 1950 sampel katai putih DA dan 1438 sampel katai putih DAZ. Parameter data yang dipilih dapat dilihat pada Tabel IV.4.

Tabel IV. 4. Parameter sampel *balanced dataset*

	DA	**DAZ**
#Jumlah	13278	1449
Teff (kK)	4.024 – 83.271	6.121 – 73.701
Log g	6.75 – 9.748	6.751 – 9.749
SNR	0.08 – 99.792	0.46 – 97.322

Model klasifikasi Random Forest kemudian diuji terhadap parameter `max_features` dan `n_estimators` untuk mendapatkan akurasi terbaik. Hasil variasi parameter untuk dataset yang seimbang ditunjukkan oleh Gambar IV.10. dan Tabel IV.5. Grafik IV.2 menampilkan hubungan antara akurasi (sumbu y) dan parameter `max_features` (sumbu x). Pada grafik tersebut, terdapat perubahan variasi parameter `n_estimators` yang ditunjukkan oleh garis. Dalam variasi `n_estimators`, terlihat bahwa akurasi mencapai puncaknya pada `n_estimators` = 500 (warna ungu). Pada `n_estimators` = 500 (akurasi tertinggi), terdapat nilai maksimal akurasi saat parameter `max_features` = 75. Oleh karena itu, dapat disimpulkan bahwa model Random Forest mencapai akurasi tertinggi ketika `n_estimators` bernilai 300 dan `max_features` bernilai 25. Oleh karena itu, kedua nilai parameter awal ini akan digunakan untuk melakukan klasifikasi dan analisis pada sampel data 1 dari LAMOST DR8. Dari gambar tersebut, dapat disimpulkan bahwa model klasifikasi

Random Forest memiliki nilai terbaik pada 25 nilai `max_features` dan 500 nilai `n_estimators`.



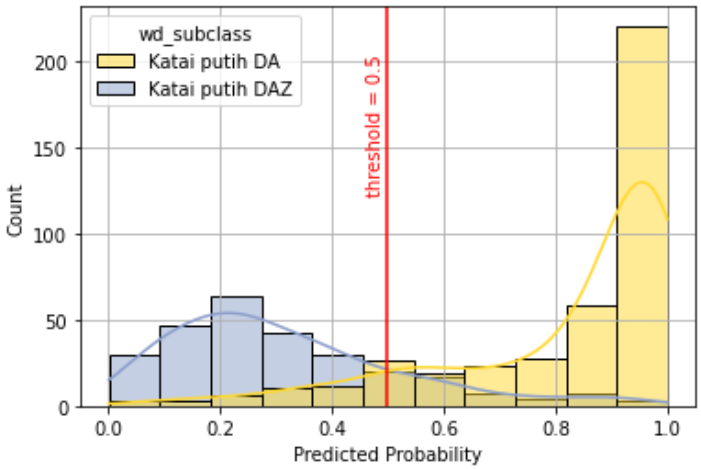
Gambar IV. 10. Variasi parameter balanced dataset. Grafik diatas menampilkan hubungan antara akurasi (sumbu y) dan parameter `max_features` (sumbu x) yang divariasikan terhadap parameter `n_estimators`.

Tabel IV. 5. Akurasi model balanced dataset

Max features	`n_estimators`					
	25	50	75	100	125	150
25	0.918450	0.924300	0.929151	0.927023	0.928454	0.929723
50	0.922239	0.927644	0.926820	0.929642	0.927855	0.929138
75	0.922460	0.925142	0.929561	0.927284	0.926622	0.932031
100	0.912299	0.924606	0.926973	0.925988	0.928166	0.929664
125	0.926703	0.929592	0.929664	0.922082	0.928760	0.927090
150	0.919035	0.923445	0.929219	0.928031	0.927383	0.928139

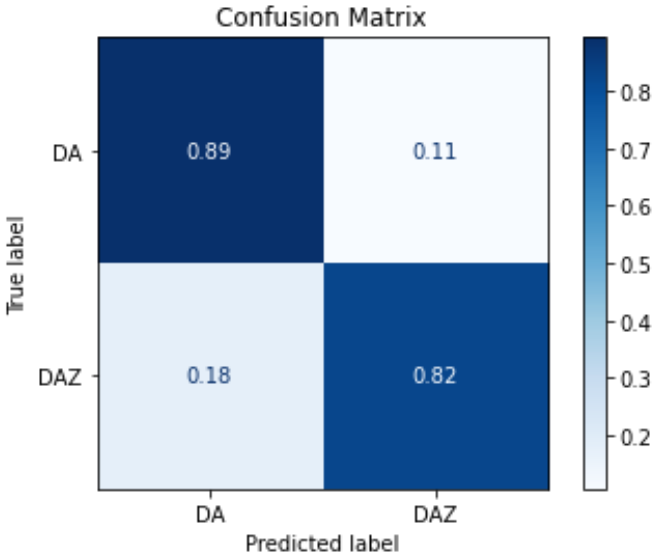
Distribusi probabilitas target subkelas katai putih hasil klasifikasi disajikan pada Gambar IV.11. Dari distribusi tersebut dapat terlihat dua kelas yang berbeda, yaitu 0 yang merupakan subkelas DAZ dan 1 yang merupakan subkelas DA. Dengan memilih threshold 0.5, sampel dengan nilai probabilitas lebih dari threshold tersebut akan diklasifikasikan sebagai kelas 1 (subkelas DA), sedangkan sampel dengan probabilitas

kurang dari threshold akan diklasifikasikan sebagai kelas 0 (subkelas DAZ).

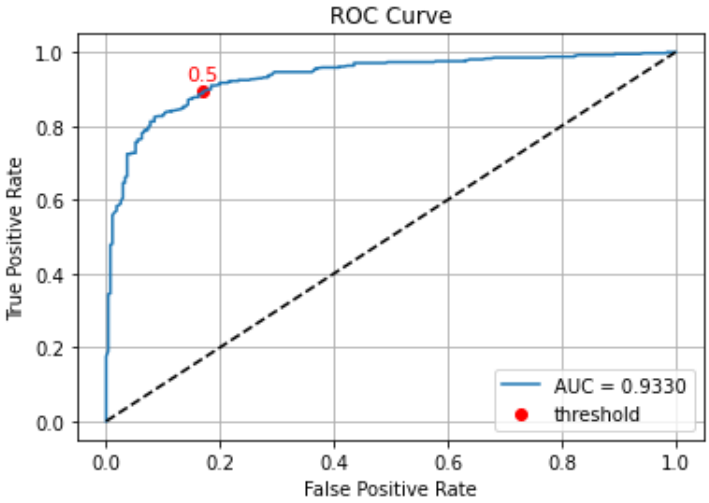


Gambar IV. 11. Distribusi probabilitas hasil klasifikasi *balanced dataset*

Hasil klasifikasi menggunakan model dengan parameter yang telah ditentukan sebelumnya dapat ditemukan melalui confusion matrix dan kurva ROC pada Gambar IV.12. dan Gambar IV.13. Berdasarkan confusion matrix, dapat disimpulkan bahwa model yang telah dibangun mampu mengklasifikasikan 89% sampel katai putih subkelas DA dengan benar dan hanya 11% yang diklasifikasikan sebagai DAZ. Untuk sampel katai putih subkelas DAZ, model yang dibangun berhasil mengklasifikasikan 82% dengan benar, sementara 18% dari sampel tersebut diklasifikasikan sebagai subkelas DA. Secara keseluruhan, hasil klasifikasi menunjukkan tingkat akurasi sekitar 93.34%, sebagaimana yang terlihat pada kurva ROC.



Gambar IV. 12. Confusion matrix *balanced dataset*



Gambar IV. 13. Kurva ROC *balanced dataset*

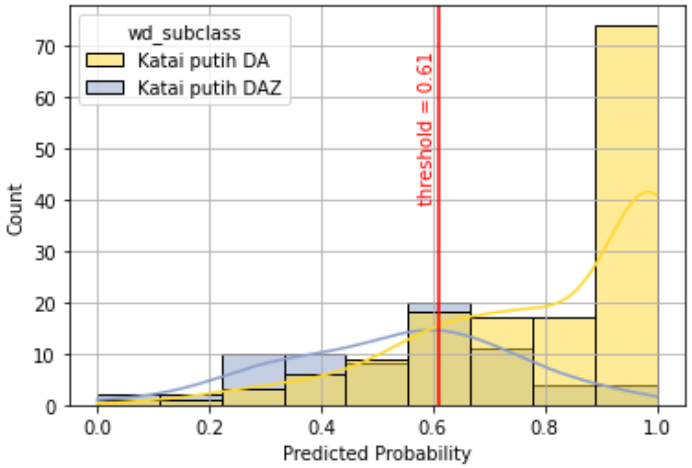
IV.1.4 Akurasi Model Klasifikasi Pada Data Gaia DR3

Setelah model diterapkan pada data LAMOST dan menghasilkan akurasi yang sangat baik, langkah selanjutnya adalah menguji model pada data Gaia DR3 dengan parameter data yang tercantum pada Tabel IV.6.

Tabel IV. 6. Parameter sampel katai putih Gaia DR3

	DA	**DAZ**
#Jumlah	758	323
Teff (kK)	6.722 – 71.430	8.670 – 73.701
Log g	6.75 – 9.725	6.754 - 9.685
SNR	3.97 – 125.61	13.09 - 116.23

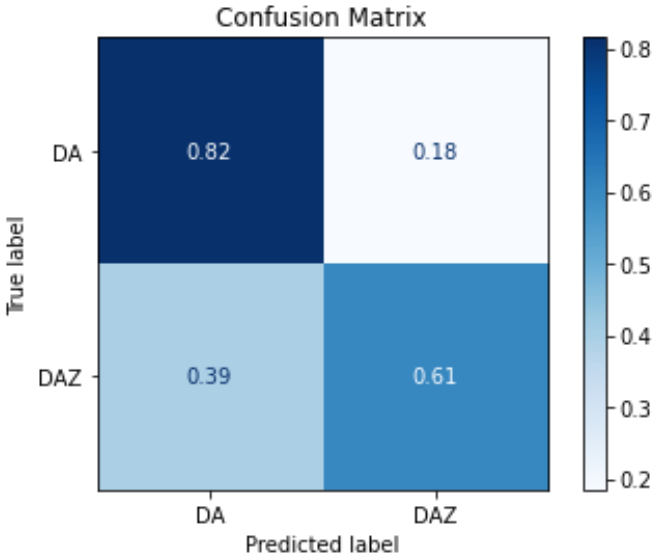
Distribusi probabilitas target subkelas katai putih hasil klasifikasi ditampilkan pada Gambar IV.14. Dari distribusi tersebut terlihat adanya dua kelas yang berbeda dengan distribusi kelasnya hampir menyatu. Hal ini menunjukkan bahwa model kurang efektif dalam mengklasifikasikan kedua kelas tersebut. Dalam hal ini, threshold 0,61 dipilih untuk memisahkan klasifikasi. Sampel dengan nilai probabilitas di atas threshold akan diklasifikasikan sebagai kelas 1 (subkelas DA), sedangkan sampel dengan probabilitas di bawah threshold akan diklasifikasikan sebagai kelas 0 (subkelas DAZ).



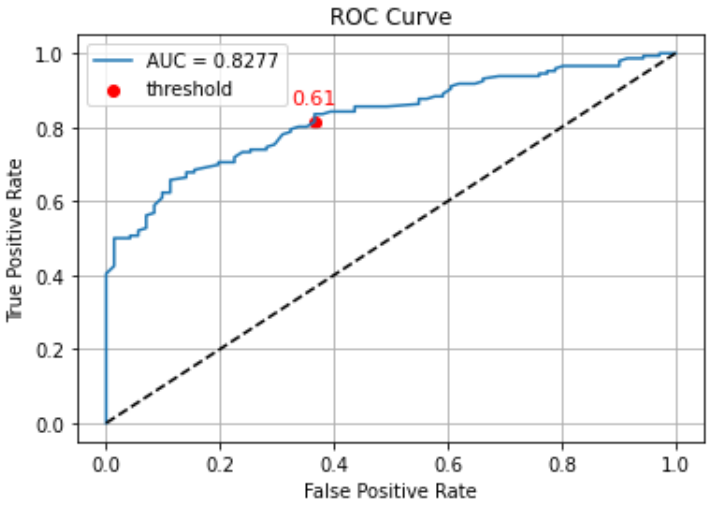
Gambar IV. 14. Distribusi probabilitas hasil klasifikasi data Gaia DR3

Hasil klasifikasi pada data Gaia DR3 dapat dilihat melalui confusion matrix dan kurva ROC pada Gambar IV.15 dan Gambar IV.16. Dari kurva ROC tersebut, dapat diketahui bahwa secara keseluruhan model berhasil mengklasifikasikan subkelas katai putih dengan akurasi sebesar 82.24%. Untuk melihat akurasi masing-masing subkelas, dapat diperiksa melalui confusion matrix pada Gambar IV.15. Dengan menggunakan threshold 0,5, diperoleh hasil bahwa model

berhasil mengklasifikasikan 82% sampel katai putih subkelas DA dengan benar dan 18% sampel katai putih subkelas DA yang salah diklasifikasikan sebagai DAZ. Sedangkan untuk subkelas DAZ, model hanya mampu mengklasifikasikan 61% sampel katai putih DAZ dengan benar dan 39% diklasifikasikan sebagai DA.



Gambar IV. 15. Confusion Matrix data Gaia DR3



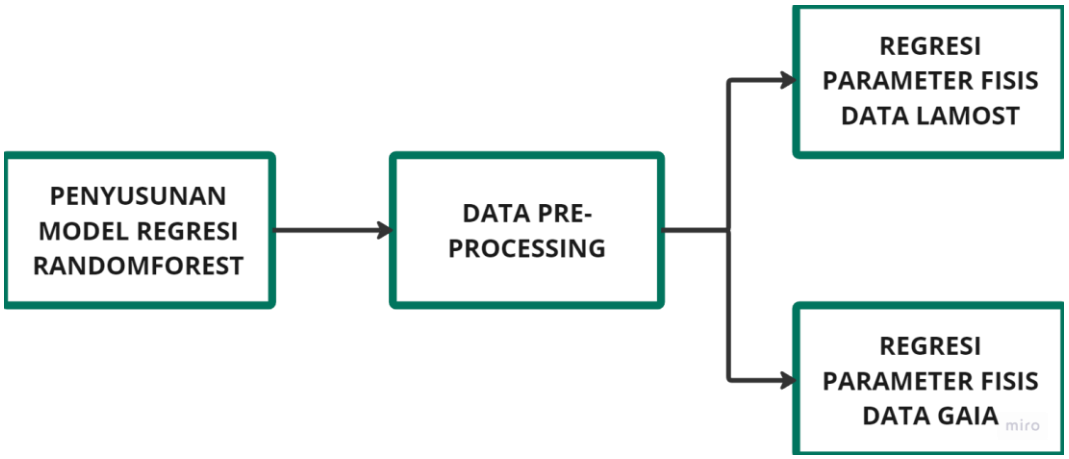
Gambar IV. 16. Kurva ROC data Gaia DR3

Nilai akurasi yang diperoleh dengan menggunakan data Gaia lebih rendah dibandingkan dengan data LAMOST. Jika dilihat dari parameter

sampel, terlihat bahwa penurunan akurasi pada model data Gaia disebabkan oleh nilai *resolving power* (R) dan SNR yang lebih rendah dibandingkan data LAMOST. Hal ini menyebabkan jumlah fitur (panjang gelombang) pada model data Gaia menjadi lebih sedikit dibandingkan dengan data LAMOST.

IV.2 Random Forest *Regression*

Dalam bagian regresi, model regresi yang telah dibuat akan diuji menggunakan parameter fisis seperti temperatur efektif dan gravitasi permukaan objek. Skema proses dapat ditinjau pada Gambar IV.17. Data pertama yang akan digunakan untuk pengujian model regresi adalah data LAMOST. Selanjutnya, model yang telah dibuat juga akan diujikan menggunakan parameter fisis dari data Gaia DR3.



Gambar IV. 17. Skema alur regresi data LAMOST DR8 dan Gaia DR3

Model regresi menggunakan metode Random Forest akan diuji pada parameter fisis sampel katai putih. Metode Random Forest regresi akan digunakan untuk memperoleh estimasi parameter temperatur efektif dan kerapatan permukaan bintang berdasarkan analisis spektrumnya. Untuk mengevaluasi keakuratan model yang telah dikembangkan, akan digunakan MAE (*Mean Absolute Error*) sebagai metrik yang mengukur rerata error antara nilai yang diprediksi oleh

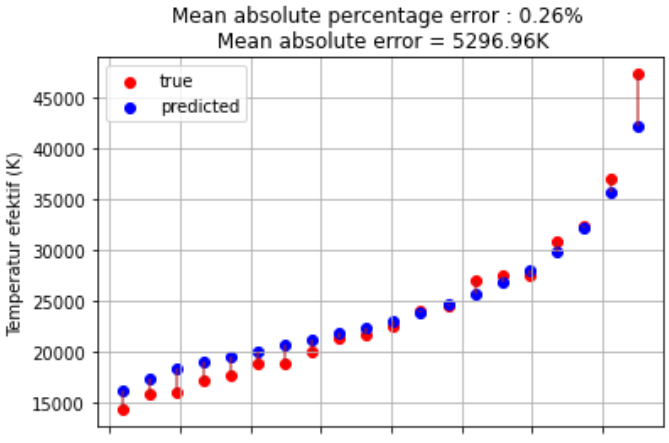
model dengan nilai sebenarnya. Penggunaan MAE dibanding MSE (Mean Squared Error) dikarenakan nilai MAE memiliki skala dimensi yang sama dengan variabel target dan lebih tahan terhadap outlier daripada MSE. Persamaan untuk MAE yaitu

$$MAE = \frac{\sum_{j=1}^{n} |y_{pred} - y_{act}|}{n} \qquad \text{(IV. 1)}$$

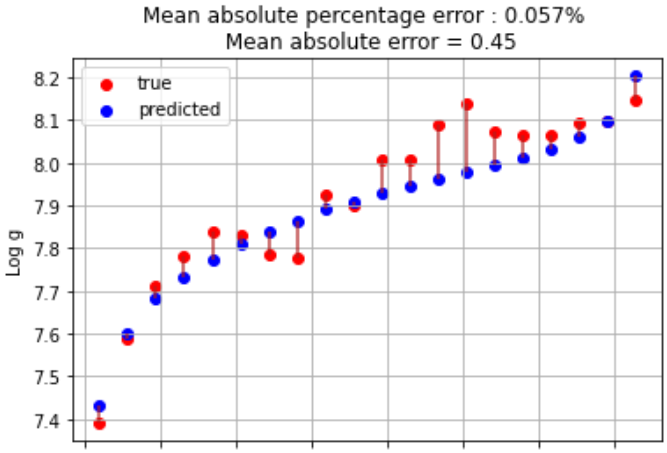
dengan n merupakan jumlah sampel.

Model regresi menggunakan metode Random Forest memanfaatkan dua parameter input, yaitu `n_estimators` (yang menentukan jumlah pohon yang akan dibangun dalam model) dan `max_depth` (yang menentukan kedalaman setiap pohon), keduanya diatur dengan nilai 100. Untuk mengevaluasi kualitas model yang dibangun, digunakan *loss-function* yang mengukur sejauh mana hasil prediksi sesuai dengan data target. Gambar *loss-function* untuk kedua parameter fisis tersebut dapat ditinjau pada Gambar IV.18.

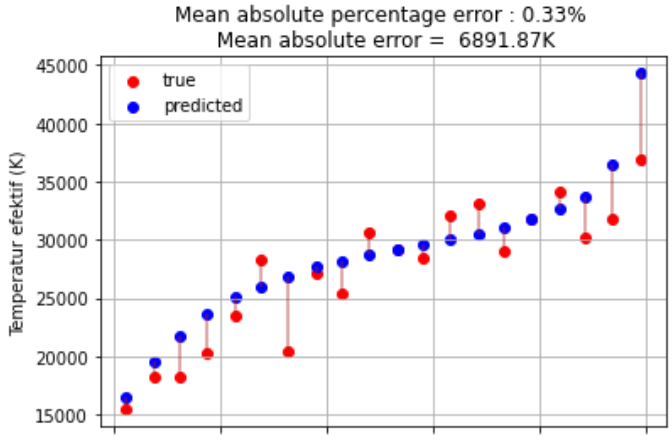
a)



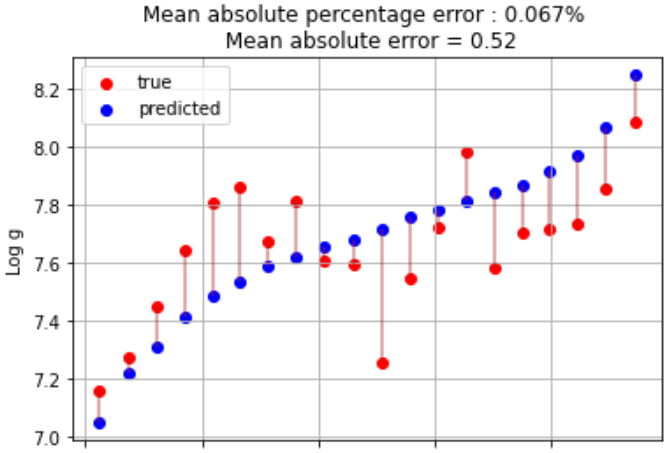
b)



c)



d)



Gambar IV. 18. *Loss-function* untuk kedua parameter fisis. a) b) merupakan loss-function data LAMOST DR8. c) d) merupakan loss-function data Gaia DR3. Titik merah merupakan nilai target yang sebenarnya dan biru merupakan nilai hasil prediksi model.

Semakin kecil nilai *loss-function* (perbedaan antara data target dan prediksi model), semakin tinggi akurasi model. Dari hasil perhitungan *loss-function,* terlihat bahwa *loss-function* untuk data LAMOST dan data Gaia sangat kecil, dengan rata-rata *loss-function* data Gaia lebih besar daripada data LAMOST. Hal ini menunjukkan bahwa akurasi model cukup baik. Hasil dari model regresi terhadap parameter fisis katai putih dapat ditemukan dalam Tabel IV.7. untuk data LAMOST dan Tabel IV.8. untuk data Gaia.

Tabel IV. 7. Hasil regresi data LAMOST DR8

Parameter	**MAE (Model Random Forest)**	**Error**
Temperatur Efektif (K)	2699.22 (0.13%)	623.58 (0.027%)
Gravitasi Permukaan	0.226 (0.028%)	0.118 (0.014%)

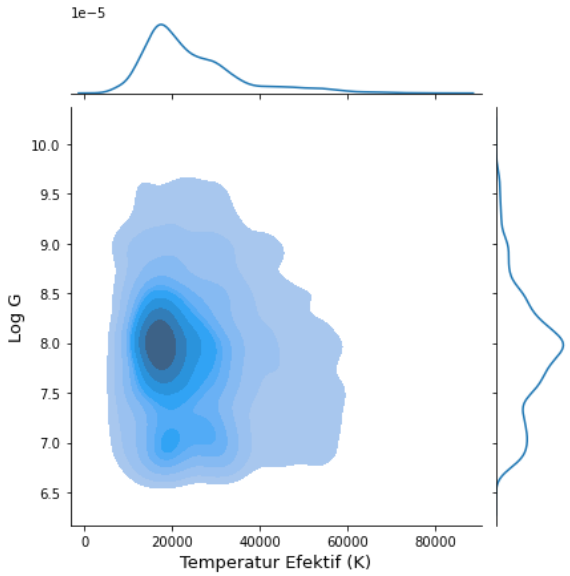
Tabel IV. 8. Hasil regresi data Gaia DR3

Parameter	**MAE (Model Random Forest)**	**Error**
Temperatur Efektif (K)	3802.68 (0.17%)	618.58 (0.026%)
Gravitasi Permukaan	0.286 (0.036%)	0.111 (0.014%)

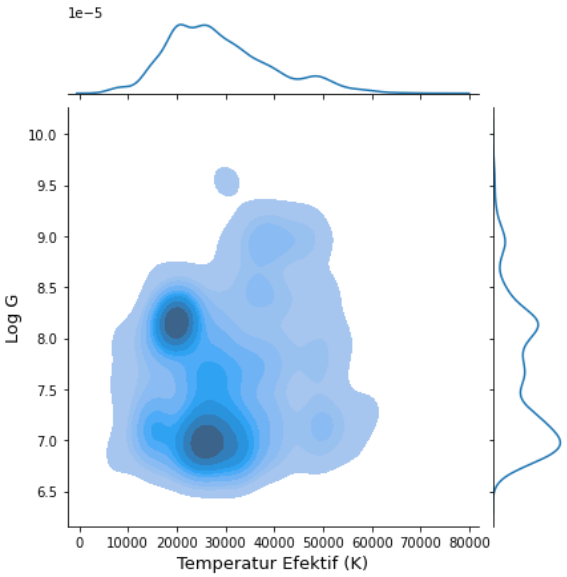
Tabel IV. 9. Perbedaan data LAMOST DR8 dan Gaia DR3

	LAMOST DR8	**Gaia DR3**
Resolusi	$\sim 1800 R_\lambda$	$\sim 20 - 70 R_\lambda$
Rerata SNR	~ 11	~ 50
Jumlah sampel	9332	1266
Jumlah fitur	3663	346

Secara umum, model menunjukkan tingkat akurasi yang lebih tinggi pada data LAMOST daripada data Gaia. Perbedaan ini dapat diatribusikan kepada beberapa faktor, seperti jumlah sampel, resolusi spektrum, dan keseragaman data. Perbedaan data LAMOST dan Gaia dapat dilihat pada Tabel IV.9. Jumlah sampel katai putih dan fitur berupa panjang gelombang yang tersedia dalam data LAMOST lebih banyak daripada data Gaia, sehingga memberikan tingkat akurasi yang lebih baik. Selain itu, resolusi spektrum yang diberikan oleh LAMOST juga lebih tinggi, memungkinkan model untuk menangkap pola yang lebih detail dalam data dan menghasilkan akurasi yang lebih tinggi. Ketika meninjau distribusi data, parameter fisis dalam data LAMOST menyebar lebih merata dalam rentang nilai yang luas, sehingga model lebih mudah menangkap pola dari data tersebut. Distribusi persebaran parameter fisis antara data LAMOST dan Gaia dapat dilihat pada Gambar IV.9 dan Gambar IV.20.



Gambar IV. 19. Distribusi temperatur efektif dan gravitasi permukaan LAMOST DR8

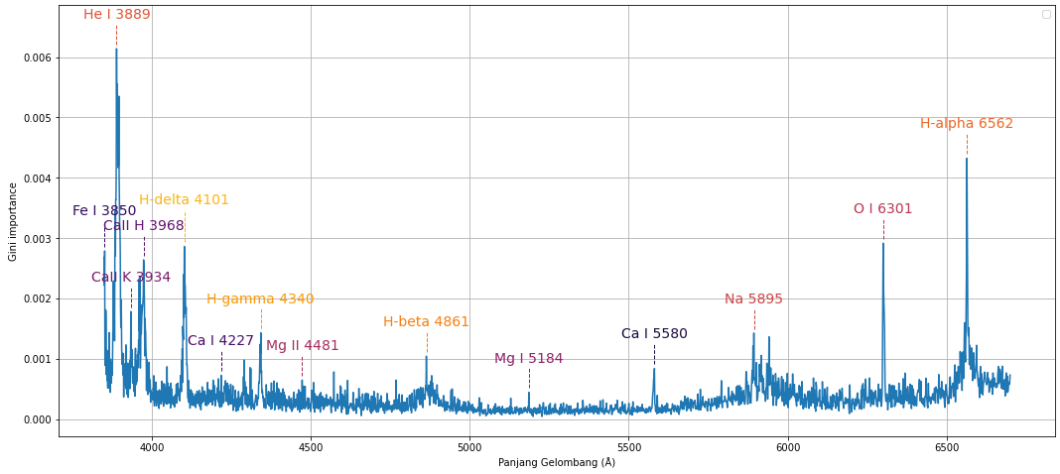


Gambar IV. 20. Distribusi temperatur efektif dan gravitasi permukaan Gaia DR3

Dari distribusi data parameter fisis dalam data Gaia pada Gambar IV.20, terlihat bahwa terdapat hubungan antara akurasi model dengan sebaran data. Akurasi model regresi untuk parameter fisis temperatur efektif lebih rendah dibandingkan dengan gravitasi permukaan. Hal ini disebabkan oleh fakta bahwa sebaran data temperatur efektif cenderung terpusat membentuk pola distribusi Gaussian tunggal. Sementara itu, sebaran data gravitasi permukaan lebih merata dan membentuk beberapa pola distribusi Gaussian.

IV.3 ANALISIS PANJANG GELOMBANG

Untuk menentukan fitur atau panjang gelombang yang memiliki pengaruh yang signifikan terhadap klasifikasi, digunakan metode feature importance yang disediakan oleh model Random Forest. Feature importance tersebut menghasilkan indeks gini untuk setiap fitur, yang mengindikasikan seberapa besar pengaruhnya terhadap klasifikasi model. Semakin tinggi indeks gini suatu fitur, semakin besar pengaruhnya dalam melakukan klasifikasi. Setelah proses klasifikasi selesai, dilakukan analisis terhadap fitur atau panjang gelombang dari data spektrum katai putih. Hasil analisis ini akan disajikan dalam Gambar IV.21.



Gambar IV. 21. Plot indeks Gini terhadap panjang gelombang

Dari informasi yang tertera pada Gambar IV.21., terlihat bahwa sebagian besar garis hidrogen Balmer memiliki tingkat indeks Gini yang lebih tinggi dibandingkan dengan garis lainnya. Garis hidrogen Balmer ini terdiri dari garis Hα (λ 6562Å), Hβ (λ 4861Å), Hγ (λ 4340Å), dan Hδ (λ 4101Å). Penemuan ini dapat dijelaskan melalui analisis spektrum subkelas katai putih DA dan DAZ. Garis serapan hidrogen Balmer pada subkelas DA cenderung memiliki intensitas yang lebih tajam dibandingkan dengan garis serapan hidrogen Balmer pada subkelas katai putih DAZ. Oleh karena itu, algoritma Random Forest akan menggunakan panjang gelombang garis hidrogen Balmer sebagai fitur

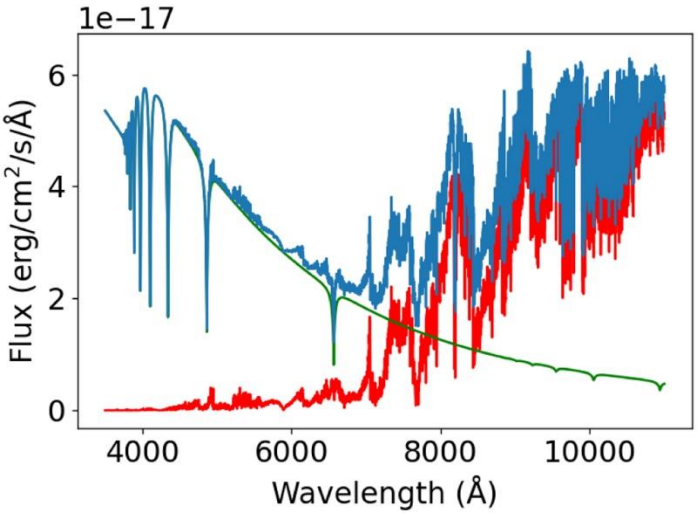
untuk mengklasifikasikan kedua jenis spektrum tersebut.

Selain itu, terdapat beberapa garis lain yang menjadi ciri khas bagi katai putih DAZ dan tidak dimiliki oleh subkelas katai putih DA. Garis-garis ini disebabkan oleh kehadiran atom-atom logam, seperti kalsium (Ca), magnesium (Mg), besi (Fe), dan lainnya. Contohnya adalah garis Ca II H (λ 3970Å) dan garis Ca II K (λ 3934Å) yang terdapat dalam sampel ini. Keberadaan garis-garis tersebut disebabkan oleh perbedaan dalam mekanisme pembentukan antara kedua subkelas katai putih tersebut. Faktor-faktor seperti sifat kimia bintang, temperatur, komposisi atmosfer, dan evolusi mempengaruhi munculnya garis-garis tersebut pada subkelas katai putih DAZ. Oleh karena itu, algoritma Random Forest menggunakan dua fitur garis tersebut untuk membedakan dan mengklasifikasikan keduanya.

Selain garis-garis tersebut, terdapat juga beberapa garis logam lainnya yang memiliki nilai indeks Gini yang tinggi, seperti garis Ca I (λ 4227Å, 5580Å), He I (λ 3889Å), O I (λ 6301Å), Mg I (λ 5184Å), Mg II (λ 4481Å), dan Fe I (λ 3850Å). Garis-garis ini akan muncul pada subkelas katai putih DAZ dan tidak terdapat pada subkelas katai putih DA. Sama seperti sebelumnya, perbedaan dalam proses pembentukan subkelas katai putih DAZ yang berbeda dari subkelas katai putih DA menjadi penyebab munculnya garis-garis logam tersebut.

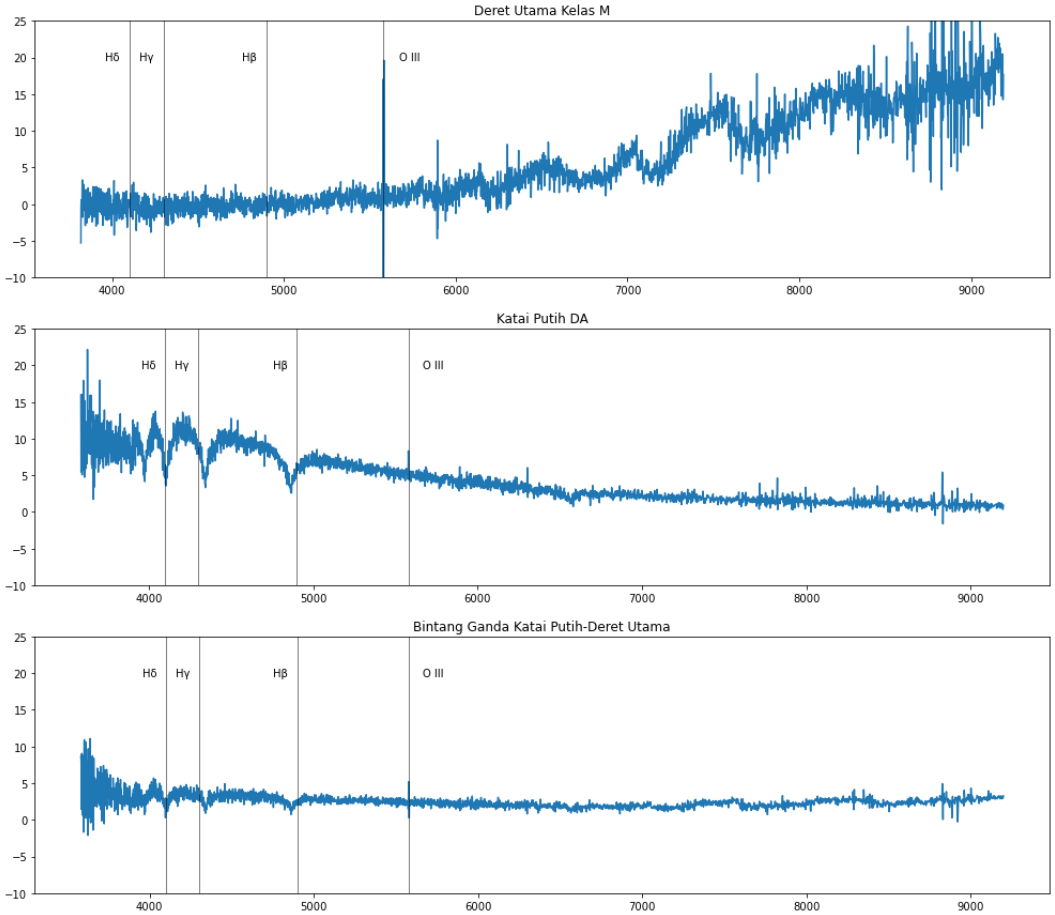
IV.4 Perbandingan dengan Literatur

Klasifikasi dengan menggunakan metode Random Forest melalui spektrum bintang juga telah dilakukan oleh Echeverry dkk. (2022) yang mengklasifikasikan bintang deret utama kelas M, katai putih, dan bintang ganda deret utama - katai putih melalui fitur spektrum dan garis elemennya. Hasil pekerjaan tersebut akan dibandingkan dengan klasifikasi yang telah dilakukan pada pengerjaan Tugas Akhir ini. Perbedaan spektrum objek klasifikasi dari kedua studi ditunjukkan pada Gambar IV.22.



Gambar IV. 22. Contoh spektrum bintang ganda katai putih-deret utama (garis biru) yang diperoleh dari spektrum katai putih (garis hijau) dan bintang deret utama (garis merah). altahu : Echeverry dkk. (2022).

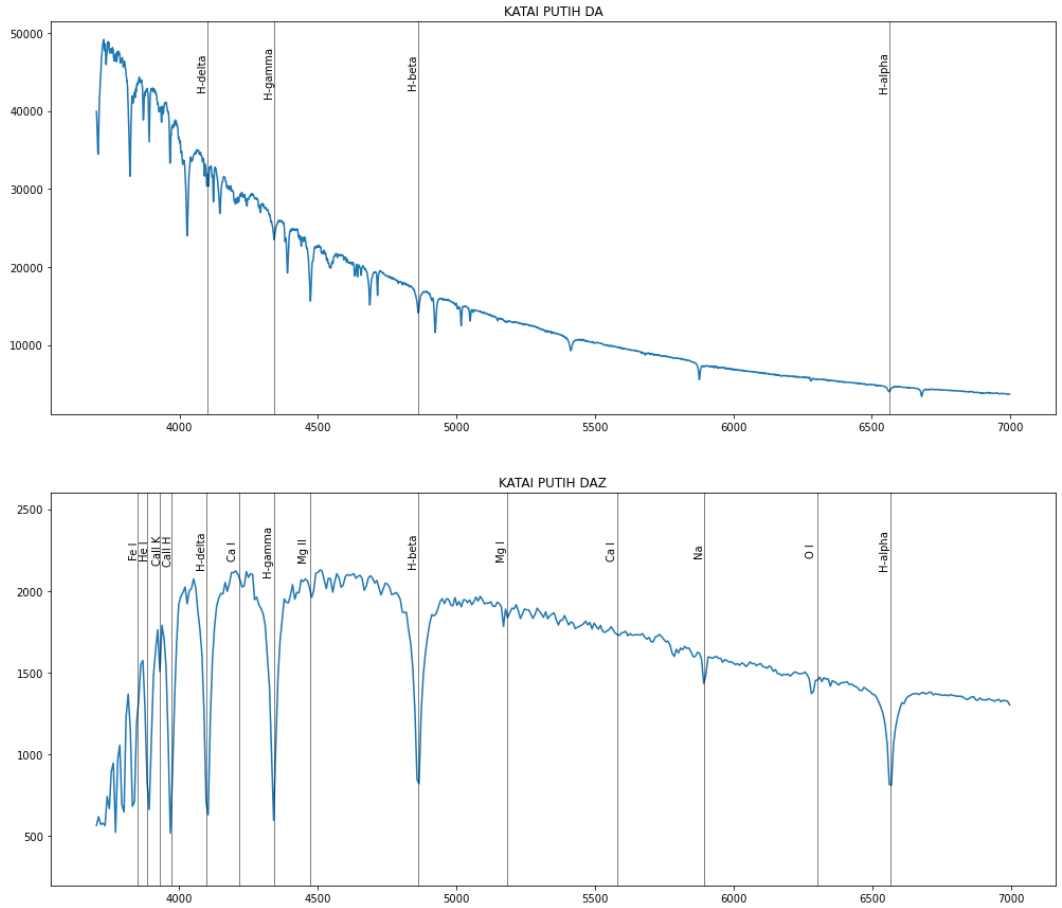
Objek klasifikasi yang digunakan dalam studi Echeverry dkk. (2022) dapat dilihat pada Gambar IV.22. Dari Gambar IV.22. terlihat bahwa metode hutan acak (Random Forest) akan mengklasifikasikan spektrum ketiga objek tersebut berdasarkan nilai fluks atau kecerlangannya. Katai putih akan menghasilkan fluks yang dominan pada panjang gelombang merah, sementara bintang deret utama akan menghasilkan fluks yang dominan pada panjang gelombang biru. Bintang ganda katai putih-deret utama akan menghasilkan fluks yang dominan pada kedua panjang gelombang (merah dan biru). Analisis lebih lanjut mengenai garis-garis elemen pada ketiga objek dapat ditemukan dalam Gambar IV.22. Dari gambar tersebut dapat diketahui bahwa katai putih dan bintang ganda deret utama akan memiliki fitur garis hidrogen Balmer yang tidak ditemukan dalam spektrum bintang deret utama, yaitu garis Balmer Hδ, Hγ, dan Hβ yang terletak pada '4 100, '4 300, dan '4 900 Å. Sementara itu, garis oksigen terionisasi dua kali pada panjang gelombang sekitar 5500 Å (OIII) akan muncul dalam spektrum bintang deret utama, tetapi tidak pada spektrum katai putih dan bintang ganda katai putih-deret utama. Garis-garis hidrogen Balmer dan oksigen terionisasi dua kali tersebut akan membantu metode Random Forest dalam melakukan klasifikasi ketiga objek tersebut.



Gambar IV. 23. Garis elemen pada ketiga objek klasifikasi. (a) Bintang deret utama. (b) Bintang katai putih. (c) Bintang ganda katai putih-deret utama.

Sementara itu, dalam pengerjaan Tugas Akhir ini, metode hutan acak (Random Forest) akan melakukan klasifikasi dari dua objek, yaitu subkelas katai putih DA dan subkelas katai putih DAZ. Perbedaan spektrum kedua objek tersebut dapat dilihat dalam Gambar IV.23. Dari gambar tersebut terlihat perbedaan secara umum dari kecerlangan kedua spektrum. Objek subkelas katai putih DA menghasilkan kecerlangan yang lebih dominan pada panjang gelombang merah, kemudian secara perlahan menurun saat mendekati panjang gelombang biru. Di sisi lain, subkelas katai putih DAZ menghasilkan fluks yang dominan hampir pada seluruh panjang gelombang. Analisis lebih lanjut menunjukkan adanya perbedaan dalam elemen-elemen garis yang muncul sebagai akibat dari perbedaan proses pembentukan kedua subkelas katai putih tersebut. Subkelas katai putih DA memiliki fitur elemen garis hidrogen Balmer yang dominan, sedangkan subkelas katai putih DAZ memiliki elemen garis hidrogen Balmer ditambah beberapa elemen berat seperti kalsium dan magnesium. Oleh karena itu, metode Random Forest akan

melakukan klasifikasi kedua objek tersebut melalui fitur elemen yang mencirikan masing-masing objek.



Gambar IV. 24. Garis elemen pada kedua objek klasifikasi. (a) Bintang katai putih subkelas DA. (b) Bintang katai putih subkelas DAZ.

Studi yang dilakukan oleh Echeverry dkk. (2022) menggunakan data SDSS sebelum menggunakan data Gaia DR3. Dalam hal ini, hasil klasifikasi menggunakan data SDSS akan dibandingkan dengan hasil klasifikasi menggunakan data LAMOST yang dilakukan pada pengerjaan Tugas Akhir ini. Terdapat beberapa perbedaan hasil yang diperoleh yang dapat dilihat dalam Tabel IV.9.

Tabel IV. 10. Perbedaan hasil klasifikasi data LAMOST dan SDSS.

Perbedaan	RF	Echeverry dkk.
Data	Large Sky Area Multi-Object Fibre Spectroscopic Telescope	Sloan Digital Sky Survey (SDSS)

	(LAMOST)	
Objek Klasifikasi	Sub-kelas katai putih	Bintang kelas spektral M, katai putih, dan bintang ganda katai putih-deret utama
Jumlah sampel	3401 (1950 katai putih subkelas DA dan 1451 katai putih subkelas DAZ)	6923 (2340 deret utama, 2031 katai putih, dan 2551 bintang ganda deret utama-katai putih
Jumlah fitur (panjang gelombang)	2410 (3847 – 6700 Å)	3760 (3850 – 9150 Å)
Rerata SNR	25 (0.046 – 100)	11 (0.45 – 100)
Resolusi	~1800	1800 - 2200
Parameter yang digunakan	`n_estimators` = 500 `max_features` = 25	`n_estimators` = 200 `max_features` = 500
Akurasi	0.93	0.96

Dari Tabel IV.9., dapat disimpulkan bahwa akurasi yang dihasilkan dari kedua data tersebut hampir sama. Akurasi yang hampir sama tersebut diperoleh dengan memilih parameter input untuk metode hutan acak (Random Forest) yang disesuaikan dengan data yang ditinjau. Hasil ini juga mengindikasikan bahwa metode Random Forest dapat diaplikasikan pada survei lain dengan memperhatikan properti data yang digunakan dan parameter input metode Random Forest.

Setelah menggunakan data LAMOST, hasil klasifikasi spektrum Gaia DR3 pada pengerjaan Tugas Akhir ini juga akan dibandingkan dengan studi yang dilakukan oleh Echeverry dkk. (2022). Studi tersebut mengklasifikasikan bintang kelas M, katai putih, dan bintang ganda deret utama-katai putih menggunakan metode hutan acak (Random Forest) berdasarkan spektrumnya. Terdapat perbedaan dalam pengerjaan ini yang dicatat dalam Tabel IV.10.

Tabel IV. 11. Perbedaan hasil klasifikasi spektrum Gaia DR3 kalibrasi internal dan eksternal.

Perbedaan	RF	Echeverry dkk.

Data	Gaia DR3 (*External Calibration*)	Gaia DR3
Objek Klasifikasi	Sub-kelas katai putih	Bintang kelas spektral M, katai putih, dan bintang ganda katai putih-deret utama
Jumlah sampel	1081 (758 katai putih subkelas DA dan 323 katai putih subkelas DAZ)	6285 (5992 deret utama, 126 katai putih, dan 167 bintang ganda deret utama-katai putih
Jumlah fitur (panjang gelombang)	343 (3360 – 10200 Å)	-
Rerata SNR	18.75	BP = 278, RP = 1036
Resolusi	22 - 70	66
Parameter yang digunakan	`n_estimators` = 500 `max_features` = 25	`n_estimators` = 200 `max_features` = 500
Akurasi	0.82	0.92

Akurasi hasil penelitian Echeverry dkk. (2022) lebih tinggi dibandingkan dengan akurasi yang diperoleh dalam penelitian Tugas Akhir ini. Selain perbedaan dalam objek klasifikasi yang diamati, perbedaan akurasi tersebut juga bisa dipengaruhi oleh faktor-faktor lain. Berdasarkan data dalam Tabel IV.10., terlihat bahwa perbedaan akurasi tersebut mungkin disebabkan oleh resolusi dan SNR dari spektrum kalibrasi eksternal Gaia DR3 yang lebih rendah daripada spektrum kalibrasi internal Gaia DR3, sehingga mengakibatkan penurunan akurasi model Random Forest. Selain itu, perbedaan akurasi juga dapat disebabkan oleh jumlah sampel yang lebih sedikit. Model Random Forest umumnya memberikan akurasi yang lebih rendah ketika digunakan dengan jumlah sampel yang terbatas.

BAB V

SIMPULAN DAN SARAN

V.1 Simpulan

Pada penelitian Tugas Akhir ini telah dilakukan klasifikasi dan regresi pada sampel katai putih Gaia DR3 dan LAMOST DR8. Bagian klasifikasi dilakukan dengan memisahkan subkelas katai putih DA dengan subkelas katai putih DAZ. Sedangkan bagian regresi telah dilakukan penentuan nilai parameter fisis katai putih, yaitu temperatur efektif dan gravitasi permukaan. Berdasarkan hasil dan analisis yang telah dilakukan, diperoleh beberapa simpulan yaitu :

1. Random Forest merupakan salah satu metode klasifikasi dan regresi yang baik digunakan pada data yang memiliki noise tinggi serta fitur yang hampir tak terhingga. Dalam klasifikasi, perfoma Random Forest meningkat ketika jumlah kelas dalam variable target seimbang. Sedangkan dalam regresi, akurasi Random Forest meningkat ketika distribusi variable target membentuk distribusi gaussian yang lebih dari satu.
2. Random Forest berhasil mengklasifikasikan 14727 objek katai putih (14036 DA dan 1772 DAZ) menggunakan berbagai fitur (dalam hal ini panjang gelombang) yang memiliki pengaruh signifikan terhadap proses klasifikasi subkelas katai putih. Panjang gelombang tersebut bersesuaian dengan garis hidrogen Balmer dan beberapa garis spektrum yang disebabkan oleh atom-atom logam, seperti Ca II H dan Ca II K yang mencirikan atmosfir subkelas katai putih DAZ.
3. Random Forest berhasil menentukan parameter fisis temperature efektif 9610 objek katai putih dan gravitasi permukaan 9332 objek katai putih menggunakan berbagai fitur (dalam hal ini panjang gelombang) yang memiliki pengaruh signifikan terhadap proses regresi parameter fisis katai putih.
4. Klasifikasi yang dilakukan pada sampel katai putih menghasilkan akurasi 93.34% untuk data LAMOST DR8 dan 82.24% untuk data

Gaia DR3. Hasil klasifikasi paling baik diperoleh untuk subkelas katai putih DA yang memiliki jumlah sampel lebih banyak dibanding dengan subkelas katai putih DAZ. Akurasi yang dihasilkan untuk data LAMOST lebih tinggi dibandingkan dengan data Gaia karena jumlah sampel katai putih dalam LAMOST lebih banyak dibanding dengan sampel katai putih pada Gaia.

5. Regresi yang dilakukan pada sampel katai putih menghasilkan *error* parameter temperatur efektif untuk data pengujian Gaia DR3 dan LAMOST DR8 masing masing yaitu 3802.68K (0.17%) dan 2699.22K (0.13%). Tingkat kesalahan tersebut lebih tinggi dibanding tingkat kesalahan yang dihasilkan oleh instrument, yaitu 618.58K (0.026%) dan 623.58K (0.027%). Sedangkan error pada gravitasi permukaan untuk data pengujian Gaia DR3 dan LAMOST DR8 masing masing yaitu 0.286 (0.036%) dan 0.226 (0.028%). Tingkat kesalahan tersebut lebih tinggi dibanding tingkat kesalahan yang dihasilkan oleh instrument, yaitu 0.111 (0.014%) dan 0.118 (0.014%).

V.2 Saran

Berikut beberapa saran yang perlu dipertimbangkan untuk penelitian selanjutnya:

1. Fitur panjang gelombang pada setiap objek data LAMOST DR8 memiliki nilai yang sama sehingga nilai kecerlangan atau fluxnya lebih akurat.
2. Dilakukan peninjauan ulang mengenai sampel Gaia DR3 yang diperoleh dari katalog LAMOST DR8.
3. Jumlah sampel katai putih yang digunakan untuk klasifikasi pada data Gaia DR3 lebih banyak sehingga model mampu menemukan pola dari spektrumnya.
4. Untuk perbandingan hasil pengerjaan tugas akhir terhadap literatur sebaiknya menggunakan pengerjaan dengan objek yang sama.
5. Dilakukan peninjauan ulang spektrum Gaia DR3 kalibrasi eksternal.
6. Spektrum katai putih Gaia DR3 menggunakan spektrum Gaia RVS.
7. Evaluasi hasil model tidak hanya menggunakan kurva ROC dan AUC, tetapi juga dengan menggunakan analisis Bayesian.

DAFTAR PUSTAKA

Althaus, L. G., Bertolami, M. M., Córsico, A. H., García-Berro, E., & Gil-Pons, P. (2005). The formation of DA white dwarfs with thin hydrogen envelopes. Astronomy & Astrophysics, 440(1), L1-L4.

Althaus, L. G., Córsico, A. H., Isern, J., & García-Berro, E. (2010). Evolutionary and pulsational properties of white dwarf stars. *The Astronomy and Astrophysics Review*, *18*, 471-566.

Althaus, L., Panei, J., Córsico, A., García-Berro, E., & Scóccola, C. (2005, July). Formation and Evolution of Hydrogen-Deficient post-AGB White Dwarfs. In *14th European Workshop on White Dwarfs* (Vol. 334, p. 61).

Ball, N. M., & Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, *19*(07), 1049-1106.

Barros, R. C., De Carvalho, A. C., & Freitas, A. A. (2015). *Automatic design of decision-tree induction algorithms*. Springer.

Breiman, L. (2001). Random Forests. *Machine learning*, *45*, 5-32.

Carrasco, J. M., Weiler, M., Jordi, C., Fabricius, C., De Angeli, F., Evans, D. W., ... & Montegriffo, P. (2021). Internal calibration of Gaia BP/RP low-resolution spectra. Astronomy & Astrophysics, 652, A86.

Carroll, B. W., & Ostlie, D. A. (2017). *An introduction to modern astrophysics*. Cambridge University Press.

Dufour, P., Fontaine, G., Liebert, J., Schmidt, G. D., & Behara, N. (2008). Hot DQ white dwarfs: Something different. The Astrophysical Journal, 683(2), 978.

Echeverry, D., Torres, S., Rebassa-Mansergas, A., & Ferrer-Burjachs, A. (2022). Random Forest classification of Gaia DR3 white dwarf-main sequence spectra: A feasibility study. *Astronomy & Astrophysics*, *667*, A144.

Feng, W., Sui, H., Tu, J., Huang, W., & Sun, K. (2018). A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images. *International journal of remote sensing*, *39*(22), 7998-8021.

Gänsicke, B. T., Koester, D., Girven, J., Marsh, T. R., & Steeghs, D. (2010). Two white dwarfs with oxygen-rich atmospheres. Science, 327(5962), 188-

190.

Gentile Fusillo, N. P., Tremblay, P. E., Cukanovaite, E., Vorontseva, A., Lallement, R., Hollands, M., ... & Jordan, S. (2021). A catalogue of white dwarfs in Gaia EDR3. *Monthly Notices of the Royal Astronomical Society*, *508*(3), 3877-3896.

Guo, J., Zhao, J., Zhang, H., Zhang, J., Bai, Y., Walters, N., ... & Liu, J. (2022). White dwarfs identified in LAMOST Data Release 5. Monthly Notices of the Royal Astronomical Society, 509(2), 2674-2688.

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. (2014). Statistics, data mining, and machine learning in astronomy. In *Statistics, Data Mining, and Machine Learning in Astronomy*. Princeton University Press.

Kepler, S. O., Koester, D., Pelisoli, I., Romero, A. D., & Ourique, G. (2021). White dwarf and subdwarf stars in the Sloan Digital Sky Survey Data Release 16. *Monthly Notices of the Royal Astronomical Society*, *507*(3), 4646-4660.

Kawka, A., Vennes, S., Dinnbier, F., Cibulková, H., & Németh, P. (2011, March). Abundance analysis of DAZ white dwarfs. In *AIP Conference Proceedings* (Vol. 1331, No. 1, pp. 238-245). American Institute of Physics.

Lacombe, P., Liebert, J., Wesemael, F., & Fontaine, G. (1983). G74-7-A true DA, F (DAZ) white dwarf. *The Astrophysical Journal*, *272*, 660-664.

Mitchell, T. M. (2007). Machine learning (Vol. 1). New York: McGraw-hill.

Montegriffo, P., De Angeli, F., Andrae, R., Riello, M., Pancino, E., Sanna, N., ... & Yoldas, A. (2022). Gaia Data Release 3: External calibration of BP/RP low-resolution spectroscopic data. arXiv preprint arXiv:2206.06205.

Mohapatra, N., Shreya, K., & Chinmay, A. (2020). Optimization of the random forest algorithm. In *Advances in Data Science and Management: Proceedings of ICDSM 2019* (pp. 201-208). Springer Singapore.

Salzberg, S., Chandar, R., Ford, H., Murthy, S. K., & White, R. (1995). Decision trees for automated identification of cosmic-ray hits in Hubble Space Telescope images. *Publications of the Astronomical Society of the Pacific*, *107*(709), 279.

Strobel, N. (2007). Properties of Stars: Color and Temperature. *Astronomy Notes. Primis/McGraw-Hill, Inc. Archived from the original on*, 06-26.

von Hippel, T., Kuchner, M. J., Kilic, M., Mullally, F., & Reach, W. T. (2007). The

new class of dusty DAZ white dwarfs. *The Astrophysical Journal*, *662*(1), 544.

Werner, K., Nagel, T., & Rauch, T. (2009, June). Spectral modeling of gaseous metal disks around DAZ white dwarfs. In Journal of Physics: Conference Series (Vol. 172, No. 1, p. 012054). IOP Publishing.

Xiong, H. (2009). Classification: Basic Concepts, Decision Trees, and Model Evaluation. *New Jersey: Rutgers University*.

Zuckerman, B., Koester, D., Reid, I. N., & Hünsch, M. (2003). Metal lines in DA white dwarfs. *The Astrophysical Journal*, *596*(1), 477.

https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c diakses pada 19/04/2023 20:00 WIB.

https://towardsdatascience.com/random-forest-classification-678e551462f diakses pada 19/04/2023 20:20 WIB.

https://machinelearningmastery.com/cross-entropy-for-machine-learning/ diakses pada 19/04/2023 20:40 WIB.

https://towardsdatascience.com/predict-vs-predict-proba-scikit-learn-bdc45daa5972 diakses pada 19/04/2023 21.00 WIB.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html diakses pada 09/11/2022 20:00 WIB.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html diakses pada 05/03/2023 10:00 WIB.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html diakses pada 10/12/2022 21:00 WIB.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html diakses pada 10/12/2022 22:00 WIB.

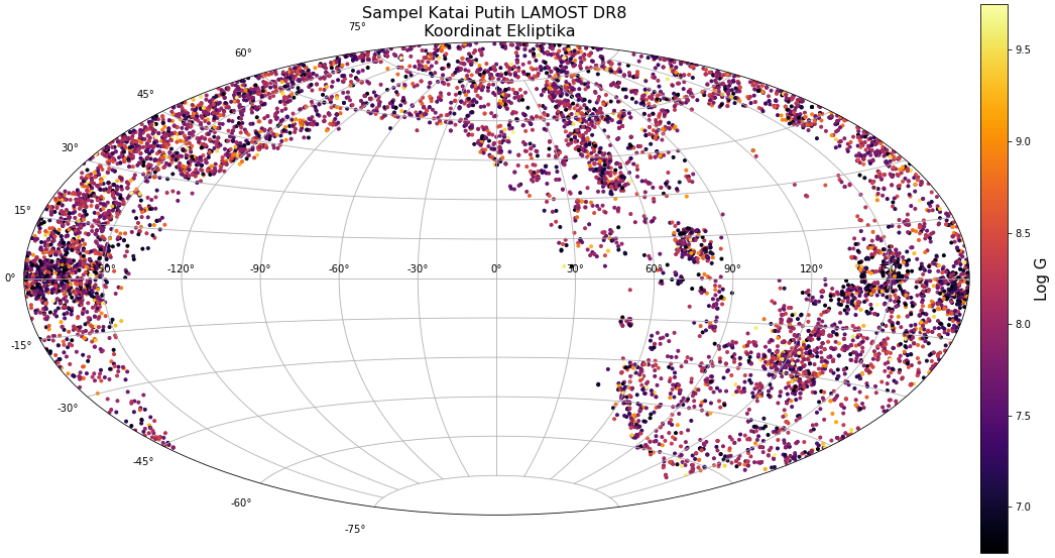
LAMPIRAN

LAMPIRAN A
PARAMETER DATA

Parameter fisis dan data bintang katai putih LAMOST dan Gaia yang digunakan dapat dilihat pada Tabel ksjd.

	DA	**DAZ**
#Jumlah	13278	1449
Teff (kK)	4.024 – 83.271	6.121 – 73.701
Log g (dex)	6.75 – 9.748	6.751 – 9.749
SNR	0.08 – 99.792	0.46 – 97.322

Sebaran objek katai putih katalog LAMOST dalam koordinat ekliptika



Gambar A. 1. Sebaran katai putih LAMOST dalam koordinat ekliptika.

Contoh tabel data Gaia DR3

Tabel A. 1. Contoh data LAMOST DR8 yang digunakan untuk klasifikasi subkelas katai putih

3847.69	3848.576	3849.463	…	6697.306	6698.848	6700.391	wd_subclass
24.72889	27.64467	24.89325	…	17.17936	17.08929	17.4439	DA
1290.424	1190.242	1224.701	…	153.5051	149.336	144.9419	DA
9819.987	9687.308	9957.726	…	1469.381	1475.331	1460.838	DA
63.67539	71.42969	54.51587	…	8.625998	10.12834	9.736727	DAZ
53.45627	51.67845	49.29684	…	4.980087	4.796206	5.550964	DAZ
1.750419	57.19115	32.45917	…	29.18537	23.25776	22.08492	DAZ

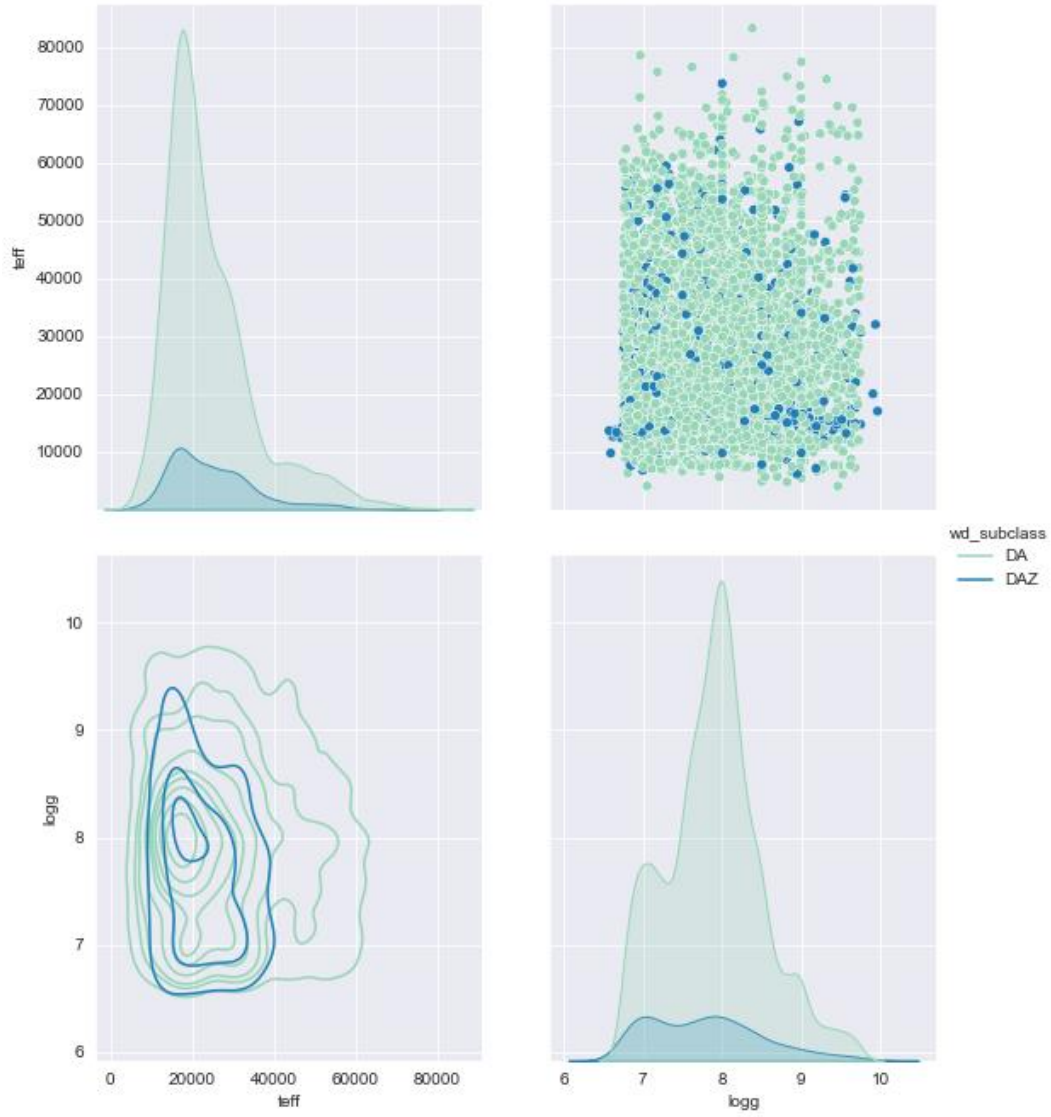
Tabel A. 2. Contoh data LAMOST DR8 yang digunakan untuk regresi gravitasi permukaan

3838.84	3839.72	3840.60	…	8910.46	8912.51	8914.56	logg
31.7974	37.1353	61.8470	…	-1.06191	8.905656	12.46156	8.263
60.4088	65.8149	42.5663	…	-3.85108	7.131863	-2.4116	7.933
92.9018	102.263	105.296	…	6.045517	7.900367	5.390872	8.063
19.4345	24.7216	11.6086	…	16.05405	20.5242	12.84386	7.017
1.62007	-20.7783	25.1175	…	4.253694	5.548764	5.837281	7.578
31.1717	35.6591	36.2266	…	1.006247	0.696943	3.159384	8

Tabel A. 3. Contoh data LAMOST DR8 yang digunakan untuk regresi temperatur efektif

3838.84	3839.72	3840.60	…	8910.46	8912.512	8914.564	teff
31.7974	37.1353	61.8470	…	-1.0619	8.905656	12.46156	24000
60.4088	65.8149	42.5663	…	-3.8510	7.131863	-2.4116	17122.64
92.9018	102.263	105.296	…	6.04551	7.900367	5.390872	14070.3
19.4345	24.7216	11.6086	…	16.0540	20.5242	12.84386	12000
1.62007	-20.7783	25.1175	…	4.25369	5.548764	5.837281	27937.78
31.1717	35.6591	36.2266	…	1.00624	0.696943	3.159384	22655.31

Distribusi kerapatan parameter fisis temperatur efektif terhadap gravitasi permukaan data LAMOST DR8



Gambar A. 2. Distribusi kerapatan parameter fisis temperatur efektif terhadap gravitasi permukaan data LAMOST DR8.

Contoh tabel data Gaia DR3

Tabel A. 4. Contoh data Gaia DR3 yang digunakan untuk klasifikasi subkelas katai putih

336	338	340	…	1016	1018	1020	wd_ subclass
470.2319	338.184	289.0806	…	455.7578	453.4698	469.6055	DAZ
301.145	194.0511	238.3259	…	316.5253	333.2371	364.1728	DA
97.02513	107.2735	123.3591	…	436.3081	430.7414	444.1011	DAZ
758.0544	694.717	624.4375	…	590.9294	599.0813	635.6432	DA
573.2198	516.8903	513.3741	…	203.5182	205.7586	220.5827	DA
257.4674	229.9346	175.1567	…	773.6465	775.5715	801.3766	DAZ

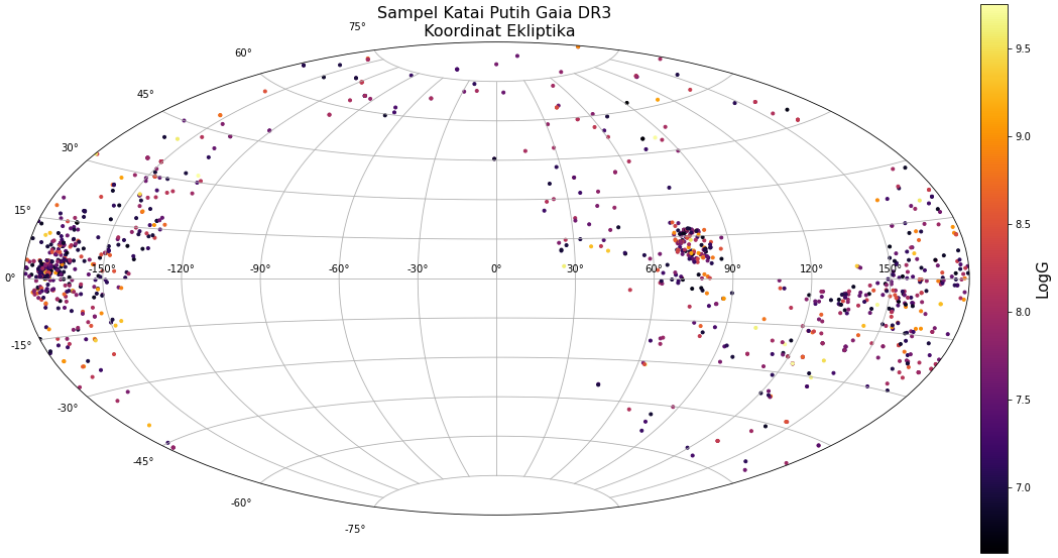
Tabel A. 5. Contoh data Gaia DR3 yang digunakan untuk regresi gravitasi permukaan

336	338	340	…	1016	1018	1020	logg
686.2544	710.1692	694.5167	…	266.5456	281.9081	314.5286	6.942
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	7.93
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	7.878
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	8
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	7.979
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	9.725

Tabel A. 6. Contoh data Gaia DR3 yang digunakan untuk regresi temperatur efektif

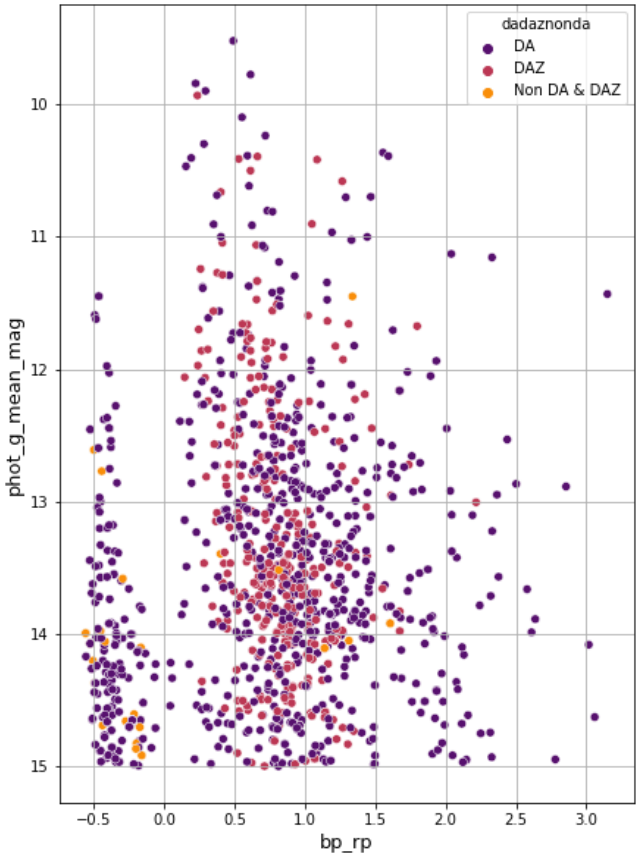
336	338	340	…	1016	1018	1020	teff
686.2544	710.1692	694.5167	…	266.5456	281.9081	314.5286	8076.02
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	28068.98
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	27816.05
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	28320.78
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	27872.4
2294.624	2373.459	2422.135	…	62.02552	62.81041	65.93398	28095.5

Sebaran objek katai putih Gaia DR3 dalam koordinat ekliptika.



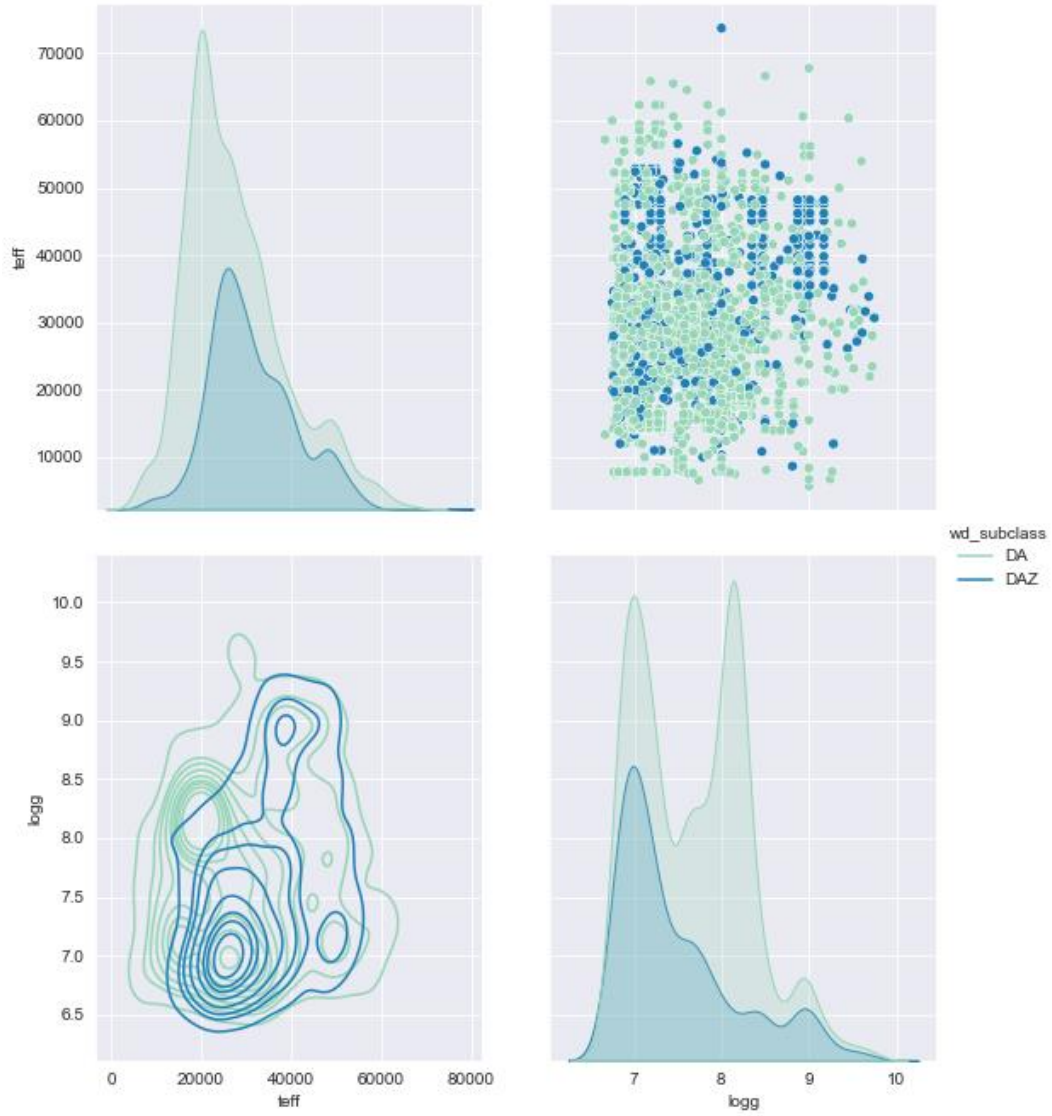
Gambar A. 3. Sebaran objek katai putih Gaia DR3 dalam koordinat ekliptika.

Diagram Hertzprung-Russel populasi katai putih Gaia DR3



Gambar A. 4. Diagram HR populasi katai putih Gaia DR3. Pada diagram HR tersbut terlihat adanya DB-Gap yang memisahkan kedua populasi katai putih.

Distribusi kerapatan parameter fisis temperatur efektif terhadap gravitasi permukaan data Gaia DR3



Gambar A. 5. Distribusi kerapatan parameter fisis temperatur efektif terhadap gravitasi permukaan data Gaia DR3.

LAMPIRAN B
TANGKAPAN LAYAR CODE PYTHON METODE DAN PENGOLAHAN DATA

PENGOLAHAN DATA

```
#Mengolah data spektrum yang diperoleh dari wget command
sc = pd.read_csv('sc')
DA = pd.DataFrame([])
for i in sc.obsid:
    data_fits = Table.read("DA/"+str(i) + ' token=F30f38a8159')
    data = []
    header = ['FLUX','IVAR','WAVELENGTH','ANDMASK','ORMASK',
              'NORMALIZATION']
    for j in range (len(header)):
        data_new = np.array((data_fits[header[j]].data)
                            .flatten())
        data.append(data_new)
    df = pd.DataFrame(data)
    df = df.T
    df.columns = header
    DA['FLUX'+str(i)] = df.FLUX
    DA['WAVELENGTH'+str(i)] = df.WAVELENGTH
```

```
#pemotongan panjang gelombang setiap spektrum objek

dbdropna = pd.read_csv('DAdatakotor.csv')
sc = pd.read_csv('sc') #obsid lamost

data = pd.DataFrame([])
for i in sc.obsid:
    df = dbdropna[['WAVELENGTH'+str(i), 'FLUX'+str(i)]]
    df = df[df['WAVELENGTH'+str(i)] != 0].dropna()
    df = df[df['WAVELENGTH'+str(i)] >= 3847]
    df = df[df['WAVELENGTH'+str(i)] <= 8771]
    a = df['FLUX'+str(i)].tolist()
    b = df['WAVELENGTH'+str(i)].tolist()

    data['FLUX'+str(i)] = pd.Series(a)
    data['WAVELENGTH'+str(i)] = pd.Series(b)
```

```
#mencari batas atas dan bawah panjang gelombang
#dari setiap spektrum objek

wvmaxti = []
wvminti = []
for i in sc.obsid:
    wvmax = data['WAVELENGTH'+str(i)].max()
    wvmin = data['WAVELENGTH'+str(i)].min()
    wvmaxti.append(wvmax)
    wvminti.append(wvmin)

datawv = {'max':wvmaxti, 'min':wvminti}
wvi = pd.DataFrame(datawv)

print('batas bawah',wvi['min'].unique())
print('batas atas',wvi['max'].unique())
print(len(a))
print(len(b))
```

```
#jumlah objek dengan panjang gelombang berbeda

from collections import Counter

print(Counter(wvi['min']).values())
print(Counter(wvi['max']).values())
```

```
#set objek dengan obsid 210043 sebagai acuan

datawv = data[['WAVELENGTH101076','FLUX101076']]
for i in sc.obsid:
    datawv['FLUX'+str(i)] = data['FLUX'+str(i)]
datawv.head()
```

KLASIFIKASI RANDOM FOREST

Train-test split

```
#Mengubah kelas target menjadi nilai 1 (DA) dan 0 (DAZ)
dadaz = ['DA']

data['wd_subclass'] = np.where(data['wd_subclass'].isin(dadaz), 1, 0)
```

```
from sklearn.model_selection import train_test_split
```

```
X = data.drop('wd_subclass', axis=1)
y = data['wd_subclass']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)
```

```
X_train.shape, X_test.shape
```

Training

```
from sklearn.ensemble import RandomForestClassifier
```

```
rfc = RandomForestClassifier(n_estimators=500, max_features = 25)
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier(max_features=25, n_estimators=500)
```

```
arr_feature_importances = rfc.feature_importances_
arr_feature_names = X_train.columns.values

df_feature_importance = pd.DataFrame(index=range(len(arr_feature_importances)),
                                     columns=['feature', 'importance'])
df_feature_importance['feature'] = arr_feature_names
df_feature_importance['importance'] = arr_feature_importances
df_all_features = df_feature_importance.sort_values(by='importance',
                                                    ascending=False)
df_all_features
```

Validation

```
y_pred_proba = rfc.predict_proba(X_test)[:][:,1]

df_actual_predicted = pd.concat([pd.DataFrame(np.array(y_test),
                                              columns=['y_actual']),
                                 pd.DataFrame(y_pred_proba,
                                              columns=['y_pred_proba'])],
                                axis=1)
df_actual_predicted.index = y_test.index
```

AUC

```
from sklearn.metrics import roc_curve, roc_auc_score
```

```
fpr, tpr, tr = roc_curve(df_actual_predicted['y_actual'],
                         df_actual_predicted['y_pred_proba'])
auc = roc_auc_score(df_actual_predicted['y_actual'],
                    df_actual_predicted['y_pred_proba'])

plt.plot(fpr, tpr, label='AUC = %0.4f' %auc)
plt.plot(fpr, fpr, linestyle = '--', color='k')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend()
```

Confusion Matrix

```
#Threshold = 0.5
y_pred = []
for i in y_pred_proba:
    if i >= 0.5 :
        y_pred.append(1)
    elif i < 0.5 :
        y_pred.append(0)
```

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
cm = confusion_matrix(y_test, y_pred, labels=[1,0], normalize = 'true')
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=['DA','DAZ'])
disp.plot(cmap='Blues')
plt.title('Confusion Matrix')
plt.show()
```

Feature Importances

```
plt.figure(figsize = [18,8])
sns.lineplot(df_all_features.feature, df_all_features.importance)

#Annotate
plt.annotate('Fe I 3850', xy=[3850.3490, 0.002791 +  .0006],  ha = 'center',
             color='#320a5e', size = 14)
plt.vlines(x=3850.3490, ymin=0.002791 + .00007, ymax=0.002791 + .0005,
           colors='#320a5e', ls='--', lw=1)

plt.annotate('He I 3889', xy=[3887.7659, 0.006140 +  .0005],  ha = 'center',
             color='#e05536', size = 14)
plt.vlines(x=3887.7659, ymin=0.006140 + .00007, ymax=0.006140 + .0004,
           colors='#e05536', ls='--', lw=1)

plt.annotate('CaII K 3934', xy=[3932.7842, 0.001787 +  .0005],  ha = 'center',
             color='#781c6d', size = 14)
plt.vlines(x=3932.7842, ymin=0.001787 + .00007, ymax=0.001787 + .0004,
           colors='#781c6d', ls='--', lw=1)

plt.annotate('CaII H 3968', xy=[3972.8313, 0.002643 +  .0005],  ha = 'center',
             color='#62146e', size = 14)
plt.vlines(x=3972.8313, ymin=0.002643 + .00007, ymax=0.002643 + .0004,
           colors='#62146e', ls='--', lw=1)

plt.annotate('H-delta 4101', xy=[4101.0977, 0.002865 +  .0007],  ha = 'center',
             color='#fbb61a', size = 14)
plt.vlines(x=4101.0977, ymin=0.002865 + .00007, ymax=0.002865 + .0006,
           colors='#fbb61a', ls='--', lw=1)
```

```
plt.annotate('Mg I 5184', xy=[5185.6133, 0.000449 +  .0005],  ha = 'center',
             color='#8f2469', size = 14)
plt.vlines(x=5185.6133, ymin=0.000449 + .00007, ymax=0.000449 + .0004,
           colors='#8f2469', ls='--', lw=1)

plt.annotate('Ca I 5580', xy=[5580.8470, 0.000845 +  .0005],  ha = 'center',
             color='#1b0c41', size = 14)
plt.vlines(x=5580.8470, ymin=0.000845 + .00007, ymax=0.000845 + .0004,
           colors='#1b0c41', ls='--', lw=1)

plt.annotate('Na 5895', xy=[5893.8643, 0.001431 +  .0005],  ha = 'center',
             color='#cf4446', size = 14)
plt.vlines(x=5893.8643, ymin=0.001431 + .00007, ymax=0.001431 + .0004,
           colors='#cf4446', ls='--', lw=1)

plt.annotate('O I 6301', xy=[6300.8643, 0.002918 +  .0005],  ha = 'center',
             color='#bc3754', size = 14)
plt.vlines(x=6300.8643, ymin=0.002918 + .00007, ymax=0.002918 + .0004,
           colors='#bc3754', ls='--', lw=1)

plt.annotate('H-alpha 6562', xy=[6564.4766, 0.004325 +  .0005],  ha = 'center',
             color='#ed6925', size = 14)
plt.vlines(x=6564.4766, ymin=0.004325 + .00007, ymax=0.004325 + .0004,
           colors='#ed6925', ls='--', lw=1)


plt.ylabel('Gini importance')
plt.xlabel('Panjang Gelombang (Å)')
plt.legend()
plt.grid()
plt.show()
```

VARIASI PARAMETER

```
max_features = [25, 50, 75, 100, 125, 150]
akurasi = []
```

```
#Mendefinisikan fungsi untuk algoritma RF sebagai fungsi max_features
#dan n_estimators

def model(n_estimators, max_features):

    from sklearn.model_selection import train_test_split
    X = data.drop('wd_subclass', axis=1)
    y = data['wd_subclass']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                        random_state=42)


    from sklearn.ensemble import RandomForestClassifier
    rfc = RandomForestClassifier(n_estimators = n_estimators,
                                 max_features = max_features,
                                 oob_score = True, class_weight = 'balanced')
    rfc.fit(X_train, y_train)
    y_pred_proba = rfc.predict_proba(X_test)[:][:,1]
    df_actual_predicted = pd.concat([pd.DataFrame(np.array(y_test),
                                                  columns=['y_actual']),
                                     pd.DataFrame(y_pred_proba,
                                                  columns=['y_pred_proba'])],
                                    axis=1)
    df_actual_predicted.index = y_test.index

    from sklearn.metrics import roc_curve, roc_auc_score
    auc = roc_auc_score(df_actual_predicted['y_actual'],
                        df_actual_predicted['y_pred_proba'])

    akurasi.append(auc)
```

```
# n = 100, 200, 300, 400, 500
for i in max_features :
    model(100,i)
```

```
d = {'max_features': max_features,
     'n500': akurasi[0:6],
    'n400' : akurasi[6:12],
    'n300' : akurasi[12:18],
    'n200' : akurasi[18:24],
    'n100' : akurasi[24:30]}
df = pd.DataFrame(data=d)
```

```
plt.figure(figsize=(8,6))
sns.lineplot(df.max_features, df.n100, label = 'n_estimators = 100')
sns.lineplot(df.max_features, df.n200, label = 'n_estimators = 200')
sns.lineplot(df.max_features, df.n300, label = 'n_estimators = 300')
sns.lineplot(df.max_features, df.n400, label = 'n_estimators = 400')
sns.lineplot(df.max_features, df.n500, label = 'n_estimators = 500')
plt.ylabel('Akurasi (AUC)')
plt.title('Variasi parameter input terhadap akurasi model')
plt.legend()
plt.grid()
plt.show()
```

REGRESI

```
loggdata = lamost[lamost.logg != -9999]
loggdata.head(3)
```

```
cek = pd.read_csv('data olah.csv')
cek.rename(columns = {'Unnamed: 0':'Wavelength'}, inplace = True)
cek.head(4)
```

```
from sklearn.model_selection import train_test_split

X = loggdata.drop(['Wavelength','logg','wd_subclass'], axis=1)
y = loggdata['logg']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)
```

```
from sklearn.ensemble import RandomForestRegressor

rfc = RandomForestRegressor(n_estimators=100, max_depth = 100,
                            oob_score=True, random_state = 0)
rfc.fit(X_train, y_train)
```

```
from sklearn.metrics import mean_absolute_error

val_pred = rfc.predict(X_test)
val_mae = mean_absolute_error(val_pred, y_test)
print("Validation MAE : {:,.9f}".format(val_mae))
```

```
y_pred = rfc.predict(X_test)
testurut = pd.DataFrame({'y_test':y_test,'y_pred' : y_pred})
testurut1 = testurut.sort_values(by='y_pred')
testurut1.reset_index(inplace=True)
testurut1['x'] = testurut1.index
```

```
#resclae dan bin nilai per objek
testurut1['interval'] = pd.cut(testurut1.x, 20)

mean   = testurut1.groupby(testurut1.interval).mean()['y_test']
mean1 = testurut1.groupby(testurut1.interval).mean()['y_pred']
median = testurut1.groupby(testurut1.interval).median()['x']
```

```
#plt.figure(figsize=(15,10))
plt.scatter(median, mean, color = "red", label = 'true')
plt.scatter(median, mean1, color = "blue", label = 'predicted')
for i in testurut1 :
    plt.vlines(x=median, ymin=mean, ymax=mean1, colors='firebrick', alpha=0.2)
plt.ylabel('Log Gravitasi Permukaan')
plt.title('Mean average percentage error : 0.35% \n Mean average error = 0.45')
plt.legend()
plt.grid()
plt.show()
```

PARAMETER DATA

```
logghr = df2[df2.logg >= 0]
plt.figure(figsize=(7,10))
g=sns.scatterplot(x=logghr['bp_rp'], y=logghr['phot_g_mean_mag'],
                  hue=logghr['dadaznonda'],palette="inferno")
g.set_xlabel('bp_rp', fontsize =13)
g.set_ylabel('phot_g_mean_mag', fontsize =13)
plt.gca().invert_yaxis()
plt.grid()
```

```
from astropy.coordinates import SkyCoord
import astropy.units as u

df2 = wd_subclass.copy()
fx = df2[df2.logg != -9999]

eq = SkyCoord(fx.ra, fx.dec, unit=u.deg)
gal = eq.galactic

fig = plt.figure(figsize=(20, 10))
ax = fig.add_subplot(111,projection="aitoff")
image = ax.scatter(gal.l.wrap_at('180d').radian, gal.b.radian,
                   c=fx.logg, cmap="inferno", marker='.')
bar = fig.colorbar(image,orientation="vertical",pad=0.01)
bar.set_label("Log G",size=15)
plt.title("Sampel Katai Putih LAMOST DR8 \n Koordinat Ekliptika",
          fontsize=16)
plt.grid()
fig.show()
```