TESIS

PENGEMBANGAN METODE *DECISION TREE* DENGAN DISKRITISASI DATA DAN *SPLITTING* ATRIBUT MENGGUNAKAN *HIERARCHICAL CLUSTERING* DAN *DISPERSION RATIO*

Dimas Ari Setyawan
NRP. 05111850010043

DOSEN PEMBIMBING

Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

NIP: 19751220 20011220 02

DEPARTEMEN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2020

[*Halaman ini sengaja dikosongkan*]

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M. Kom)
di
Institut Teknologi Sepuluh Nopember
Oleh:

DIMAS ARI SETYAWAN
NRP: 05111850010034

Tanggal Ujian: 16 Juli 2020
Periode Wisuda: Oktober 2020

Disetujui oleh:
Pembimbing:

1. Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
NIP: 19751220 20011220 02 .....

Penguji:

1. Prof. Dr. Agus Zainal Arifin, S.Kom.,M.,Kom.
NIP: 19720809 19951210 01 .....

2. Dr. Ahmad Saikhu, S.Si., MT.
NIP: 19710718 20060410 01 .....

3. Dr. Diana Purwitasari, S.Kom., M.Sc.
NIP: 19780410 20031220 01 .....

Kepala Departemen Teknik Informatika
Fakultas Teknologi Elektro dan Informatika Cerdas

Dr. Eng. Chastine Fatichah, S.Kom., M.Kom
NIP: 19751220 20011220 02

[*Halaman ini sengaja dikosongkan*]

PENGEMBANGAN METODE *DECISION TREE* DENGAN DISKRITISASI DATA DAN *SPLITTING* ATRIBUT MENGGUNAKAN *HEIRARCHICAL CLUSTERING* DAN *DISPERSION RATIO*

Nama Mahasiswa : Dimas Ari Setyawan
NRP : 05111850010043
Pembimbing I : Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

ABSTRAK

Pada umumnya metode *decision tree* menggunakan *information gain* pada proses *splitting*, tetapi dalam perkembangannya *information gain* mempunyai kelemahan ketika dataset memiliki atribut unik seperti *id-product* untuk setiap *record* dan distribusi kelas yang *imbalance*. Pada umumnya data yang digunakan untuk klasifikasi *decision tree* memiliki 2 tipe yaitu numerik dan nominal. Proses pelatihan *decision tree* pada tipe data numerik meggunakan proses diskritisasi data untuk memperoleh *interval* data sehingga memudahkan dalam pembangunan tree. Diskritisasi data numerik dengan *equal interval* memiliki kelemahan yaitu distribusi data yang tidak seimbang.

Kelemahan metode *information gain* pada proses *splitting* dapat ditangai dengan penggunaan metode *dispersion ratio* yang tidak tergantung pada distribusi kelas, tetapi pada distribusi frekuensi. Metode *dispersion ratio* hanya bisa digunakan untuk *splitting* data tipe nominal. Sehigga data yang bertipe numerik akan dilakukan diskritisasi data menggunakan metode *hierarchical clustering* untuk memperoleh *cluster* data yang seimbang. Oleh karena itu, penelitian ini mengembangkan metode *decision tree* dengan disktritisasi data dan splitting atribut menggunakan *hierarchical clustering* dan *dispersion ratio*. Data yang digunakan pada penelitian ini diambil dari UCI *machine learning repository*, yang memilki dua tipe data numerik dan nominal. Ada dua tahap pada penelitian ini yaitu, pertama data yang bertipe numerik dilakukan diskritisasi menggunakan *hierarchical clustering* dengan 3 metode yaitu *single link*, *complete link*, dan *average link*. Kedua, data hasil diskritisasi digabung kembali kemudian dilakukan pembentukan tree dengan *splitting* atribut menggunakan *dispersion ratio* dan di evaluasi dengan *7-fold cross validation*.

Dari hasil evaluasi yang diperoleh, proses diskritisasi data dengan *hierarchical clustering* dapat meningkatkan prediksi sebesar 14,6% dibandingkan dengan data tanpa diskritisasi. Sedangkan proses *splitting* atribut dengan *dispersion ratio* dari data hasil diskritisasi *hierarchical clustering* dapat meningkatkan prediksi sebesar 6,15 %.

Kata kunci : *Decision Tree, Hierarchical Clustering, Dispersion Ratio*

DECISION TREE METHOD DEVELOPMENT WITH DATA DISCRETIZATION USING HEIRARCHICAL CLUSTERING AND DISPERSION RATIO

By : Dimas Ari Setyawan
Student Identity Number : 05111850010043
Supervisor : Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

ABSTRACT

Generally, the decision tree method uses the information gain in the splitting process, but in its development the information gain has weaknesses when the dataset has unique attributes such as ID-product for each record and the distribution of the imbalance class. Generally, the data used for the decision tree classification has two types: numeric and nominal. The process of decision tree training in the numeric data type uses the data diskritization process to obtain the data interval to facilitate the development of the tree. Numerical data diskritization with equal intervals has a weakness i.e. unbalanced data distribution.

The downside of the information gain method in the splitting process can be handled by the use of a dispersion ratio method that does not depend on the class distribution, but on the frequency distribution. The dispersion ratio method can only be used to splitting nominal type data. The numeric data Sehigga will be performed data diskritization using the hierarchical clustering method to obtain a balanced data cluster. Therefore, this research develops the decision tree methods with data disktritization and splitting attributes using hierarchical clustering and dispersion ratio. The data used in this research is derived from the UCI machine learning Repository, which has two types of numerical and nominal data. There are two stages in this research that is, the first numeric-type data is done discretization using hierarchical clustering with 3 methods i.e. single link, complete link, and average link. Second, the data is disritated result recombined then done tree formation with splitting attribute using dispersion ratio and evaluation with 7-fold cross validation.

From the results of the evaluation, data diskritization process with hierarchical clustering can increase the prediction by 14.6% compared to data without disritization. While the splitting attribute process with the dispersion ratio of the results data is discrete hierarchical clustering can increase the prediction by 6.15%.

Keywords : Decision Tree, Hierarchical Clustering, Dispersion Ratio

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

Bab ini menjelaskan mengenai beberapa hal dasar dalam pembuatan penelitian yang meliputi latar belakang, perumusan masalah, tujuan, manfaat, kontribusi penelitian, dan batasan masalah.

1.1 Latar Belakang

Laju perkembangan penelitian, berjalan seirama dengan banyaknya algoritma baru terutama di bidang *machine learning*. Beberapa jenis dari *machine learning* adalah *supervised learning, unsupervised learning*, dan *reinforcement learning*. *Machine learning* bertujuan untuk menganalisis data sebagai bahan belajar sebuah mesin sehingga output yang dihasilkan mampu mengkategorikan data (Herrera Semenets et al., 2017).

Klasifikasi merupakan salah satu jenis dari *machine learning* dengan mererapkan metode *supervised learning*. Metode klasifikasi yang ada yaitu *decision tree algorithm, naïve bayes classifier, neural network, k-nearest neighbour, support vector machine*, dan lain sebagainya (Roy et al., 2019). Menurut (Wang et al., 2014) *decision tree* merupakan metode klasifikasi yang memiliki proses seleksi fitur. Model *tree* yang terbentuk biasanya menggunakan fungsi *Information Gain* (IG) untuk memisahkan antar fitur. *Decision tree* dengan IG mempunyai kekurangan jika dataset berisi atribut kunci seperti *Product-ID*, karena akan dipilih sebagai atribut pemisah dan menghasilkan partisi yang besar (Roy, 2016). IG juga bergantung pada distribusi kelas sehingga jika kelas *imbalance* maka nilai *true positive* dan *false positive* sama (Roy et al., 2019).

Kekurangan pada *decision tree* dengan IG dapat diatasi dengan konsep seleksi fitur yang signifikan. Hal ini telah dilakukan oleh (Roy, 2016) dengan menggunkan *Correlation Ratio* pada proses pemisahan fitur pada data kesehatan. Penggunaan *correlation ratio* masih memiliki kekurangan untuk data yang *non-liner*. Pada penelitian lainnya (Roy et al., 2019) melakukan perbaikan metode pada tahap signifikansi fitur menggunakan metode *dispersion ratio*. Data yang

digunakan pada tahap *dispersion ratio* lebih beragam. *Correlation ratio* dan *dispersion ratio* hanya digunakan untuk proses *splitting tree* sehingga cocok untuk data yang bersifat kategorikal atau nominal. Namun untuk data numerik proses distriktisasi menggunakan konsep *clustering*.

Data numerik memiliki nilai atribut yang sangat banyak sehingga membutuhkan diskritisasi untuk memperoleh interval data (Dash et al., 2011). Selain itu, di dunia nyata data sering kali bertipe numerik sehingga untuk merubah fitur menjadi diskrit menggunakan prosedur *preprosesing* data yaitu diskritisasi (Xu et al., 2010). Diskritisasi juga mengurangi kebutuhan *storage* pada sistem, mempercepat proses *mining* data dan meningkatkan akurasi dari klasifikasi (Xu et al., 2010)

Pada penelitian (Roy, 2016) dan (Roy et al., 2019) proses diskritisasi digunakan untuk data yang bersifat numerik atau kontinu. Menerapkan metode *clustering k-means* untuk proses diskritisasi dengan alasan diskritisasi *equal interval* mengakibatkan nilai distribusi yang tidak seimbang (Dash et al., 2011). Diskritisasi dengan *k-means* mampu menangani nilai batas yang lebih baik dibandingkan dengan *equal interval* biasa (Maslove et al., 2013). Diskritisasi dengan *k-means* sangat tergantung dengan nilai *k* dan inisalisasi *centroid* awal (Kindhi et al., 2018). Selain penentuan nilai *k* dan *centroid* awal yang dapat mengubah jumlah *cluster*, adanya *outlier* data juga mempengaruhi *cluster*.

Diskritisasi dengan *k-means* yang dilakukan oleh (Roy et al., 2019), ternyata juga memiliki masalah yaitu rendahnya akurasi model yang terbentuk. Data yang memiliki atribut numerik lebih banyak daripada nominal akan menghasilkan akurasi rendah contohnya data *Bank Marketing, Thyroid (Allbp), Thyroid (Allhypo), Thyroid (Allrep)* dan *Mammography*. Data bertipe nominal dan numerik mempunyai kompleksitas yang tinggi akan mempengaruhi akurasi dari model karena disktritisasi *K-means* memiliki kelemahan untuk *cluster* yang ukuran dan densitasnya berbeda serta harus berbentuk bulat. *Hierarchical clustering* mengelompokkan data dengan melihat jarak kedekatan data sehingga menghasilkan bentuk cluster yang tidak harus bulat. *Hierarchical clustering* akan menghasilkan jumlah cluster yang tetap pada setiap kali proses *cluster*. Sehingga menghasilkan

cluster yang lebih konstan dan tidak dipengaruhi oleh inisialiasasi awal seperti *k-means* (Horng et al., 2011).

Diskritisasi data yang beratribut numerik menggunakan *k-means* pada penelitian Roy (Roy, 2016; Roy et al., 2019) bertujuan untuk membuat *cluster* yang ukuran dan densitas sama. Diskritisasi dengan metode lain, seperti *equal* interval juga membuat bentuk data selalu berganti, sesuai dengan panjang interval data. Bentuk *decision tree* yang dibangun menggunakan metode *Information Gain* atau *Gini Index* menggasilkan model yang memiliki bias tertentu karena tergantung pada distribusi kelas. Akibatnya, fitur yang digunakan tidak terlalu signifikan. Sehingga diperlukan metode diskritisasi data numerik untuk menghasilkan interval data yang seimbang dan membentuk *tree* yang tidak tergantung pada distribusi kelas.

Oleh karena itu, penelitian ini mengusulkan metode diskritisasi data menggunakan algoritma *hierarchical clustering* untuk tipe data numerik serta proses *splitting decision tree* menggunakan *dispersion ratio*. Tahapan awal, data akan dipisah antara tipe numerik dengan nominal. Data yang bertipe numerik akan didiskritisasi dengan *hierarchical clustering.* Setelah itu data digabung kembali dengan data tipe nominal. Tahapan pembentukan *tree* adalah dengan menggunakan data gabungan serta penerapan metode *dispersion ratio* pada tahap *splitting* atribut.

1.2 Perumusan Masalah

Rumusan masalah yang diangkat dalam penelitian ini adalah sebagai berikut.

1. Bagaimana menerapkan metode diskritisasi data numerik dengan *hierarchical clustering*?
2. Bagaimana menerapkan metode *splitting decision tree* dengan metode *dispersion ratio*?
3. Bagaimana pengaruh penerapan metode *decision tree* dengan diskritisasi dan *splitting* atribut menggunakan *hierarchical clustering* dan *dispersion ratio*?

1.3 Tujuan

Tujuan yang akan dicapai dalam penelitian ini adalah mengembangkan metode klasifikasi *decision tree* dengan menerapkan disktritisasi data numerik dengan *hierarchical clustering* dan *splitting* atribut pada pembentukan *tree* menggunakan *dispersion ratio*.

1.4 Manfaat

Manfaat dari hasil penelitian ini adalah membantu melakukan pembentukan model *decision tree* pada klasifikasi dataset yang *imbalance.* Proses pembentukan *decision tree* menggunakan *dispersion ratio* dengan data numerik yang telah di diskritisasi menggunakan *hierarchical clustering.*

1.5 Kontribusi Penelitian

Kontribusi pada penelitian ini adalah pengembangan model *Decision tree* dengan diskritisasi data menggunakan *hierarchical clustering* untuk menghasilkan interval data yang seimbang dan *splitting* atribut menggunakan *dispersion ratio* unuk membentuk *tree* yang tidak tergantung pada distribusi kelas, tetapi pada distribusi frekuensi.

1.6 Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Menggunakan data dari UCI *machine learning repository.*
2. Data *classification* dengan *integer* dan *categorical*.

BAB 2
KAJIAN PUSTAKA

Bab ini akan menjelaskan tentang kajian pustaka yang terkait dengan landasan penelitian. Pustaka yang terkait adalah seputar *machine learning*, *decision tree, hierarchical clustering, correlation ratio, dispersion ratio, discretization*, dan metode evaluasi.

2.1 *Machine Learning*

Machine learning adalah bagian dari ilmu kecerdasan buatan atau *arificial intelligence* yang fokus pada pembentukan model pembelajaran di mesin dengan menerapkan algoritma matematika dan statistik (Marsland, 2015). Penggunaan algoritma pada *machine learning* minimal memiliki 2 tujuan yaitu, pertama untuk mendapatkan pengetahuan dari data sehingga model yang terbentuk dapat merepresentasikan pola data dan kedua memprediksi kejadian di masa depan dari data yang telah di proses (Putra, 2019).

Penerapan *machine learning* dalam beberapa tahun terakhir telah mengalami peningkatan dalam kehidupan sehari-hari. Rekomendasi film yang harus ditonton, makanan apa yang dipesan atau produk mana yang akan dibeli dan banyak situs web maupun perangkat modern memiliki algoritma *machine learning*. Situs web yang kompleks seperti *Facebook*, *Amazon*, atau *Netflix*, sangat mungkin bahwa setiap bagian situs berisi beberapa model *machine learning*.

Proses pembentukan *machine learning* (Marsland, 2015) ada beberapa tahapan yaitu :

A. *Data collection and preparation*, jadi proses pengumpulan data dan persiapan data merupakan tahap awal yang sangat menentukan bentuk dari data, mulai dari besarnya data, banyaknya *noise* pada data, jumlah fitur pada data, dan tipe data.

B. *Feature selection* merupakan tahapan seleksi fitur yang ada pada data, karena pemilihan fitur sangat membantu dalam proses klasifikasi dengan hanya memilih fitur yang penting.
C. *Algorithm choice* atau pemilihan algoritma yang tergantung dengan bentuk dari data.
D. *Parameter and Model Selection,* kebanyakan algoritma sudah memiliki parameter yang diatur secara manual.
E. *Training,* merupakan tahapan setelah dataset, algoritma dan parameter ditentukan dalam pembentukan model untuk memprediksi data baru.
F. *Evaluation,* merupakan tahapan terakhir sebelum suatu model *machine learning* digunakan. Proses evaluasi bisa berupa akurasi dari model dalam memprediksi.

Tingkat keberhasilan suatu algoritma *machine learning* dalam melakukan pengambilan keputusan tergantung dengan model yang terbentuk dari data. Model yang dapat menggambarkan atau memvisualisasikan data dari *input* dan *output* secara benar memiliki tingkat keberhasilan lebih besar dalam melakukan pengambilan keputusan (Müller & Guido, 2015). Pembangunan model pada *machine learning* memiliki 2 aspek yang terpenting yaitu *training* dan *testing*. *Training* adalah proses pembentukan model dari dataset yang telah ditentukan dan *testing* adalah proses pengujian kinerja dari model yang terbentuk. Ada 4 tipe dari *machine learning* menurut (Marsland, 2015) yaitu :

a. *Supervised learning* adalah proses *training* yang telah ditentukan target atau responnya. Sehingga algoritma yang digunakan akan membentuk model berdasarkan contoh data yang disediakan. Penerapan *supervised learning* untuk metode klasifikasi dan regresi. Contoh dari *supervised learning* yaitu : Pertama, mengidentifikasi kode pos dari angka yang ditulis tangan pada surat, sebagai *input* hasil *scan* dari tulisan tangan dan *output* label angka kode pos yang tertulis pada surat. Kedua, menentukan jenis tumor melalui gambar, dengan *input* adalah gambar, dan *output* adalah jenis tumor. Ketiga, mendeteksi aktivitas penipuan dalam transaksi kartu kredit, sebagai input

adalah catatan transaksi kartu kredit dan output adalah kemungkinan penipuan atau tidak.

b. *Unsupervised learning* adalah proses training yang belum ditentukan target atau responnya. Sehingga algoritma yang digunkan akan mengidentifikasi kedekatan atau kemiripan masing-masing data dan akan dikelompokan menurut kedekatannya. Penerapan *unsupervised learning* untuk metode clustering. Contoh dari *unsupervised learning* yaitu: pertama, mengidentifikasi topik dalam berita, dimana dari dataset yang memiliki *text* berita akan dicari tema atau topik yang dimaksud dari berita tersebut. Kedua, pemetaan pembeli ke dalam sebuah kelompok yang memiliki kemiripan. Ketiga, deteksi pola akses yang abnormal pada sebuah website, dengan mengidientifikasi bug yang ada untuk mencari pola akses yang berbeda dari umumnya.
c. *Reinforcement learning* merupakan gabungan dari *supervised learning* dan *unsupervised learning*. Algoritma akan diberikan jawaban atau *response* yang salah sehingga harus melakukan pendekatan data dengan nilai.
d. *Evolutionary learning* merupakan konsep pembelajaran yang melihat dari cara mahluk hidup dalam melakukan tingkat adaptasi terhadap lingkungan untuk bertahan hidup, dan peluang memiliki keturunan. Sehingga dapat dimodelkan dalam komputer.

2.1.1 *Classification*

Classification atau klasifikasi merupakan metode pembelajaran terawasi atau *supervised learning* yang melakukan klasifikasi data baru berdasarkan pembelajaran dari beberapa data latih yang telah diberi label dan kelas secara benar (Mouthami, 2013). Klasifikasi memiliki tujuan untuk memprediksi label kelas yang telah ditentukan. Label kelas pada klasifikasi ada 2 yaitu *binary clasification* dan *multiclass classification*. *Binary classification* adalah klasifikasi yang memiliki 2 kelas label, seperti contoh kelas label “ya” atau “tidak”. *Multiclass classification* merupakan klasifikasi yang memiliki lebih dari 2 label kelas.

Penggunaan klasifikasi sangat membantu dalam proses penentuan keputusan. Seperti contoh pada Tabel 2.1 Contoh data pembeli perahu (Müller & Guido, 2015)tentang data pembeli yang akan membeli sebuah perahu. Data tersebut diperoleh dari pelanggan yang membeli perahu dan yang tidak tertarik untuk membeli. Klasifikasi dilakukan dengan tujuan untuk promosi tepat sasaran terhadap pelanggan yang benar-benar tertarik untuk membeli perahu. Setelah proses klasifikasi beberapa aturan diperoleh, yaitu umur lebih dari 45, memiliki anak kurang dari 3 atau tidak pernah cerai, maka pelanggan akan membeli perahu. Semakin baik model klasifikasi yang terbentuk akan lebih baik dalam proses prediksi data uji.

Algoritma yang digunakan untuk mengimplementasikan klasifikasi disebut dengan *classifier*. Pemilihan *classifier* tergantung dengan dataset yang ada. Karena setiap *classifier* memiliki keunggulan dan kelemahan dalam melakukan klasifikasi. Menurut (Roy et al., 2019) ada beberapa *classifier* yang umum digunakan yaitu *decision tree algorithm, naïve bayes classifier, neural network algorithm, k-nearest neighbour algorithm,* dan *support vector machine.*

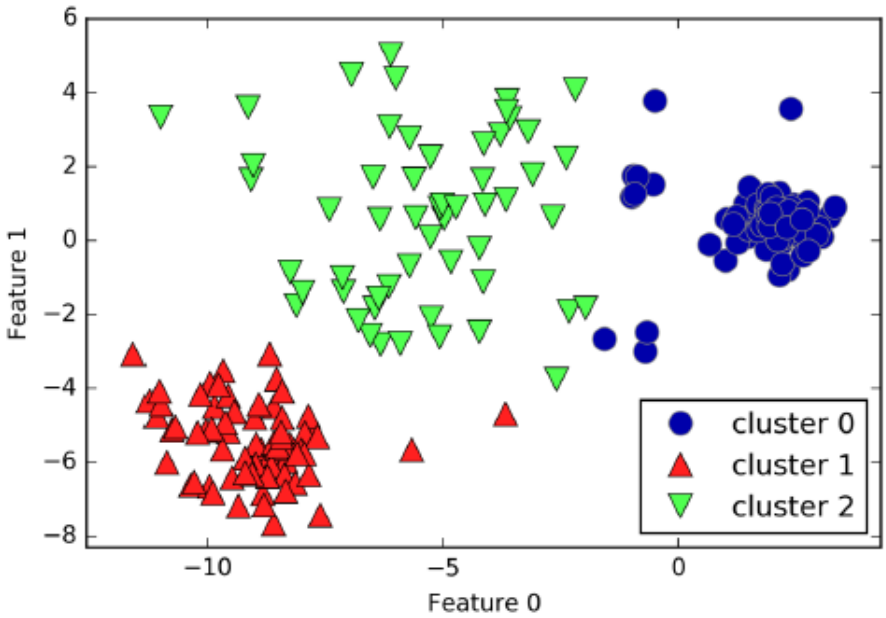
Tabel 2.1 Contoh data pembeli perahu (Müller & Guido, 2015)

Age	Owns Cars	Owns House	Children	Martial Status	Own a dog	Buy a boat
66	1	Yes	2	Widowed	No	Yes
52	2	Yes	3	Married	No	Yes
22	0	No	0	Married	Yes	No
25	1	No	1	Single	No	No
44	0	No	2	Divorced	Yes	No
39	1	Yes	2	Maried	Yes	No
26	1	No	2	Single	No	No
40	3	Yes	1	Maried	Yes	No
53	2	Yes	2	Divorce	No	Yes
64	2	Yes	3	Divorce	No	No
58	2	Yes	2	Maried	Yes	Yes
33	1	No	1	Single	No	No

2.1.2 *Clustering*

Clustering merupakan metode pembelajaran yang tidak terawasi atau *unsupervised learning*, dimana pengelompokan data dilakuakan berdasarkan tingkat kemiripan dan data tidak memiliki label atau kelas. Tingkat kemiripan data dapat dilihat dari jarak data, dimana untuk mengukur jarak bisa menggunakan *euclidean distance* atau *manhattan distance*. Semakin dekat jarak antar data maka data tersebut menjadai satu *cluster* dan sebaliknya semakin jauh jarak data maka berbeda cluster.

Penerapan *clustering* pada saat *upload* gambar di media sosial dijelaskan pada penelitian (Müller & Guido, 2015). Media sosial akan mengelompokkan gambar ke dalam satu group yang memiliki kesamaan. Tetapi dalam prosesnya media sosial tidak tahu gambar mana yang menunjukan siapa dan tidak tahu berapa banyak gambar yang ada. Karena pendekatan yang dilakukan oleh sistem media sosial tersebut adalah mengekstraksi semua wajah dan membaginya kedalam kelompok-kelompok wajah yang memiliki kemiripan.



Gambar 2.1 Contoh Cluster dengan *Density* yang Berbeda (Müller & Guido, 2015)

Pemilihan metode *clustering* juga tergantung dengan *density* atau kerapatan data. Gambar 2.1 menjelaskan bahwa untuk *cluster* 0 dan 1 memiliki kerapatan yang hampir sama, namun untuk *cluster* 2 kerapatannya lebih lebar. Selain kerapatan data, bentuk data juga mempengaruhi pemilihan metode *cluster*, ada beberapa metode *cluster* yang kurang mampu untuk melakukan *clustering* data ketika data berbentuk tabular. *Clustering* dibedakan menjadi 2 macam berdasarkan hasil *cluster* menurut (Rastogi, Guha, & Shim, 2001). Pertama, *partitional* yaitu penentuan jumlah *cluster* sesuai dengan jumlah *k*, contoh algortimanya yaitu *k-means, k-medoid* dan sebagainya. Kedua, *hierarchical* yaitu mengelompokan data berdasarkan struktur taksonomi (Ros & Guillaume, 2019), contohnya *hierarchical clustering, agglomerative clustering* dan sebagainya.

2.2 *Decision Tree*

Decision tree merupakan salah satu pembelajaran yang merepresentasikan pengetahuan dalam bentuk aturan (*rule*) klasifikasi (Horng et al., 2011). Klasifikasi dengan menggunakan metode *decision tree* berguna untuk melakukan klasifikasi untuk data set yang memiliki jumlah variable yang banyak. Algortima *decision tree* memiliki konsep *wrapper* dimana model klasifikasi yang terbentuk sudah memiliki seleksi fitur pada proses pembentukannya (Roy et al., 2019). Menurut (Pandya, 2015) algoritma *decision trees* melakukan pemecahan dataset ke dalam subset yang lebih kecil sehingga akan lebih mempermudah dalam proses pembelajaran atau *learn*. Selain itu *decision tree* juga bisa menangani data set yang bertipe numerik dan kategorikal.

Pembangunan *association rule* pada *decision tree* menggunakan konsep pembentukan pohon keputusan yang berulang, dimana *parent* dari pohon akan melukukan *splitting root* dan membentuk *leaf tree* sampai *leaf* tidak mempunyai cabang. Atribut yang sudah dipilih sebagai *leaf* atau *parent* tidak akan dicoba pada proses percabangan di cabang tertentu. Contoh pada Gambar 2.2 dan Tabel 2.1, *decision tree* yang dibuat dengan algortima C4.5 yang menggunakan perhitungan *information gain* menghasilkan *decision tree play tennis*. *Association rule* pada *play tennis* menggambarkan salah satu keunggulan dari penggunaan klasifikasi

dengan *decision tree*. Fitur *temperature* tidak diperhitungkan dalam kemungkinan *play tennis* yang berarti sudah ada proses seleksi fitur yang terjadi ketika pembentukan *rule.* Dari *rule* tersebut dapat diketahui bahwa kemungkinan permainan tennis bisa terjadi ketika :

1. *Outlook = sunny* dan *Humadity = normal*
2. *Outlook = overcast*
3. *Outlook = rain* dan *Wind = weak*

Tabel 2.2 Data klasifikasi *play tennis*

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Data yang digunakan pada klasifikasi *decision tree* tidak sepenuhnya bertipe nominal sesuai Tabel 2.2, tetapi ada juga yang bertipe numerik atau *continuous.* Data bertipe numerik akan memiliki fitur yang jumlahnya sama dengan banyak data. Jadi untuk mengurangi jumlah fitur pada data numerik ada beberapa cara, antara lain *binary split, multi-way split, equal interval*, diskritisasi dll.

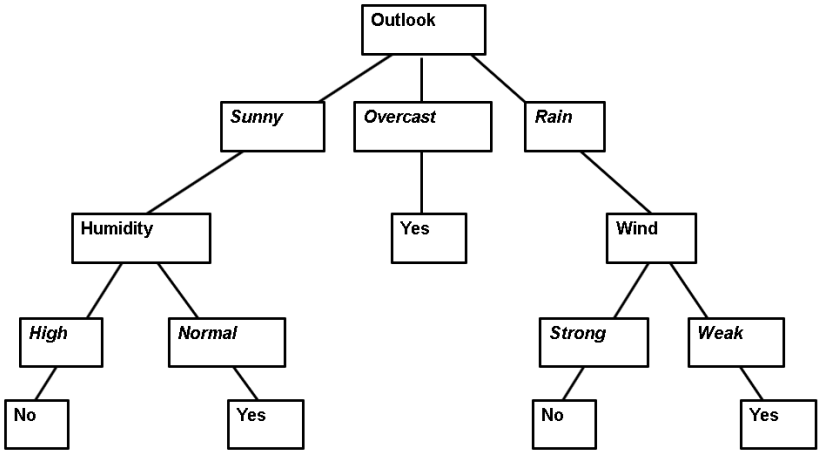
Pada umumnya metode *splitting* pada *decision tree* menggunakan algoritma *Information Gain* dimana atribut yang memiliki nilai *gain* tertinggi akan dipilih (Rutkowski et al., 2013). Perhitungan *splitting* atribut *decision tree* dengan IG menggunkan perhitungan *gain entropy* seperti persamaan (2.1) dan (2.2). Nilai

$p(j|t)$ pada persamaan (2.1) merupakan frekuensi kelas j didalam atribut t. Pada persamaan (2.2) nilai $Entropi(p)$ merupakan nilai entropi dari *parent node* dan nilai k merupakan partisi dari split atribut p.

$$Entropi\ (t) = -\sum_t p(j|t) \log p(j|t) \quad (2.1)$$

$$GAIN_{split} = Entropi(p) - \sum_{i=1}^{k} \frac{n_i}{n} * Entropi\ (i) \quad (2.2)$$

Penerapan IG pada algoritma C4.5 memiliki beberapa kekurangan untuk proses *splitting* data yang bertipe numerik (Putra, 2019). Data bertipe numerik membutuhkan proses diskritisasi atau *binary decision* untuk menghasilkan jarak data. *Binary decision* menghasilkan dua data split, sedangkan diskritisasi bisa menghasilkan dua atau lebih data split.



Gambar 2.2 decison tree hasil dari perhitungan C4.5 pada data Tabel 2.2

2.3 *Hierarchical Clustering*

Hierarchical clustering adalah metode pengelompokan data ke dalam bentuk pohon hierarki yang disebut *dendrogram* (Jafarzadegan et al., 2019). *Hierarchical clustering* membentuk *cluster* yang terbawah atau *root node* sebagai *cluster* setiap data. Sedangkan *cluster* diatasnya dibentuk oleh kedekatan antar data dibawahnya. Menurut (Dash et al., 2011; Jafarzadegan et al., 2019) ada dua cara untuk membentuk *hierarchical clustering* yaitu *divisive* dan *agglomerative*.

Gambar 2.3 menunjukkan semua data point ada pada bagian bawah mulai dari 0 - 11, kemudian membentuk *tree* yang menghubungkan data. Proses

penggabungan antar data dapat dilihat dari kedekatan data, misalnya data 1 dan 4 kemudian 6 dan 9 dan seterusnya sampai menjadi satu *cluster*. Ketika data dibagi menjadi 2 *cluster* maka data 1,4,3,2,8 menjadi satu dan 5,0,11,10,7,6,9 menjadi cluster lain. Jika kluster dibagi menjadi 3 maka 1,4,3,2,8 menjadi satu cluster, 5,0,11 menjadi satu cluster, dan 10,7,6,9 menjadai satu cluster. Dendogram yang digambarkan pada gambar 2.3 sama dengan gambar garis data pada Gambar 2.4. Perbedaannya terletak dari cara penyajian data hasil *clustering*.

Divisive atau *top-down* merupakan metode memecah data dimulai dengan satu *cluster* yang berisi semua data, kemudian dipecah menjadi beberapa *cluster* lainnya. *Agglomerative* atau *down-top* merupakan metode membuat *cluster* untuk setiap data dan kemudian dijadikan cluster baru yang dilihat dari kedekatan data. Kedekatan antar data yang bisa dijadikan satu *cluster* bisa dihitung dengan tiga metode, yaitu *single link*, *complete link*, dan *average link* seperti persamaan (2.3), (2.4), dan (2.5).

$$single\ d_{uv} = \max\{d_{uv}\}, d_{uv} \in D \quad (2.3)$$

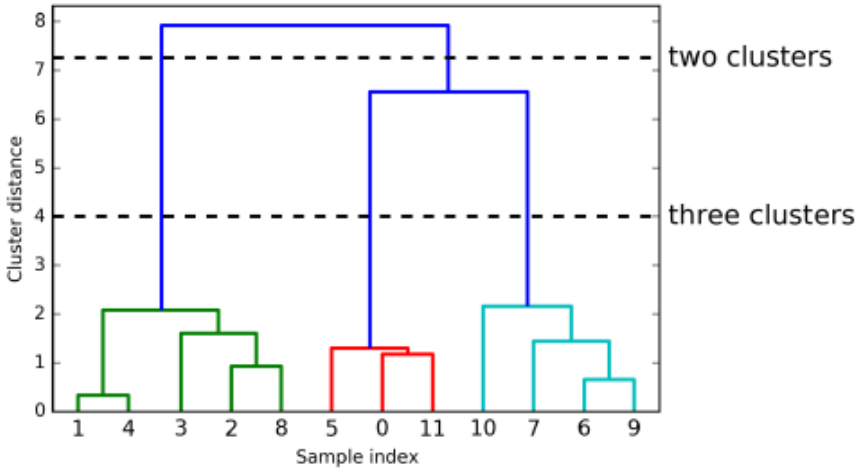
$$complate\ d_{uv} = \min\{d_{uv}\}, d_{uv} \in D \quad (2.4)$$

$$average\ d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D \quad (2.5)$$

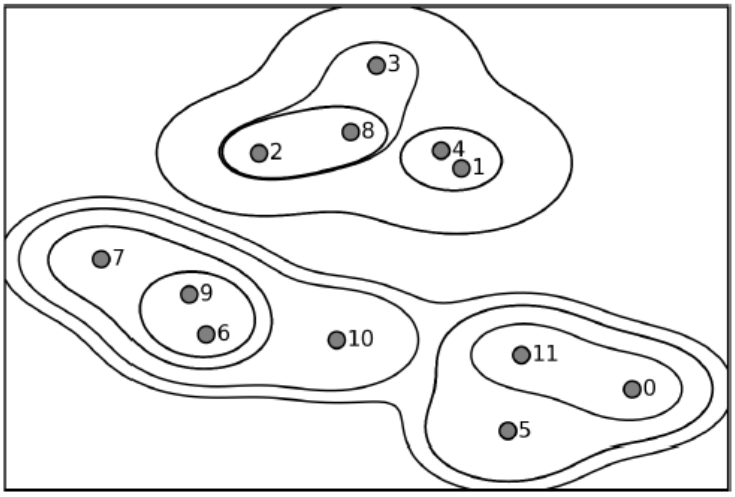
$$D = \sum_{i=0}^{n} |b_i - a_i| \quad (2.6)$$

$$D(a,b) = \sqrt{\sum_{i=0}^{n} (b_i - a_i)^2} \quad (2.7)$$

Perhitungan tingkat kemiripan dengan *single link* dihitung berdasarkan nilai kemiripan terbesar $\max\{d_{uv}\}$ di antara anggota *cluster* sesuai persamaan (2.3). *complete link* merupakan perhitungan tingkat kemiripan data berdasarkan nilai terkecil $\min\{d_{uv}\}$ di antara anggota cluster sesuai persamaan (2.4). Sedangkan *average link* merupakan perhitungan tingkat kemiripan data berdasarkan jarak rata-rata $\text{average}\{d_{uv}\}$ diantara anggota cluster seperti persamaan (2.5). Perhitungan jarak antar data pada metode *hierarchical clustering* bisa menggunakan algoritma *manhattan distance* (2.6) dimana nilai jarak D merupakan jumlah selisih dari $b_i - a_i$, atau menggunakan metode *euclidean distance* (2.7) dimana nilai jarak D merupakan akar kuadrat dari jumlah selisih $(b_i - a_i)^2$.



Gambar 2.3 Dendogram dari *Hierarchical clustering* (Müller & Guido, 2015)



Gambar 2.4 *Hierarchical clustering* dengan garis plot data (Müller & Guido, 2015)

Cluster yang terbentuk dari metode *single link* memiliki kelebihan, yaitu dapat menangani bentuk *cluster non-elips*, tetapi memiliki kekurangan yaitu *sensitive* terhadap *noise* dan *outlier*. *Cluster* dengan metode *complete link* memiliki kelebihan, yaitu sedikit terpengaruh oleh adanya *outlier* dan *noise*, serta memiliki kekurangan bahwa *cluster* besar, maka akan cenderung memecah cluster tersebut. *Cluster* dengan metode *average link* memiliki kelebihan sama dengan *complete link* dan memiliki kekurangan untuk *cluster* yang berbentuk *globular*.

2.4 *Correlation Ratio* (CR)

Dalam kasus *data mining*, pengertian *correlation ratio* adalah tingkat hubungan atau pengaruh satu atribut data terhadap kumpulan data (Roy, 2016). Keterkaitan hubungan atribut dengan kumpulan data harus bersifat *linier* atau juga bisa *non-linier* dengan beberapa keterkaitan hubungan mendasar antar atribut (Li et al., 2015).

Metode *correlation ratio* dapat digunakan untuk membagi sampel *dataset* ke dalam beberapa kategori yang berbeda. Atribut yang dinilai signifikan atau berpengaruh dalam data adalah atribut yang dapat mengidentifikasi minimal satu kelas hasil dengan nilai rata-rata atribut dan rata-rata pada semua kelas berbeda. Atribut yang tidak bisa mengidentifikasi kelas hasil, termasuk ke dalam atribut yang tidak signifikan dan apabila nilai atribut signifikan diganti maka hasil kelas akan berganti juga.

$$\forall y \in Y | \{S_y = \left(x_{jy}^{(1)}, \dots, x_{jy}^{(n)}\right); j = 1, \dots l_y\} \tag{2.8}$$

$$\forall y \in Y | \, \bar{x}_y^{(i)} = \frac{\sum_{j=1}^{ly} x_{jy}^{(i)}}{l_y} \tag{2.9}$$

$$\bar{x}^{(i)} = \frac{\sum_{y \in Y} \sum_{j=1}^{ly} x_{jy}^{(i)}}{l} = \frac{\sum_{y \in Y} l_y \, \bar{x}_y^{(i)}}{l} \tag{2.10}$$

$$Cr_i^2 = \frac{\sum_{y \in Y} l_y (\bar{x}_y^{(i)} - \bar{x}^{(i)})^2}{\sum_{y \in Y} \sum_{j=1}^{ly} (\bar{x}_y^{(i)} - \bar{x}^{(i)})^2} \tag{2.11}$$

Persamaan *correlation ratio* untuk data yang linier dapat ditunjukan oleh persamaan (2.8), (2.9), dan (2.10), sedangkan untuk data yang non linier dapat menggunkan persamaan (2.11). Misalnya ada data set sebanyak l dan Y merupakan hasil kelas label, maka untuk membuat partisi kelas menggunakan persamaan (2.8) dengan S_y adalah himpunan semua hasil y dan $x_{jy}^{(i)}$ adalah nilai dari atribut ke i. Untuk menghitung rata-rata nilai atribut ke i dalam setiap kelas, digunakan persamaan (2.9), sedangkan untuk menghitung rata-rata semua atribut ke i digunakan persamaan (2.10).

2.5 *Dispersion Ratio* (DR)

Penggunaan *dispersion ratio* bertujuan untuk mengurangi keterbatasan dari *correlation ratio*. DR merupakan perbaikan metode dari CR (Roy et al., 2019), yang mana dapat diterapkan untuk menemukan hubungan antar data nominal atau kategorikal.

Dispersion ratio untuk sebuah atribut didefinisikan sebagai akar kuadrat dari rasio dua komponen yang tersusun dari: pembilangnya, penyebaran nilai signifikan atribut terhadap kelas. Penyebutnya merupakan nilai penyebaran atribut terhadap semua kelas.

$$DR_i = \sqrt{\frac{\sum_{y \in Y} n_y \left(\bar{m}_y^{(i)} - \bar{m}^{(i)}\right)^2}{\sum_{y \in Y} \sum_{j=1}^{y} \left(\bar{v}_{jy}^{(i)} - \bar{m}^{(i)}\right)^2}} \tag{2.12}$$

Persamaan (2.12) merupakan nilai *dispersion ratio* atribut i, dimana y merupakan kelas lebel dan Y adalah kumpulan dari kelas sehingga $y \in Y$. Nilai n_y merupakan jumlah data untuk label kelas. Nilai $\bar{m}_y^{(i)}$ adalah nilai signifikan atribut ke i dari kelas tersebut, sedangkan $\bar{m}^{(i)}$ adalah nilai signifikan semua atribut terhadap kelas dan $\bar{v}_{jy}^{(i)}$ merupakan nilai signifikan atribut ke *j* sampai *i* terhadap kelas.

Proses perhitungan DR dapat dicontohkan untuk data sampel pada Tabel 2.3 dengan label kelas *buy* dan *don't buy*, fitur *income* yang memiliki 3 atribut yaitu *Low*, *Medium* dan *High.* Frekuensi nilai-nilai atribut yang berbeda sehubungan dengan setiap kelas diberikan di setiap sel tabel. Nilai frekuensi maksimum untuk atribut dalam kelas tertentu digunakan untuk menghitung nilai signifikan dari atribut terhadapt kelas. Dalam contoh $\bar{m}_{buy}^{(i)}$ dan $\bar{m}_{d'buy}^{(i)}$ adalah nilai signifikan yang menunjukkan bobot relatif dari atribut di masing-masing kelas *Buy* dan *Don't Buy*. Nilai signifikan keseluruhan $\bar{m}^{(i)}$ atribut adalah rasio penjumlahan dari frekuensi maksimum dari dua kelas dan jumlah total instance dalam dua kelas. Sehingga didapat untuk nilai masing – masing seperti Tabel 2.4.

Tabel 2.3 Contoh dataset DR (Roy et al., 2019)

Class	Income		
	Low	Medium	High
Buy	2	3	5
Don't Buy	4	2	1

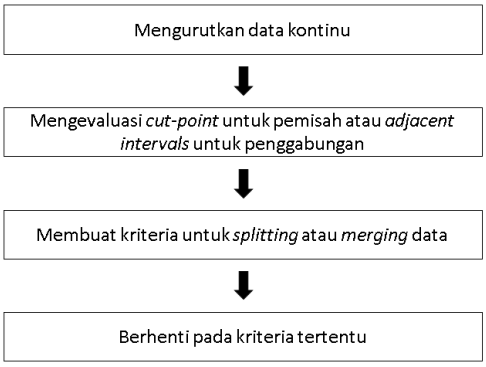
Tabel 2.4 Perhitungan DR Tabel 2.3

$\overline{m}_{buy}^{(i)}$	$\overline{m}_{d'buy}^{(i)}$	$\overline{m}^{(i)}$	DR income
0.5	0.571	0.529	0.209

2.6 *Discretization*

Discretization atau diskritisasi dapat digunakan sebagai tahap *preprocessing* data pada proses *data mining* dan *machine learning* (Maslove et al., 2013). Banyak metode *machine learning* yang menggunakan tahap tersebut untuk menangani data numerik. Diskritisasi memberikan probabilitas kondisional data numerik menjadi nilai kategorikal berdasarkan perhitungan dataset. Disktritisasi juga dapat menurunkan interval data sehingga meningkatkan model klasifikasi yang menggunakan *rule set* (Lustgarten et al., 2011). Diskritisasi juga dapat membantu mengetahui hubungan *non liner* di dalam dataset, termasuk asosiasi diskontinu pada distribusi frekuensi (Muhlenbach et al., 2009).

Preproses disktritisasi data adalah metode *preprocessing* yang bertujuan untuk mengurangi jumlah perbedaaan nilai pada data numerik dengan memberikan rentang data (Jiang et al., 2009). Data hasil diskritisasi akan digunakan untuk proses lebih lanjut antara lain pembentukan model dan proses *data mining*. Menurut (Gama & Pinto, 2014) proses diskritisasi ada 4 tahap sesuai Gambar 2.5.



Gambar 2.5 Tahap proses diskritisasi (Gama & Pinto, 2014)

Menurut (Dash et al., 2011), tujuan proses diskritisasi adalah untuk menentukan *cut point* yang tepat dalam membagi data menjadi beberpa interval kecil. Proses diskritisasi memiliki dua tugas, yaitu menemukan jumlah interval diskrit dan menemukan lebar atau batas dari interval. Tahapan diskritisasi bisa diawali dengan mengurutkan data numerik, kemudian memilih batas atau *landmark* di antara seluruh dataset. Pemilihan landmark bisa dengan cara *top-down* yang dimulai dengan data kosong dan dilakuakan pembagian interval atau *bottom-up* yang dimulai dangan data lengkap dan dilakaukan penggabungan interval.

2.7 Metode Evaluasi

Cross Validation (CV) merupakan metode statsitik yang dapat digunakan untuk mengevaluasi performa dari model *machine learning.* Proses validasi dengan CV menggunakan seluruh dataset yang dibagi menjadi 2 bagian, yaitu dataset untuk pembelajaran (*training subset*) dan dataset untuk testing atau validasi (*validation subset*). Model atau algoritma *machine learning* dibentuk atau dilatih oleh dataset pembelajaran dan kemudian akan divalidasi oleh dataset validasi.

Metode CV pada umumnya ada 3 yaitu *hold-out* CV, *k-fold* CV dan *leave-one-out* CV. Metode *Hold-out* CV adalah metode yang membagi dataset menjadi dua bagian secara acak, yaitu data latih dan data validasi. Misalnya 70% data acak sebagai data latih dan 30% sisanya sebagai data validasi. Proses *hold-out* CV tidak menggunakan semua data untuk memvalidasi model, sehingga sangat tergantung pada pembagian dataset. *K-fold* CV membagi dataset menjadi *k* himpunan bagian yang hampir sama. Proses *K-fold* CV dilakukan dengan iterasi sebanyak k kali dimana setiap iterasi himpunan dataset latih dan dataset testing selalu diganti.

Leave-one-out CV (LOOCV) merupakan kondisi ekstrem dari *k-fold* CV dimana k sama dengan jumlah titik data dalam dataset n (Cheng & Pecht, 2012). Proses iterasi satu titik data yang terpilih sebagai dataset validasi dan semua titik data yang tersisa menjadi data latih.

Tabel 2.5 Pembagian dataset *cross validation* dengan *k-fold* = 5

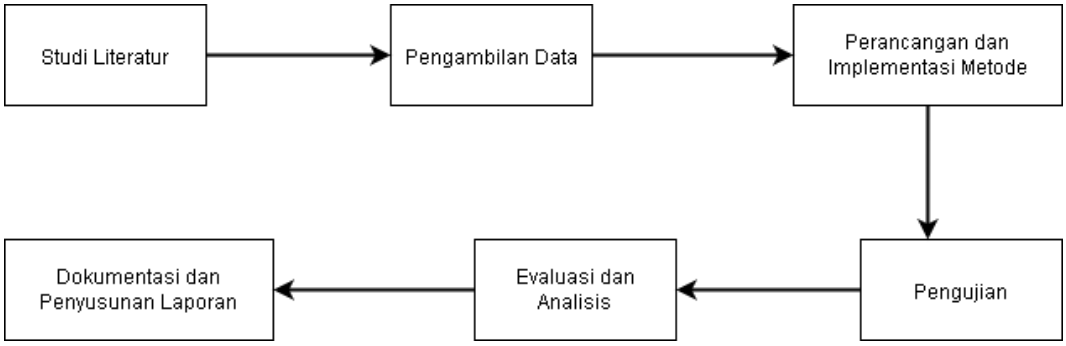
Iterasi	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	Data V	Data L	Data L	Data L	Data L
2	Data L	Data V	Data L	Data L	Data L
3	Data L	Data L	Data V	Data L	Data L
4	Data L	Data L	Data L	Data V	Data L
5	Data L	Data L	Data L	Data L	DataV

Berdasarkan Tabel 2.5 CV dengan *k-fold* = 5 menghasilkan 5 pembagian dataset. Iterasi 1 menghasilkan himpunan dataset 1 yang dijadikan data validasi dan himpunan dataset lain menjadi data latihnya. Proses iterasi dan pemilihan data validasi sesuai dengan k.

[*Halaman ini sengaja dikosongkan*]

BAB 3
METODOLOGI PENELITIAN

Bab ini akan memaparkan tentang metodologi penelitian yang digunakan pada penelitian, terdiri dari (1) studi literatur, (2) pengambilan data, (3) perancangan dan implementasi metode, (4) pengujian, (5) evaluasi dan analisis, dan (6) dokumentasi dan penyusunan laporan. Diagram alir metodologi penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Alur metodologi penelitian

Penjelasan tahapan metode penelitian pada Gambar 3.1 akan diterangkan secara terperinci pada subbab berikut.

3.1. Studi Literatur

Tahap awal pada penelitian ini adalah melakukan kajian beberapa studi literatur yang berkaitan dengan topik penelitian. Studi literatur yang digunakan bersumber dari jurnal, konferensi, dan buku. Referensi digunkan sebagai landasan masalah serta solusi dari permasalahan tersebut. Berdasarkan studi literatur dapat diambil informasi sebagai berikut:

1. *Classification* menggunakan metode *decision tree* menghasilkan model yang telah terseleksi fiturnya (Wang et al., 2014).
2. *Decision tree* menggunkaan metode IG memiliki bias ketika data beratribut banyak (Roy, 2016).
3. Metode *dispersion ratio* (Roy et al., 2019) dan *correlation ratio* (Roy, 2016) dapat diggunkan sebagai metode splitting atribut pada *decision tree.*

4. Diskritisasi atribut numerik dapat dilakukan dengan metode *clustering* (Maslove et al., 2013).

3.2. Pengambilan Data

Data pada penelitian ini menggunakan dataset yang telah disediakan oleh UCI *machine learning repository*. Data yang digunakan merupakan data yang sama seperti pada penelitian (Roy et al., 2019) dengan sedikit pengurangan data karena disesuaikan dengan penelitian. Seperti data *mushroom* dan *hayes* roth, semua atribut pada data karakternya nominal, sehingga untuk proses diskritisasi tidak dapat dilakukan. Tabel 3.1 merupakan daftar data yang digunakan pada penelitian ini. Jenis data yang digunakan beragam seperti data *real, categorical*, dan *integer*.

Tabel 3.1 Dataset UCI *machine learning repository*

No	Data	Character	Instances	Attributes	Class	Number Class
1	Ecoli	Integer, Real	336	7	8	(143:77:52:35:20:5:2:2)
2	Adult	Categorical, Integer	48842	14	2	(12210:36632)
3	Dermatology	Categorical, Integer	366	33	6	(112:61:72:49:52:20)
4	Bank Marketing	Categorical, Integer	4521	16	2	(521:4000)
5	Zoo	Categorical, Integer	101	17	7	(41:20:5:13:4:8:10)
6	Credit Approval	Categorical, Integer, Real	690	15	2	(307:383)
7	Statlog(heart)	Real, Categorical	270	13	2	(151:119)

Pada penelitian sebelumnya (Roy et al., 2019), jumlah dataset yang digunakan ada 16. Sedangkan untuk penelitian ini menggunakan dataset dengan tipe gabungan antara tipe kategorikal dengan tipe *integer* atau *real*. Data tambahan untuk penelitian ini yaitu: *Adult, Dermatology, Credit Approval dan Zoo.* Data yang hanya memiliki atribut kategorikal tidak digunakan pada penelitian ini karena dalam tahap diskritisasi.

Pada beberapa data memiliki kriteria yang dapat mendukung penggunaan dari metode *dispersion ratio* dengan adanya data dengan kelas label yang

imbalance. Seperti contoh data *ecoli, adult, dermatology,bank marketing*, *zoo, credit approval*, dan *satlog(heart).* Data ecoli merupakan dataset yang atributnya bertipe numerik semua, sedangkan data yang lain, atributnya bertipe gabungan antara numerik dan nominal.

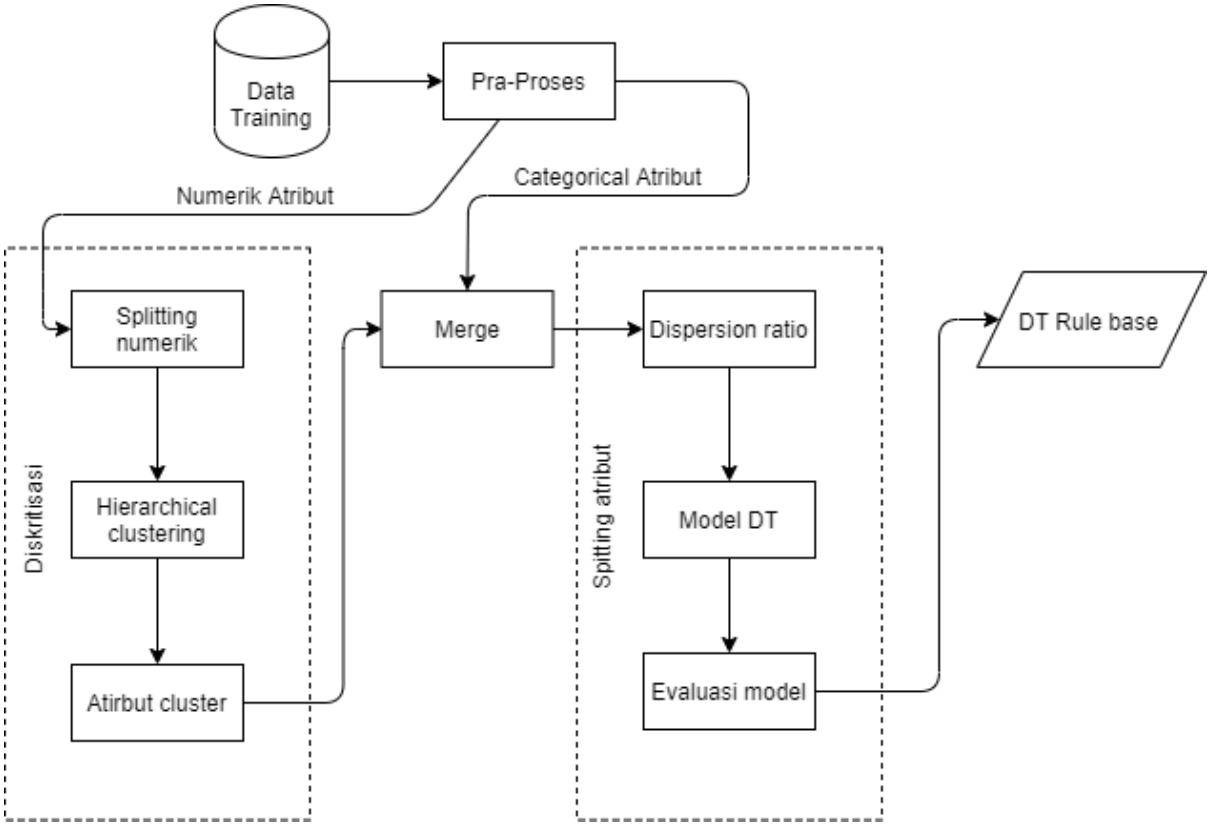
3.3. Perancangan dan Implementasi Metode

Diagram alir perancangan metode pada klasifikasi dengan *decision tree* terdiri dari beberapa tahap sesuai Gambar 3.2. Tahap pertama, data *training* dipisah menjadi 2 bagian yaitu data yang bertipe numerik dan data yang bertipe *categorical*. Tahap kedua, proses diskritisasi data yang bertipe numerik dengan metode *hierarchical clustering.* Tahap selanjutnya, merupakan data hasil *clustering* dengan 3 metode *single link, complete link,* dan *average link.* Tahap berikutnya, dilakukan *merge* data numerik hasil clustering dengan data *categorical.* Tahap selanjutnya, dilakukan *splitting* data hasil merger dengan metode *dispersion ratio*. Tahap terakhir, model *decision tree* yang terbentuk dilakukan evaluasi dan dibentuk *rule base* dari klasifikasi. Tahap-tahap usulan akan dijelaskan pada subbab berikut:

3.3.1 Pra-Proses

Pra-proses dilakukan pada tahap awal rancangan model. Tahap pra-proses dilakukan untuk membagi data menjadi dua, yaitu data yang bertipe numerik dan data yang bertipe nominal yang digambarkan dengan Gambar 3.2. Data yang bertipe numerik akan diproses kedalam tahap diskritisasi. Data nominal digunakan untuk proses *splitting* yang telah digabungkan ulang dengan data numerik yang telah diskritisasi.

Tahapan pra-proses yang memisahkan data numerik dan nominal bertujuan untuk mempermudah dalam proses diskrtisasi. Pemilihan data numerik ditentukan secara manual, dengan menentukan atribut yang bertipe numerik. Data ecoli yang tipe data atributnya numerik semua, maka langsung dilakukan proses disktritisasi, tanpa aadanya pra-proses. Pra-proses dilakukan hanya untuk data yang memiliki tipe gabungan antara numerik dan nominal.

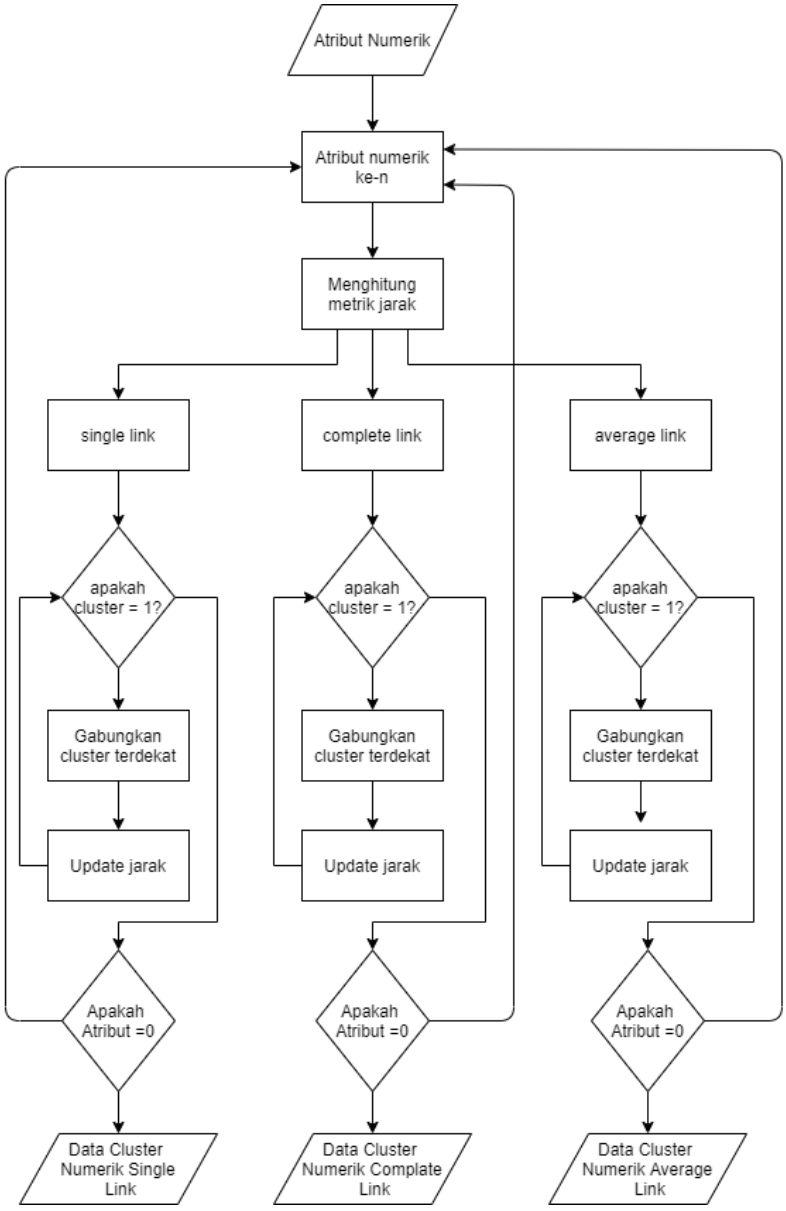


Gambar 3.2 Diagram Alir metode usulan

3.3.2 Diskritisasi

Tahap diskritisasi dilakukan untuk data yang bertipe numerik yang dijelaskan pada Gambar 3.3. Proses disktritisasi pada penelitian ini menggunakan metode *hierarchical clustering*. Proses *hierarchical clustering* dilakukan dengan cara *agglomerative* (*down-top*) jadi *cluster* akan dibentuk dari semua data kemudian di *update cluster* dengan mencari kedekatan antar data sampai hanya terbentuk satu *cluster.*

Setiap data atribut numerik ke n dihitung jarak kedetatan antar data di dalam metrik jarak. Metrik jarak yang terbentuk akan ditentukan clusternya dengan 3 metode yang berbeda yaitu: *single link* dengan mencari nilai *max* dari metrik, *complete link* dengan mencari nilai *min* dari metrik dan *average link* dengan mencari nilai rata-rata dari metrik.



Gambar 3.3 Proses Diskritisasi dengan *Hierarchical Clustering*

Setiap hasil *cluster* yang terbentuk dari masing-masing metode dijadikan input data pada tahap *merge*. Pada data *single link* dilakukan tahap *merge* dengan data nominal kemudian di hitung nilai *dispersion ratio* untuk membentuk *decision tree* dan dihitung akurasinya. Tahap tersebut juga dilakukan untuk metode *complete link* dan *average link*. Kemudian dibandingkan hasil akurasi *decision tree* yang dibangun dari disktritisasi *hierarchical clustering* dengan *single link, complete link,* dan *average link.*

Proses *hierarchical clustering* dilakukan sebanyak 6 kali setiap data. Proses tersebut meliputi : data 2 *cluster single link*, data 2 *cluster average link*, data 2 *cluster complete link*, data 3 *cluster single link*, data 3 *cluster average link*, dan data 3 *cluster complete link*. Pada dataset tanpa diskritisasi dilakukan langsung

3.3.3 Dispersion Ratio

Perhitungan *dispersion ratio* untuk mengetahui penyebaran nilai dari masing-masing atribut terhadap *class*. Penyebaran nilai ini dihitung dari nilai maksimum setiap atribut terhadap *class* seperti Gambar 3.4, m_i melambangkan nilai signifikan maksimum per *class* label. Jumlah nilai maksimal atribut pada setiap *class* label dibagi dengan total semua data akan menghasilkan nilai m. Persamaan 2.12 nilai DR_i untuk setiap atribut adalah akar kuadrat dari $\sum_{y \in Y} n_y \left(\bar{m}_y^{(i)} - \bar{m}^{(i)} \right)^2$ dibagi dengan $\sum_{y \in Y} \sum_{j=1}^{y} \left(\bar{v}_{jy}^{(i)} - \bar{m}^{(i)} \right)^2$.

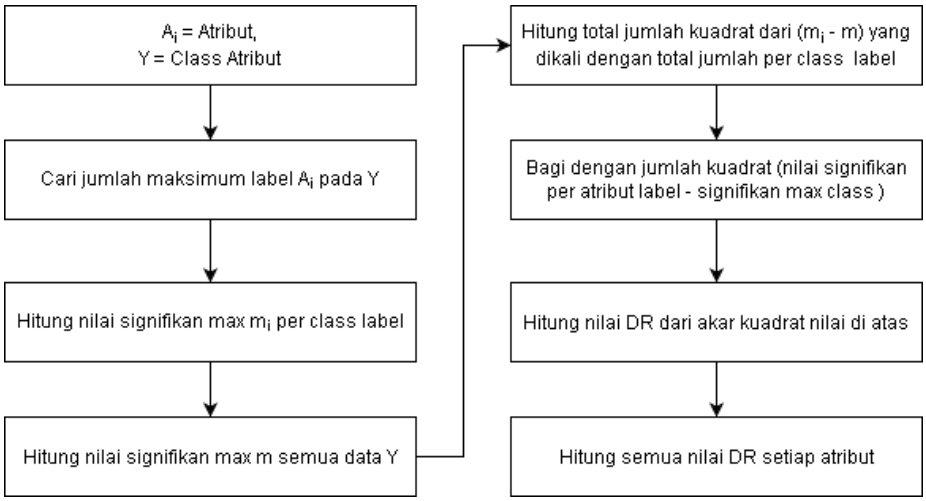
Tabel 3.2 Contoh Class Play Tennis

Atribut Outlook	Play Tennis	
	Yes	No
Sunny	2	3
Overcast	4	0
Rain	3	2

Contoh perhitungan DR pada Tabel 3.2 merupakan jumlah data untuk setiap label atribut untuk masing-masing *class* atribut. *Class* label pada tabel memilik 2 nilai yaitu *yes* dan *no*, sehingga nilai signifikan maksimal atribut memiliki dua m_{yes} dan m_{no}. Proses menghitung DR dari Tabel 3.2 sebagai berikut:

- $m_{yes} = \frac{4}{9} = 0.44\ m_{no} = \frac{3}{5} = 0{,}6\ m = \frac{7}{14} = 0.5$
- $DR^2 = \frac{numerator}{denumerator}$
- $numerator = \left(9 * \left(\frac{4}{9} - \frac{7}{14}\right)^2\right) + \left(5 * \left(\frac{3}{5} - \frac{7}{14}\right)^2\right) = 0.078$

- $denumerator = \left(\frac{2}{14} - \frac{7}{14}\right)^2 + \left(\frac{4}{14} - \frac{7}{14}\right)^2 + \left(\frac{3}{14} - \frac{7}{14}\right)^2 + \left(\frac{3}{14} - \frac{7}{14}\right)^2 + \left(\frac{0}{14} - \frac{7}{14}\right)^2 + \left(\frac{2}{14} - \frac{7}{14}\right)^2 = 0.714$
- $DR_{outlook} = \sqrt{\frac{num}{denum}} = \sqrt{\frac{0.078}{0.714}} = 0.33$



Gambar 3.4 Diagram Alir Dispersion Ratio

3.3.4 Spliting Atribut

Splitting atribut pada *decision tree* diawali dengan inisialisasi *root node* yang menggambarkan keseluruhan himpunan data. Dataset akan di *splitting* dengan menggunakan metode *dispersion ratio.* Pembentukan *splitting decision tree* pada setiap level dihitung dengan mencari nilai *dispersion ratio* tertinggi untuk masing-masing atribut terhadap *class* atribut. Cabang yang sesuai dengan *subtree* dari *root node* diberi label dengan nilai berbeda dan *child nodes* dibangun dari *subtree* dari *root node.* Pada Tabel 3.3 telah ditunjukan algoritma *splitting decision tree* dimana jika ada partisi yang memiliki label kelas yang sama untuk semua data maka *leaf node* memiliki label yang sama dengan label *class* yang sesuai. Jika data partisi kosong maka label kelas mayoritas pada *parent* digunakan untuk memberi label pada *leaf node*. Proses tersebut diulangi sampai semua node memiliki label *class* yang sesuai.

Tabel 3.3 Algortima *Splitting Decision Tree* , diambil dari (Roy et al., 2019)

Constructing DR based Decision Tree
Input: D, N **Output**: A decision tree 1: Create initially the root node associating the whole dataset. 2: Choose the best attribute based on Dispersion Ratio 3: Split the dataset based on the attribute chosen in previous step. 4: for each subset obtained after splitting do a) if all instances are of same class, create leaf node with that class label b) if the subset is empty then assign majority class of the parent node in the associated leaf node; c) if instances belong to different class label, then go to step 2. 8: **end for**

3.4. Uji Coba dan Analisis Hasil

Setelah tahapan perancangan dan implementasi metode, maka pada tahap ini dilakukan uji coba untuk mengetahui akurasi prediksi pada model *decision tree* yang terbentuk. Metode pengujian model *decision tree* menggunakan *cross validation* dengan nilai *k-fold* =7. Dataset untuk pengujian memiliki skema yaitu : a) dataset asli atau tanpa adanya proses diskrtitisasi terlebih dahulu b) dataset disktritisasi *hierarchical clustering* dengan *cluster* k = 2, c) dataset diskritisasi *hierarchical clustering* dengan *cluster* k=3. Berikut merupakan skema uji dataset :

1. Data asli atau tanpa disktritisasi dengan *hierarchical clustering* digunakan untuk proses pembentukan model *decision tree* (untuk data numerik menggunakan *binary split*) dengan *dispersion ratio* dan *information gain.*
2. Data hasil dari diskritisasi dengan *single link* digunakan untuk proses pembentukan model *decision tree* dengan *dispersion ratio* dan *information gain.*

3. Data hasil dari diskritisasi dengan *complete link* digunakan untuk proses pembentukan model *decision tree* dengan *dispersion ratio* dan *information gain.*
4. Data hasil dari diskritisasi dengan *average link* digunakan untuk proses pembentukan model *decision tree* dengan *dispersion ratio* dan *information gain.*
5. Tahap 1,2,3, dan 4 dilakukan uji coba untuk setiap data

[*Halaman ini sengaja dikosongkan*]

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Hasil Penelitian

Uji coba yang dilakukan pada penelitian ini akan dibahas secara detail pada bab ini. Tahapan skenario pengujian akan dilampirkan dengan output dari setiap pengujian. Data uji juga dilampirkan untuk mendukung penjelasan terkait uji coba yang telah dilakukan.

4.1.1 Lingkungan Uji Coba

Lingkungan uji coba perangkat lunak dan perangkat keras yang digunakan untuk implementasi terhadap skenario pengujian dapat dilihat pada Tabel 4.1. Lingkungan perangkat keras digunakan sebagai fasilitas untuk mengimplementasikan tahapan penelitian. Tahapan penelitian dilakukan dengan membuat baris kode sesuai dengan tahapan uji coba pada perangkat lunak yang telah tersedia.

Tabel 4.1 Lingkungan Uji Coba

Lingkungan	Spesifikasi
Perangkat Keras	Komputer laptop : HP Pavilion 15
	Prosesor : AMD Ryzen 5
	RAM : 8 GB
	VGA : GTX 1050 3 GB
Perangkat Lunak	Sistem Operasi : Windows 10
	Tools : Python 3.7
	Library : pandas, numpy, sklearn

4.1.2 Hasil Pra-Proses

Tahapan pra-proses merupakan tahapan untuk memisahkan atribut data berdasarkan tipe data numerik dan nominal. Tabel 4.3 menggambarkan jumlah atribut data bertipe numerik dan data bertipe nominal. Data yang disajikan

merupakan data yang diperoleh dari UCI *machine learning repository* dengan penjelasan pada Tabel 4.2.

Tabel 4.2 Data Pra-Proses

Data	Atribut		Total *Instance*	*Class*
	Numerik	Nominal		
ecoli	7 (mcg, gvh, lip, chg, aac, alm1, alm2)	-	336	cp (143), im (77), pp (52), imU (35), om (20), omL (5), imL(2), imS (2)
adult	6 *(age, fnlwgt, education-num, capital-gain, capital-loss*, dan *hours-per-week)*	8 *(workclass, education, marital-status, occupation, relationship, race, sex*, dan *native-country)*	48842	">50K" (12210) dan "<=50K" (36632).
dermatology	33 (*erythema, scaling, definite borders, itching, koebner, polygonal, follicular, oral mucosal, knee and elbow , scalp, age, melanin, eosinophils, PNL*	1 *(family history)*	366	*psoriasis* (112), *seboreic dermatitis* (61), *lichen planus* (72), *pityriasis rosea* (49), *cronic dermatitis* (52), dan

	infiltrate, fibrosis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing, elongation, suprapapillary epidermis, spongiform pustule, munro microabcess, focal hypergranulosis, granular layer, vacuolisation, spongiosis, saw-tooth, follicular horn plug, perifollicular parakeratosis, inflammatory monoluclear, dan *band-like infiltrate*)			*pityriasis rubra pilaris* (20)
bank marketing	8 *(age, balance, day, duration, campaign, pdays,* dan *poutcome)*	9 (*job, marital, education, default, housing, loan, contact,*	4521	"yes" (521) dan "no" (4000)

		month, dan *poutcome*)		
zoo	1 (*legs)*	14 (*hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic*, dan *catsize*)	101	tipe satu(41), tipe dua(20), tipe tiga(5), tipe empat(13), tipe lima(4), tipe enam(8), dan tipe tujuh(10)
credit approval	6 (A2, A3, A8, A11, A14 dan A15)	9 (A1, A4, A5, A6, A7, A9, A10, A12 dan A13)	690	"+"(307) dan "-"(383)
statlog heart	7 (*age, resting blood pressure, serum cholestoral in mg/dl, maximum heart rate, oldpeak, major vessels*, dan *the slope of the peak)*	6 (*sex, fasting blood sugar > 120 mg/dl, exercise induced angina, chest pain type, resting electrocardiographic*, dan *thal*)	270	*absence*(151) dan *presence*(119)

Dataset yang digunakan adalah kombinasi antara atribut numerik dan nominal seperti *adult, dermatology, bank marketing, zoo, credit approval, statlog heart* atau bisa data dengan tipe numerik saja seperti *ecoli* . Data atribut numerik hasil pra-proses akan digunakan untuk tahap diskritisasi menggunakan *hierarchical clustering.*

Tipe data numerik pada dataset *bank marketing* meliputi data interval, seperti atribut *day, durations, campaign, pdays,* dan *previous,* dimana atribut tersebut memiliki interval data numerik tertentu. Selain itu, tipe data numerik untuk data rasio juga terdapat pada dataset, seperti *age* dan *balance* yang berarti data diperoleh dari pengukuran dan mempunyai titik nol yang *absolute*. Sementara itu, semua atribut tipe kategorikal pada dataset *bank marketing* termasuk data nominal.

Tabel 4.3 Pemisahan Atribut Berdasarkan Tipe Data

No	Dataset	Jumlah Atribut	Atribut Numerik	Atribut Nominal
1	Ecoli	7	7	-
2	Adult	14	6	8
3	Dermatology	34	33	1
4	Bank Marketing	16	7	9
5	Zoo	15	1	14
6	Credit Approval	15	6	9
7	Statlog(heart)	13	7	6

4.1.3 Hasil Spliting Atribut dengan *Information Gain*

Prose *splitting* atribut pada pembentukan model *decision tree* menggunakan algoritma *information gain* dilakukan untuk proses perbandingan nilai prediksi antara IG dan DR. Tahap skenario uji coba pada *splitting* atribut dengan IG sama dengan skenario uji coba pada DR. Skenario uji coba dilakukan dengan menggunakan metode *cross validation* dengan nilai *k-flod* sama dengan 7. Skema uji yang dilakukan yaitu : a) prediksi klasifikasi tanpa *hierarchical clustering,* b) prediksi klasifikasi dengan *hierarchical clustering,* c) prediksi

klasifikasi dengan *hierarchical clustering* berdasarkan jumlah *cluster*. Berikut hasil skema uji coba *splitting* atribut dengan *information gain* :

a. Prediksi klasifikasi dengan IG tanpa menggunakan diskritisasi *hierarchical clustering*. Nilai prediksi maksimal didapat dari proses *cross validation* dengan *k-flod* = 7 sesuai pada tebel 4.4.
b. Prediksi klasifikasi dengan IG menggunakan data diskritisasi *hierarchical clustering single link, average link* dan *complete link*. Nilai prediksi maksimal diperoleh dari proses *cross validation* dengan *k-flod* 7 seperti pada tebel 4.5. Prediksi *single lin*k unggul untuk dataset *adult* dan *zoo*, dan c*omplete link* pada dataset *dermatology*. Dataset *bank marketing* dan *satlog heart* untuk *single link* dan *complete link* nilai prediksinya sama. Dataset *ecoli* untuk nilai prediksi tertinggi pada metode *single link* dan *average link*. Dataset *credit approval* nilai prediksinya sama untuk ketiga metode.
c. Prediksi klasifikasi dengan IG menggunakan data diskritisasi *hierarchical clustering* dengan jumlah *cluster k* = 2 dan *k* = 3. Nilai prediksi maksimal didapat dari proses *cross validation* dengan *k-flod* 7 sesuai pada tebel 4.6. Dataset dermatology, zoo dan bank marketing nilai prediksi tertinggi untuk cluster 2, untuk dataset satlog heart nilai tertinggi pada cluster 3, sedangkan *ecoli, adult* dan *credit approval* nilai prediksinya sama.

Tabel 4.4 Hasil Spliting IG tanpa *hierarchical clustering*

No	Dataset	*Non Hierarchical Clustering*
1	Ecoli	66.66 %
2	Adult	43.78 %
3	Dermatology	84.61 %
4	Bank Marketing	49.84 %
5	Zoo	93.33 %
6	Credit Approval	80.61 %
7	Statlog(heart)	64.11 %

Tabel 4.5 Hasil Spliting IG dengan *hierarchical clustering*

No	Dataset	*Single link*	*Average link*	*Complete link*
1	Ecoli	**95.83** %	**95.83** %	93.75 %
2	Adult	**79.96** %	79.83 %	79.19 %
3	Dermatology	57.69 %	65.38 %	**67.30** %
4	Bank Marketing	83.12 %	83.59 %	**84.52** %
5	Zoo	**93.33** %	92.85 %	92.85 %
6	Credit Approval	77.55 %	77.55 %	77.55 %
7	Statlog(heart)	**87.17** %	82.05 %	**87.17 %**

Tabel 4.6 Hasil Spliting IG berdasarkan jumlah *cluster*

No	Dataset	*k*=2	*k*=3
1	Ecoli	95.83 %	95.83 %
2	Adult	79.96 %	79.96 %
3	Dermatology	**67.30 %**	65.38 %
4	Bank Marketing	**84.52 %**	84.21 %
5	Zoo	**93.33 %**	92.85 %
6	Credit Approval	77.55 %	77.55 %
7	Statlog(heart)	84.61 %	**87.17 %**

4.1.4 Hasil Spliting Atribut dengan *Dispersion Ratio*

Prose *splitting* atribut pada pembentukan model *decision tree* menggunakan algoritma *dispersion ratio*. Skenario uji coba dilakukan dengan menggunakan metode *cross validation* dengan nilai *k-flod* sama dengan 7. Skema uji yang dilakukan yaitu : a) prediksi klasifikasi tanpa *hierarchical clustering*, b) prediksi klasifikasi dengan *hierarchical clustering*, c) prediksi klasifikasi dengan hierarchical clustering berdasarkan jumlah *cluster*. Berikut hasil skema uji coba *splitting* atribut dengan *information gain* :

a. Prediksi klasifikasi dengan DR tanpa menggunakan diskritisasi *hierarchical clustering*. Nilai prediksi maksimal diperoleh dari proses *cross validation* dengan *k-flod* 7 seperti pada tebel 4.7.
b. Prediksi klasifikasi dengan DR menggunakan data diskritisasi *hierarchical clustering single link, average link* dan *complete link*. Nilai prediksi maksimal diperoleh dari proses *cross validation* dengan *k-flod* 7 seperti pada tebel 4.8. Prediksi dengan *single link* unggul untuk 2 dataset yaitu : *ecoli* dan *adult*, prediksi *complete link* untuk dataset *satlog heart*, prediksi *average link* untuk dataset *bank marketing*, dan untuk dataset *credit approval* nilai prediksinya sama, sedangkan dataset *dermatology* dan *zoo* nilai prediksi sama anatara *average link* dan *complete link*.
c. Prediksi klasifikasi dengan DR menggunakan data diskritisasi *hierarchical clustering* dengan jumlah *cluster k* = 2 dan *k* = 3. Nilai prediksi maksimal diperoleh dari proses *cross validation* dengan *k-flod* 7 seperti pada tebel 4.9. *cluster* 2 lebih tinggi nilai prediksi untuk dataset *ecoli* dan *bank marketing*, *cluster* 3 untuk dataset *adult*, *dermatology* dan *satlog*, sedangkan untuk dataset *zoo* dan *credit approval* nilai prediksinya sama.

Tabel 4.7 Hasil Spliting DR tanpa *hierarchical clustering*

No	Dataset	*Non Hierarchical Clustering*
1	Ecoli	81.25 %
2	Adult	35.79 %
3	Dermatology	82.69 %
4	Bank Marketing	54.64 %
5	Zoo	93.33 %
6	Credit Approval	93.87 %
7	Statlog(heart)	74.35 %

Tabel 4.8 Hasil Spliting DR dengan *hierarchical clustering*

No	Dataset	*Single link*	*Average link*	*Complete link*
1	Ecoli	**97.91 %**	95.83 %	93.75 %
2	Adult	**79.14 %**	79.06 %	78.21 %
3	Dermatology	80.76 %	**90.38 %**	**90.38 %**
4	Bank Marketing	82.48 %	**82.94 %**	82.63 %
5	Zoo	92.85 %	**93.33 %**	**93.33 %**
6	Credit Approval	94.89 %	94.89 %	94.89 %
7	Statlog(heart)	87.17 %	82.05 %	**89.47 %**

Tabel 4.9 Hasil Spliting DR berdasarkan jumlah *cluster*

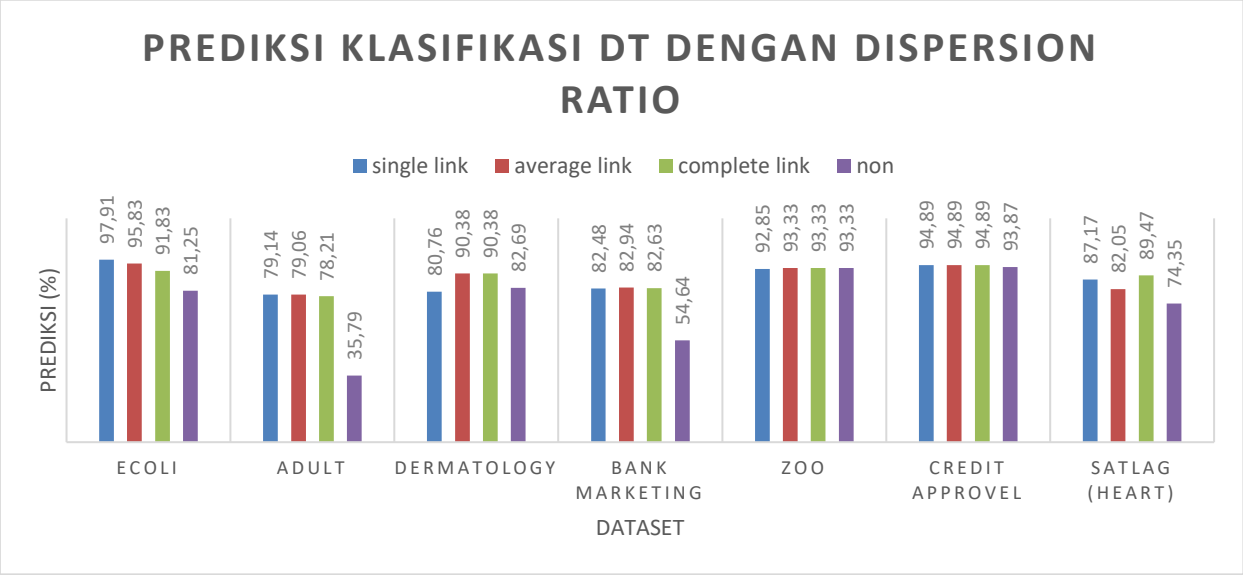
No	Dataset	*k*=2	*k*=3
1	Ecoli	95.83 %	**97.91 %**
2	Adult	**79.14 %**	79.11 %
3	Dermatology	71.69 %	**90.38 %**
4	Bank Marketing	82.01 %	**82.94 %**
5	Zoo	93.33 %	93.33 %
6	Credit Approval	94.89 %	94.89 %
7	Statlog(heart)	**89.47 %**	87.17 %

4.2 Pembahasan

Pada sub-bab ini menjelaskan tentang analisis hasil yang diperoleh dari hasil uji coba. Analisis dari skema uji dilakukan berdasarkan pengaruh diskrtitisasi data dan splitting atribut.

4.2.1 Analisis Diskritisasi dengan Hierarchical Clustering

Hasil uji coba pembentukan model *decision tree* dengan menggunakan data hasil diskrtitisasi *hierarchical clustering* memperoleh beberapa hasil.



Gambar 4.1 Diagram prediksi klasifikasi DT dengan dan tanpa diskritisasi

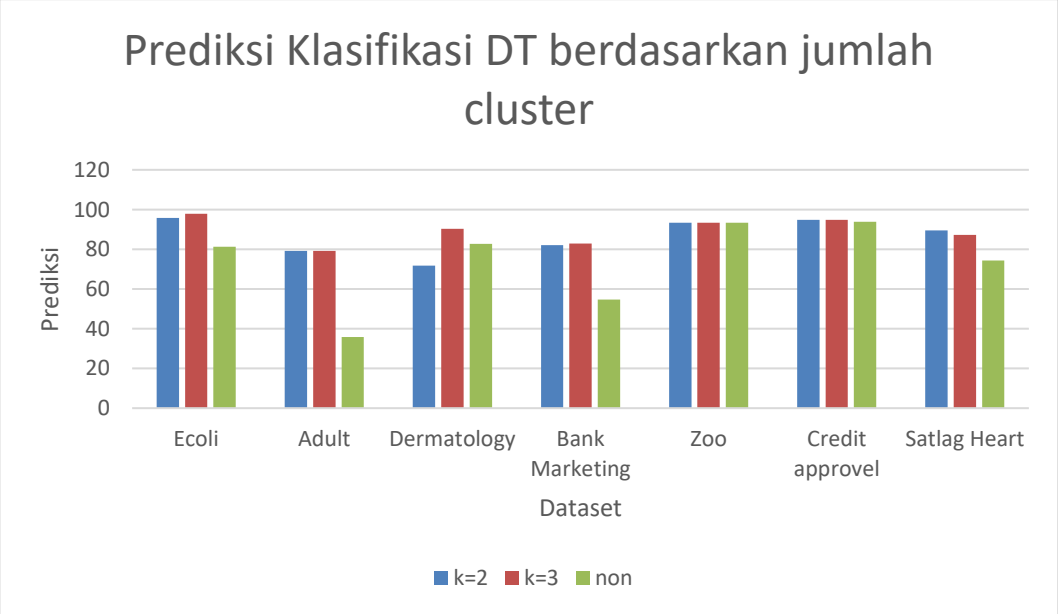
Prediksi klasifikasi model *decision tree* sesuai Gambar 4.1, bahwa data hasil diskritisasi dengan *hierarchical clustering* nilainya lebih baik daripada data yang tidak dilakukan proses diskritisasi *hierarchical clustering*. Hal ini terjadi karena, dataset dengan atribut yang bertipe numerik telah dilakukan proses *cluster* data, sehingga fitur atribut yang digunakan bisa lebih signifikan. Selain itu, untuk data yang tidak melakukan proses *hierarchical clustering* pada atribut numerik, tahapan diskritisasi yang dilakukan hanya menggunakan *binary split*.

Selisih rata-rata nilai prediksi antara prediksi cluster terkecil dengan non cluster adalah 12.45 % dan selisih prediksi cluster terbesar dengan *non cluster* adalah 15.73 %. Perbandingan rata-rata selisih prediksi antara metode *hierarchical clustering* yang telah di uji coba dengan *non cluster* menghasilkan prediksi *complete link* yang memiliki nilai selisih terbesar yaitu 14.97 %, sedangkan *average link* rata-rata selisihnya adalah 14.65 % dan *single link* 14.18 %. Prediksi dengan *single link* unggul 2 dataset yaitu *ecoli* dan *adult*, sedangkan *average link* pada data *bank marketing,* dan *complete link* pada data *satlog heart*.

Jumlah *cluster* dalam pembentukan data diskritisasi juga mempengaruhi nilai prediksi. Gambar 4.2menjelaskan bahwa data dengan jumlah *cluster* yang berbeda, nilai prediksi yang berbeda juga. Perbandingan nilai prediksi dari jumlah *cluster* dapat diperoleh dari selisih prediksi *cluster* dengan prediksi *non cluster*.

Sehingga diperoleh bahwa dataset dengan *cluster* 3 atau *k*= 3 memiliki nilai rata-rata selisih prediksi dengan *non cluster* lebih baik yaitu 15.68 % dibandingkan dataset yang menggunakan *cluster* 2 atau *k*=2 dengan nilai selisih rata-rata adalah 12.92 %.

Perbedaan terbesar antara data yang dilakukan diskritisasi dengan data non-diskritisasi terdapat pada dataset *adult* dan *bank marketing*. Selisih rata-rata prediksi untuk dataset *adult* mencapai 43.2 %, sedangkan dataset *bank marketing* sekitar 28.04 %. Hal tersebut dipengaruhi dengan beberapa atribut numerik yang memiliki *range* data yang tinggi. Seperti contoh pada dataset *bank marketing* atribut *balance* memiliki range 71188 sampai -3313. Sedangkan untuk dataset *adult*, atribut *fnlwgt* memiliki range diantara 12285 samapai 1484705. *Range* data yang tinggi pada atribut numerik, mengakibatkat prediksi *decision tree* dengan dataset non *cluster* memiliki interval data yang tidak seimbang sehingga nilai prediksinya rendah. Sebaliknya, penggunaan metode *hierarchical clustering* dapat meningkatkan prediksi dan menghasilkan cluster yang lebih seimbang.

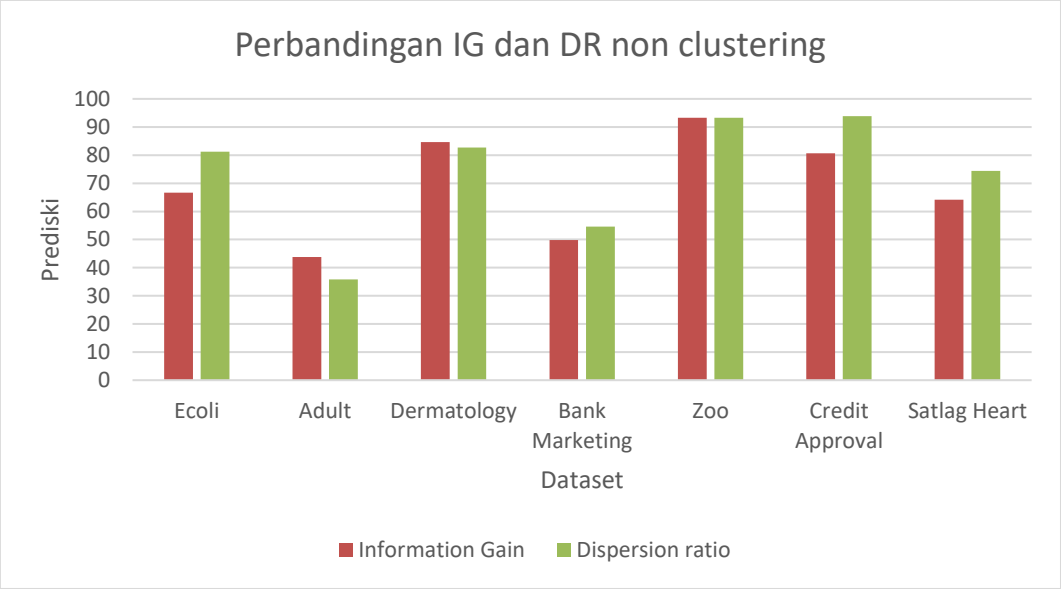


Gambar 4.2 Diagram Prediksi kalsifikasi DT berdasarkan jumlah kluster

4.2.2 Analisis *Spliting* Atribut menggunakan *dispersion ratio*

Hasil dari pembentukan model *decision tree splitting* atribut menggunakan *dispersion ratio* dan *information gain* dengan data yang digunakan belum dilakukan

diskritisasi menggunkan *hierarchical clustering* disajikan dalam diagram Gambar 4.3. Pada diagram tersebut dapat diketahui bahwa selisih prediksi dari *dispersion ratio* dengan *information gain* pada setiap dataset *non cluster* yaitu : *ecoli* = 14.59, *adult* = -7.99, *dermatology* = -1.92, *bank marketing* = 4.8, *zoo* = 0, *credit approval* = 13.26 dan *satlog heart* = 10.24. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 4.71 %. Jadi *splitting* dengan *dispersion ratio* untuk dataset *non cluster* unggul 4 dataset yaitu : *ecoli*, *bank marketing*, *credit approval* dan *satlog heart*. Sedangkan *information gain* unggul 2 dataset yaitu : *adult* dan *dermatology*, dan untuk dataset *zoo* nilai prediksinya sama.

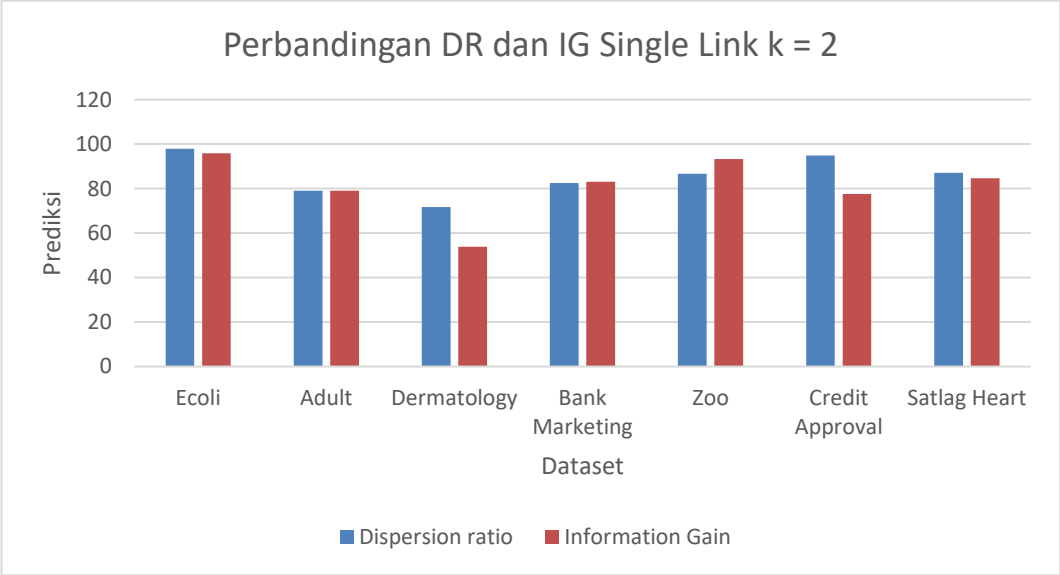


Gambar 4.3 Diagram perbandingan IG dan DR *non clustering*

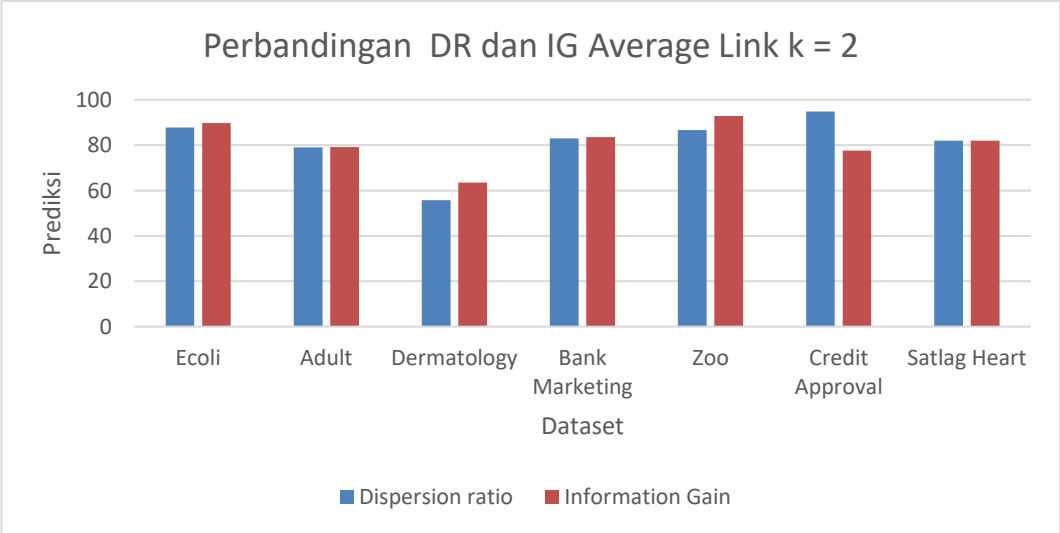
Selain perbandingan nilai prediksi antara *dispersion ratio* dan *information gain* dengan dataset yang belum dilakukan disktritisasi *hierarchical clustering*. Hasil skema uji pada dataset yang sudah dilakukan *hierarchical clustering* dengan beberapa skema seperti dataset dengan 2 kluster, dataset dengan 3 kluster serta metode *hierarchical clustering* menggunakan *single link*, *average link* dan *complete link.*

Dataset dengan jumlah *cluster* 2 dan menggunakan metode *single link* dalam pembentukan cluster menghasilkan nilai prediksi sesuai Gambar 4.4. Pada gambar tersebut dapat dilihat selisih prediksi antara DT yang menggunakan

dispersion ratio dengan *information gain* antara lain : data *ecoli* memiliki selisih 2.08, data *adult* memiliki selisih 0.07, data *dermatology* memiliki selisih 17.85, data *bank marketing* memiliki selisih -0.64, data zoo memiliki selisih -6.67, data *credit approval* memiliki selisih 17.34 dan data *satlog heart* memiliki selisih 2.56. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 4.65 %. Jadi prediksi dataset diskritisasi dengan 2 *cluster single link* untuk pembentukan DT, metode *dispersion ratio* unggul 5 dataset yaitu *ecoli*, *adult*, *dermatology*, *credit approval* dan *satlog heart*, sedangkan *information gain* unggul pada dua data, yaitu *zoo* dan *bank marketing*.

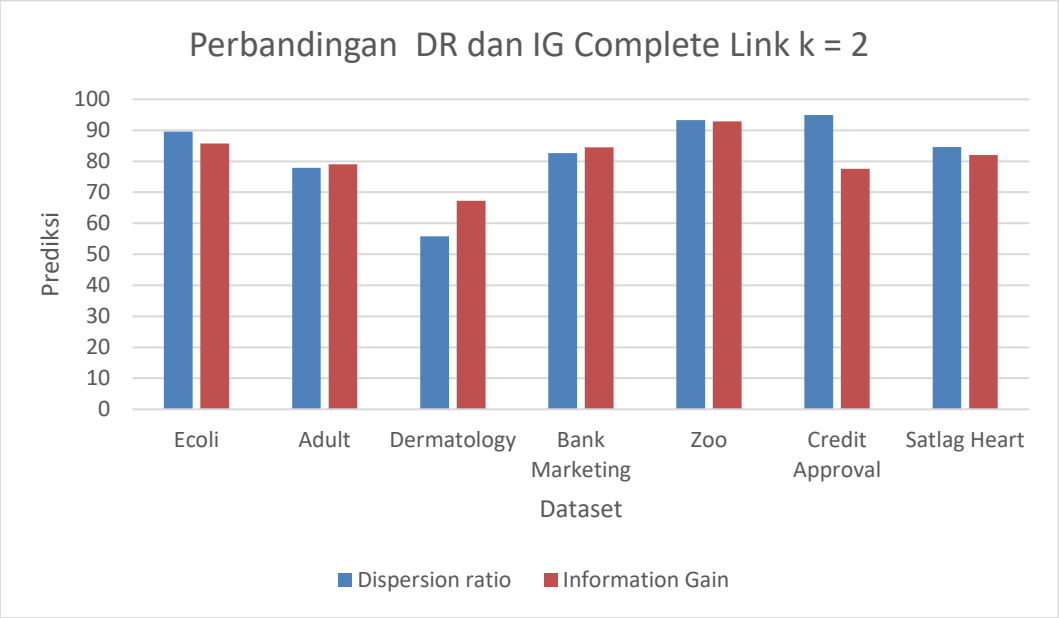


Gambar 4.4 Diagram perbandingan DR dan IG *single link*, dengan 2 *cluster*



Gambar 4.5 Diagram perbandingan DR dan IG *average link*, dengan 2 *cluster*

Dataset dengan jumlah *cluster* 2 dan menggunakan metode *average link* dalam pembentukan *cluster* menghasilkan nilai prediksi seperti Gambar 4.5. Pada gambar tersebut dapat dilihat selisih prediksi antara DT yang menggunakan *dispersion ratio* dengan *information gain* antara lain : data *ecoli* memiliki selisih -2.04, data *adult* memiliki selisih -0.06, data *dermatology* memiliki selisih -7.73, data *bank marketing* memiliki selisih -0.65, data zoo memiliki selisih -6.16, data *credit approval* memiliki selisih 17.34 dan data *satlog heart* memiliki selisih 0. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 0,09 %. Meskipun rata-rata selisih prediksi lebih tinggi *dispersion ratio*, tetapi prediksi dengan *information gain* unggul 5 data yaitu *ecoli*, *adult*, *bank marketing*, *dermatology* dan *zoo*, sedangkan *dispersion ratio* unggul pada dataset *credit approval*.



Gambar 4.6 Diagram perbandingan DR dan IG *complete link*, dengan 2 *cluster*

Dataset dengan jumlah *cluster* 2 dan menggunakan metode *complete link* dalam pembentukan cluster menghasilkan nilai prediksi sesuai Gambar 4.6. Pada gambar tersebut dapat dilihat selisih prediksi antara DT yang menggunakan *dispersion ratio* dengan *information gain* seperti : data *ecoli* memiliki selisih 3.87, data *adult* memiliki selisih -1.16, data *dermatology* memiliki selisih -11.54, data *bank marketing* memiliki selisih -1.89, data zoo memiliki selisih 0.48, data *credit approval* memiliki selisih 17.34 dan data *satlog heart* memiliki selisih 2.56.

Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 1.38 %. Jumlah prediksi dataset dengan *dispersion ratio* unggul 4 data yaitu *ecoli, zoo, credit approval* dan *satlog heart*, sedangkan data *dermatology*, *bank marketing,* dan *adult* prediksi dengan information gain lebih unggul.

Skema uji untuk dataset dengan jumlah cluster sama dengan 2, menggunakan metode *single link, average link* dan *complete link* pada proses prediksi DT dengan *information gain* dan *dispersion ratio* disajikan di Tabel 4.10. Hasil dari skema tersebut, dihitung dari selisih rata-rata hasil prediksi antara *dispersion ratio* dan *information gain* adalah metode *single link* mendapatkan selisih terbesar yaitu 4.65, *complete link* dengan nilai 1.38 dan *average link* dengan 0.09. Jadi prediksi *dispersion ratio* dataset *single link* unggul 4 data yaitu : *ecoli*, *dermatology* ,*satlog heart* dan *credit approval*. Prediksi *dispersion ratio* dataset *average link* unggul satu data, yaitu *credit approval*, *dispersion ratio* dataset *complete link* unggul 2 data, *zoo* dan *credit approval*. Sedangkan prediski *information gain* unggul pada dataset *adult* dengan *cluster average link*, dataset *zoo* dengan *cluster single link,* dan dataset *bank marketing* dengan *cluster complete link.*

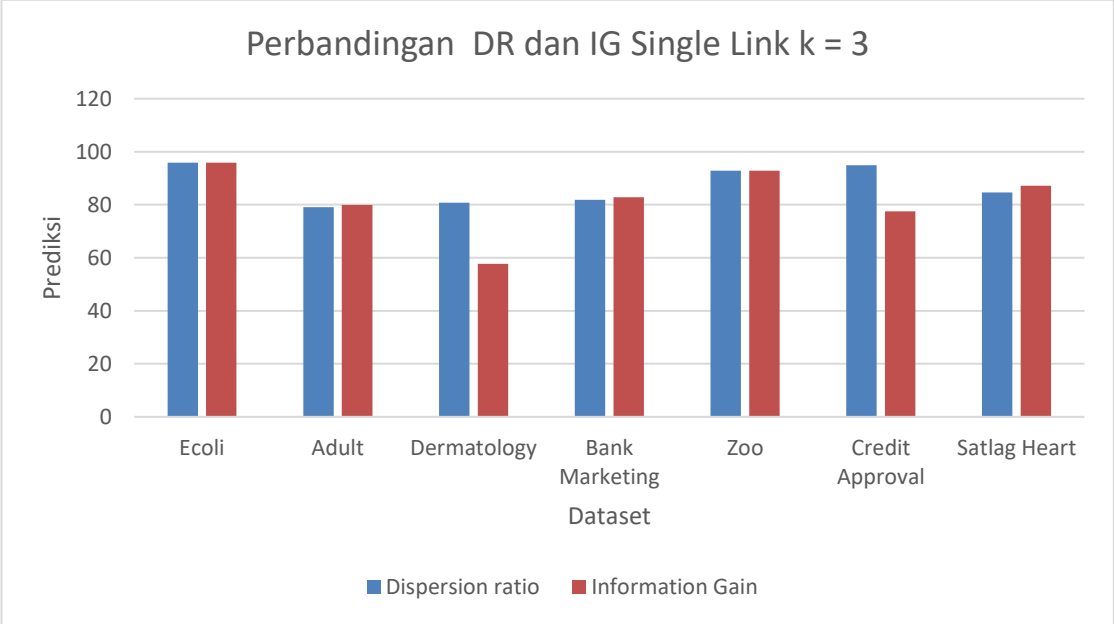
Tabel 4.10 Klasifikasi dengan nilai *k*=2

Dataset	*Dispersion ratio*			*Information gain*		
	single	*average*	*complete*	*single*	*average*	*complete*
Ecoli	**97.91**	87.75	89.58	95.83	89.79	85.71
Adult	79.10	79.06	77.85	79.04	**79.12**	79.01
Dermatology	**71.69**	55.76	55.76	53.84	63.46	67.30
Bank Marketing	82.48	82.94	82.63	83.12	83.59	**84.52**
Zoo	86.66	86.66	**93.33**	**93.33**	92.85	92.85
Credit Approval	**94.89**	**94.89**	**94.89**	77.55	77.55	77.55
Satlog Heart	**87.17**	82.05	84.61	84.61	82.05	82.05

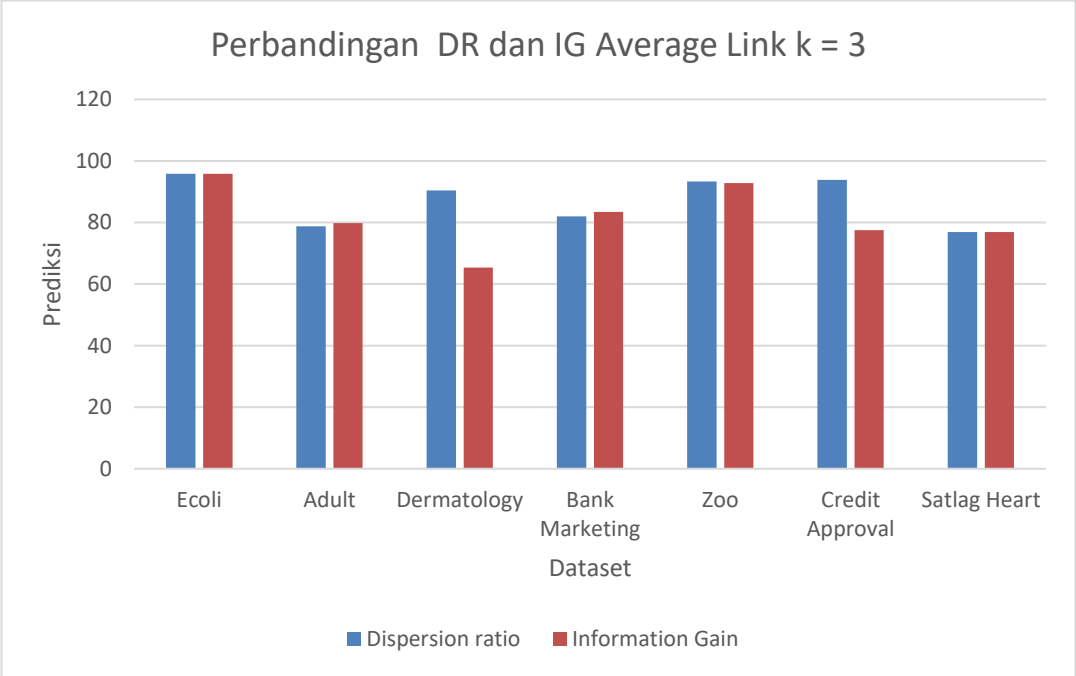
Prediksi dataset *dermatology single link* dengan metode *dispersion ratio* unggul dengan selisih yang tinggi, yaitu 12,47 daripada metode yang lain. Perbedaan yang tinggi ini disebabkan karena dataset *dermatology* memiliki jumlah atribut numerik yang lebih tinggi dari atribut nominal. Terdapat 33 atribut numerik dari total 34 atribut pada dataset *dermatology*.

Uji coba yang dilakukan pada dataset dengan jumlah k=3 atau 3 cluster pada metode *single link* tergambar pada Gambar 4.7. Hasil dari perbandingan selisih pada prediksi DT dengan *dispersion ratio* dan *information gain* yaitu : data *ecoli* memiliki selisih 0, data *adult* memiliki selisih -0.82, data *dermatology* memiliki selisih 23.07, data *bank marketing* memiliki selisih -0.95, data *zoo* memiliki selisih 0, data *credit approval* memiliki selisih 17.34 dan data *satlog heart* memiliki selisih -2.56. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 5.15 %. Prediksi dengan *dispersion ratio* hanya unggul di dua dataset yaitu *dermatology* dan *credit approval*, sedangkan *information gain* unggul pada 3 dataset *adult*, *bank marketing* dan *satlog heart*. Dataset *zoo* dan *ecoli* nilai prediksinya sama.

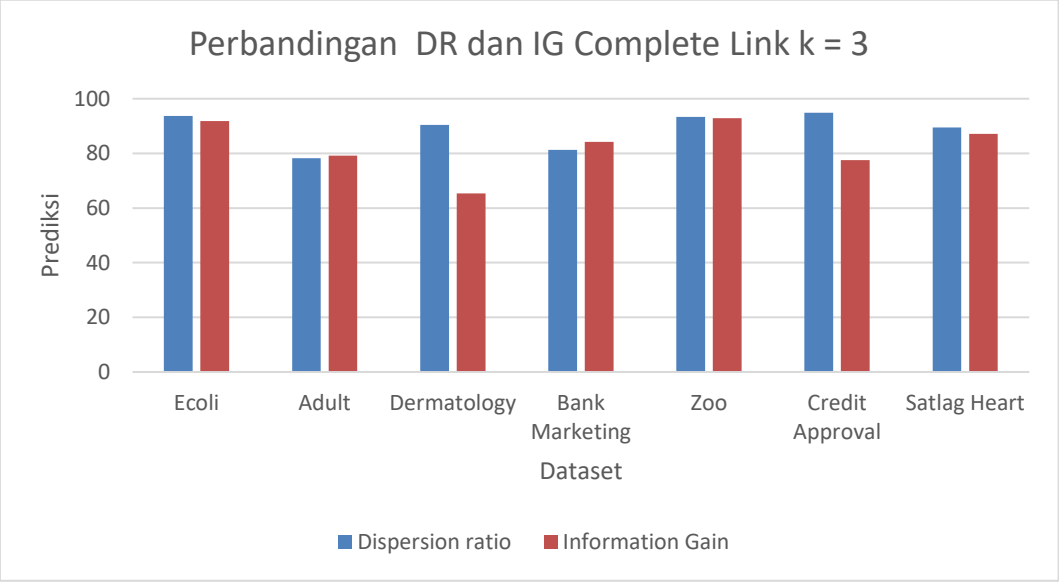
Dataset dengan jumlah k=3 atau 3 *cluster* pada metode *average link* tergambar pada Gambar 4.8. Hasil dari perbandingan selisih pada prediksi DT dengan dispersion ratio dan information gain yaitu : data *ecoli* memiliki selisih 0, data *adult* memiliki selisih -1.05, data *dermatology* memiliki selisih 25, data *bank marketing* memiliki selisih -1,42, data zoo memiliki selisih 0.48, data *credit approval* memiliki selisih 16.32 dan data *satlog heart* memiliki selisih 0. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 5.62 %. Prediksi *dispersion ratio* untuk dataset *average link* dengan 3 *cluster* unggul pada 3 datset yaitu : *dermatology, zoo*, dan *credit approval*. Prediksi dengan *information gain* unggul pada dataset *adult* dan *bank marketing*, sedangkan pada dataset *ecoli* dan *satlog heart* nilai prediksinya sama.



Gambar 4.7 Diagram perbandingan DR dan IG *single link*, dengan 3 *cluster*



Gambar 4.8 Diagram perbandingan DR dan IG *average link*, dengan 3 *cluster*



Gambar 4.9 Diagram perbandingan DR dan IG *complete link*, dengan 3 *cluster*

Dataset dengan jumlah k=3 atau 3 *cluster* pada metode *complete link* tergambar pada Gambar 4.9. Hasil dari perbandingan selisih pada prediksi DT dengan *dispersion ratio* dan *information gain* yaitu : data *ecoli* memiliki selisih 1.92, data *adult* memiliki selisih -0.98, data *dermatology* memiliki selisih 25, data *bank marketing* memiliki selisih -2.95, data zoo memiliki selisih 0.48, data *credit approval* memiliki selisih 17.34 dan data *satlog heart* memiliki selisih 2,3. Sehingga diperoleh rata-rata selisih nilai prediksi untuk DT *dispersion ratio* dengan DT *information gain* adalah 6.15 %. Jadi prediksi dengan *dispersion ratio* lebih unggul pada 6 dataset yaitu : *ecoli, adult, dermatology, zoo, credit approval* dan *satlog heart*, sedangkan untuk hanya *information gain* dua dataset, yaitu *bank marketing* dan *adult.*

Skema uji untuk dataset dengan jumlah cluster sama dengan 3, menggunakan metode *single link, average link* dan *complete link* pada proses prediksi DT dengan *information gain* dan *dispersion ratio* disajikan di Tabel 4.11 Klasifikasi dengan nilai *k*=3. Hasil dari skema tersebut, dihitung dari selisih rata-rata hasil prediksi antara *dispersion ratio* dan *information gain* adalah metode *complete link* mendapatkan selisih terbesar yaitu 6.51, *average link* dengan nilai 5.88 dan *single link* dengan 5,29. Jadi prediksi *dispersion ratio* dataset *complete link* unggul pada 4 dataset yaitu *dermatology, zoo, credit approval*, dan *satlog*

heart. Prediksi *dispersion ratio* dataset *average link* unggul pada 3 dataset juga yaitu *ecoli, dermatology,* dan *zoo,* dan *dispersion ratio* dataset *single link* hanya unggul pada 2 dataset *ecoli* dan *credit approval*. Sedangkan prediksi dengan *information gain* unggul pada dataset *ecoli* dengan metode cluster *single link dan average link,* pada dataset *adult* dengan cluster *single link,* dan dataset *bank marketing* dengan *cluster complete link.*

Selisih predisksi terbesar terdapat pada dataset *dermatology* dengan rata-rata selisih adalah 24.35 %. Prediksi *dispersion ratio* yang tidak bergantung pada distribusi kelas mempengaruhi nilai prediksi. Hal tersebut terjadi karena, dataset *dermatology* memiliki multi class label yang *imbalance* dengan rincian ada 6 kelas label, yaitu kelas *psoriasis* (112), *seboreic dermatitis* (61), *lichen planus* (72), *pityriasis rosea* (49), *cronic dermatitis* (52), dan *pityriasis rubra pilaris* (20). Sama seperti data *dermatology,* data *ecoli* dan *zoo* juga memilki multikelas label yang *imbalance* dan prediksi dengan metode *dispersion ratio* lebih baik daripada dengan *information gain.*

Tabel 4.11 Klasifikasi dengan nilai *k*=3

Dataset	*Dispersion ratio*			*Information gain*		
	single	*average*	*complete*	*single*	*average*	*complete*
Ecoli	**95.83**	**95.83**	93.75	**95.83**	**95.83**	91.83
Adult	79.14	78.78	78.21	**79.96**	79.83	79.19
Dermatology	80.76	**90.38**	**90.38**	57.69	65.38	65.38
Bank Marketing	81.86	82.01	81.26	82.81	83.43	**84.21**
Zoo	92.85	**93.33**	**93.33**	92.85	92.85	92.85
Credit Approval	**94.89**	93.87	**94.89**	77.55	77.55	77.55
Satlog Heart	84.61	76.92	**89.47**	87.17	76.92	87.17

Tabel 4.12 Rata-rata selisih prediksi DR dan IG

Cluster	*Single link*	*Average link*	*Complete Link*
K=2	4.65	0.09	1.38
K=3	5.15	5.62	6.15

Tabel 4.13 jumlah prediksi tertinggi

cluster	*Dispersion Ratio*			*Information Gain*		
	Single	*Average*	*Complete*	*Single*	*Average*	*Complete*
K=2	4	1	2	1	1	1
K=3	2	3	4	2	1	1

Selisih rata-rata untuk prediksi *dispersion ratio* dengan *information gain* sesuai pada Tabel 4.12, dan jumlah prediksi tertinggi per dataset sesuai pada Tabel 4.13. Kedua table tersebut menggambarkan bahwa prediksi gabungan antara diskritisasi data menggunakan *hierarchical clustering* dan *splitting* atribut DT dengan *dispersion ratio* menghasilkan nilai prediksi yang lebih baik daripada *splitting* atribut menggunakan *information gain*. Pada dataset dengan *cluster* 2, *single link* memiliki selih rata-rata terbesar dan jumlah prediksi tertinggi yaitu unggul pada 4 dataset. Sedangkan untuk dataset dengan cluster 3, *complete link* mimiliki selisih rata-rata tertinggi dan jumlah prediksi per dataset, 4 kali lebih tinggi.

Prediksi yang dilakukan dengan metode *dispersion ratio* mempunyai kekurangan ketika dataset memiliki jumlah *instance* besar seperti dataset *adult* dan *bank marketing*. Jumlah *instance* pada dataset *adult* 48842 dan dataset *bank marketing* 4521. Pada dataset yang memiliki jumlah instance kurang dari seribu seperti *ecoli, dermatology, zoo, credit approval* dan *satlog heart,* prediksi dengan metode *dispersion ratio* lebih baik dari pada *information gain.* Selain jumlah *instance* yang mempengaruhi prediksi, dataset dengan *binary class* yang *imbalance* juga mengakibatkan predisksi dengan *dispersion ratio* lebih rendah dibandingkan dengan *information gain*. Contohnya, dataset *adult* dengan distribusi kelas positif

12210 dan kelas negatif 36632, serta dataset *bank marketing* dengan distribusi kelas positif 521 dan kelas negatif 4000.

[*Halaman ini sengaja dikosongkan*]

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat ditarik pada penelitian ini adalah sebagai berikut :

1. Diskrtitisasi *hierarchical clustering* sangat cocok untuk dataset yang memiliki atribut numerik dengan interval data yang tinggi, seperti dataset *bank marketing* yang memiliki selisih predisksi 28.08 % dan dataset *adult* dengan selisih prediksi 43.02 % .
2. Prediksi menggunakan *dispersion ratio* untuk dataset yang multikelas dan *imbalance* seperti dataset *dermatology, zoo,* dan *ecoli* lebih unggul daripada menggunakan metode *information gain*.
3. Metode *decision tree* yang menggabungkan proses *splitting* menggunakan dispersion ratio dan diskritisasi data menggunakan *hierarchical clustering complete link* dengan jumlah *cluster* 3 menghasilkan perbaikan prediksi rata-rata terbesar yaitu 6.15 %.
4. Prediksi *decision tree* menggunakan metode *dispersion ratio* memiliki kelemahan ketika dataset memilki *binary class* yang *imbalance.*

5.2 Saran

Saran yang dapat diberikan dari hasil uji coba dan analisis yang telah dilakukan yaitu :

a. Studi lebih lanjut mengenai *decision tree* dengan *dispesion ratio* pada dataset *binary class* yang *imbalance*.
b. Pengembangan metode *hierarchical clustering* untuk diskritisasi data numerik.

[*Halaman ini sengaja dikosongkan*]

DAFTAR PUSTAKA

Cheng, S., & Pecht, M. (2012). Using cross-validation for model parameter selection of sequential probability ratio test. *Expert Systems with Applications*, *39*(9), 8467–8473. https://doi.org/10.1016/j.eswa.2012.01.172

Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology*, *June*.

Gama, J., & Pinto, C. (2014). Discretization from Data Streams : Applications to Histograms and Data Mining. *Symposium on Applied Computing*, *January 2006*. https://doi.org/10.1145/1141277.1141429

Herrera Semenets, V., Garcıa, O. A. P., Leon, R. H., Berg, J. van den, & Doerr, C. (2017). A Data Reduction Strategy and its Application on Scan and Backscatter Detection Using Rule-based Classifier. *Expert Systems With Applications*. https://doi.org/10.1016/j.eswa.2017.11.041

Horng, S., Yang, F., & Lin, S. (2011). Expert Systems with Applications Hierarchical fuzzy clustering decision tree for classifying recipes of ion implanter. *Expert Systems With Applications*, *38*(1), 933–940. https://doi.org/10.1016/j.eswa.2010.07.076

Jafarzadegan, M., Safi-esfahani, F., & Beheshti, Z. (2019). Combining hierarchical clustering approaches using the PCA method. *Expert Systems With Applications*, *137*, 1–10. https://doi.org/10.1016/j.eswa.2019.06.064

Jiang, S., Li, X., Zheng, Q., & Wang, L. (2009). Approximate Equal Frequency Discretization Method. *2009 WRI Global Congress on Intelligent Systems*, *3*(May), 514–518. https://doi.org/10.1109/GCIS.2009.131

Kindhi, B. Al, Sardjono, T. A., Purnomo, M. H., & Verkerke, G. J. (2018). Hybrid K-Means, Fuzzy C-Means, and Hierarchical Clustering for DNA Hepatitis C Virus Trend Mutation Analysis. *Expert Systems With Applications*. https://doi.org/10.1016/j.eswa.2018.12.019

Li, A., Kumar, A., Ha, Y., & Corporaal, H. (2015). Microprocessors and

Microsystems Correlation ratio based volume image registration on GPUs. *Microprocessors and Microsystems*, *39*(8), 998–1011. https://doi.org/10.1016/j.micpro.2015.04.002

Lustgarten, J. L., Visweswaran, S., Gopalakrishnan, V., & Cooper, G. F. (2011). *Application of an efficient Bayesian discretization method to biomedical data*.

Marsland, S. (2015). *MACHINE A LEARNING An Algorithmic Perspective Second Edition*. CRC Press.

Maslove, D. M., Podchiyska, T., & Lowe, H. J. (2013). *Discretization of continuous features in clinical datasets*. 544–553. https://doi.org/10.1136/amiajnl-2012-000929

Mouthami, M. K. (2013). Sentiment Analysis and Classification Based On Textual Reviews. *International Conference on Information Communication and Embedded Systems (ICICES)*.

Muhlenbach, F., Rakotomalala, R., Muhlenbach, F., Rakotomalala, R., Attributes, C., & Wang, J. (2009). *Discretization of Continuous Attributes*. 397–402.

Müller, A. C., & Guido, S. (2015). Introduction to Machine Learning with Python and Scikit-Learn. In *O'Reilly Media, Inc.* http://kukuruku.co/hub/python/introduction-to-machine-learning-with-python-andscikit-learn

Pandya, R. (2015). *C5 . 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning*. *117*(16), 18–21.

Putra, J. W. G. (2019). *Pengenalan Pembelajaran Mesin dan Deep Learning*. *July*.

Ros, F., & Guillaume, S. (2019). *A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise*. *128*, 96–108. https://doi.org/10.1016/j.eswa.2019.03.031

Roy, S. (2016). CRDT : Correlation Ratio Based Decision Tree Model for Healthcare Data Mining. *IEEE 16th International Conference on Bioinformatics and Bioengineering CRDT:* https://doi.org/10.1109/BIBE.2016.21

Roy, S., Mondal, S., Ekbal, A., Sankar, M., & Felix, D. (2019). Dispersion Ratio based Decision Tree Model for Classification. *Expert Systems With Applications*, *116*, 1–9. https://doi.org/10.1016/j.eswa.2018.08.039

Rutkowski, L., Pietruczuk, L., Duda, P., & Jaworski, M. (2013). *Decision Trees for Mining Data Streams Based on the McDiarmid ' s Bound*. *25*(6), 1272–1279.

Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). *An Improved Feature Selection Based on Effective Range for Classification*. *2014*.

Xu, E., Liangshan, S., Yongchang, R., Hao, W., & Feng, Q. (2010). *2010 Asia-Pacific Conference on Wearable Computing Systems A New Discretization Approach of Continuous Attributes*. 141–143. https://doi.org/10.1109/APWCS.2010.40

[*Halaman ini sengaja dikosongkan*]

LAMPIRAN

Lampiran Dataset

Data	Atribut		Total *Instance*	*Class*
	Numerik	Nominal		
ecoli	7 (mcg, gvh, lip, chg, aac, alm1, alm2)	-	336	cp (143), im (77), pp (52), imU (35), om (20), omL (5), imL(2), imS (2)
adult	6 *(age, fnlwgt, education-num, capital-gain, capital-loss*, dan *hours-per-week)*	8 *(workclass, education, marital-status, occupation, relationship, race, sex*, dan *native-country)*	48842	">50K" (12210) dan "<=50K" (36632).
dermatology	33 (*erythema, scaling, definite borders, itching, koebner, polygonal, follicular, oral mucosal, knee and elbow , scalp, age, melanin, eosinophils, PNL*	1 *(family history)*	366	*psoriasis* (112), *seboreic dermatitis* (61), *lichen planus* (72), *pityriasis rosea* (49), *cronic dermatitis* (52), dan

	infiltrate, fibrosis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing, elongation, suprapapillary epidermis, spongiform pustule, munro microabcess, focal hypergranulosis, granular layer, vacuolisation, spongiosis, saw-tooth, follicular horn plug, perifollicular parakeratosis, inflammatory monoluclear, dan *band-like infiltrate*)			*pityriasis rubra pilaris* (20)
bank marketing	8 *(age, balance, day, duration, campaign, pdays,* dan *poutcome)*	9 (*job, marital, education, default, housing, loan, contact,*	4521	"yes" (521) dan "no" (4000)

		month, dan *poutcome*)		
zoo	1 (*legs)*	14 (*hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic*, dan *catsize*)	101	tipe satu(41), tipe dua(20), tipe tiga(5), tipe empat(13), tipe lima(4), tipe enam(8), dan tipe tujuh(10)
credit approval	6 (A2, A3, A8, A11, A14 dan A15)	9 (A1, A4, A5, A6, A7, A9, A10, A12 dan A13)	690	"+"(307) dan "-"(383)
statlog heart	7 (*age, resting blood pressure, serum cholestoral in mg/dl, maximum heart rate, oldpeak, major vessels*, dan *the slope of the peak)*	6 (*sex, fasting blood sugar > 120 mg/dl, exercise induced angina, chest pain type, resting electrocardiographic*, dan *thal*)	270	*absence*(151) dan *presence*(119)