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

Dimas Ari Setyawan[1)] dan Chastine Fathicah[2)]
[1, 2)] Departemen Teknik Informatika, Institut Sepuluh Nopember
Surabaya, Indonesia 60111
e-mail: dimas.ariawan16@gmail.com[1)], chastine@if.its.ac.id[2)]

ABSTRAK

Penelitian ini mengembangkan metode klasifikasi decision tree dengan splitting atribut menggunakan dispersion ratio yang memiliki keunggulan pada dataset yang imbalance, serta penggunaan hierarchical clustering pada tahap diskritisasi data numerik. Hal ini terjadi karena information gain sebagai metode splitting mempunyai kelemahan ketika dataset memiliki distribusi kelas yang imbalance. Metode hierarchical clustering digunakan pada tahap diskritisasi data numerik bertujuan untuk memperoleh cluster data yang seimbang, dibandingkan dengan binary split. Ada dua tahap pada penelitian ini yaitu, pertama data yang bertipe numerik akan dilakukan diskritisasi menggunakan hierarchical clustering dengan 3 metode yaitu single link, complete link, dan average link. Kedua, data hasil diskritisasi akan digabung kembali kemudian dilakukan pem-bentukan tree dengan splitting atribut menggunakan dispersion ratio dan di evaluasi dengan 7-fold cross validation. Data yang digunakan pada penelitian ini diambil dari UCI machine learning repository sejumlah 7 dataset, yaitu ecoli, adult, dermatology, bank marketing, zoo, credit approval, dan satlog heart. Hasil yang diperoleh menunjukakan bahwa diskritisasi data dengan hierarchical clustering dapat meningkatkan prediksi sebesar 14,6% dibandingkan dengan data tanpa diskritisasi. Proses splitting atribut dengan dispersion ratio dari data hasil diskritisasi hierarchical clustering dapat meningkatkan prediksi sebesar 6,15 %.

***Kata kunci**: Decision tree, dispersion ratio, hierarchical clustering.*

ENHANCEMENT OF DECISION TREE METHOD BASED ON HIERARCHICAL CLUSTERING AND DISPERSION RATIO

Dimas Ari Setyawan[1)] and Chastine Fathicah[2)]
[1, 2)] Department of Informatics, Institut Sepuluh Nopember
Surabaya, Indonesia 60111
e-mail: dimas.ariawan16@gmail.com[1)], chastine@if.its.ac.id[2)]

ABSTRACT

This research developed a decision tree classification method with splitting attributes using a dispersion ratio that has advantages in imbalance datasets, as well as the use of hierarchical clustering at the stage of numerical data discretization. This happens because, information gain as a splitting method has a weakness when the dataset has a imbalance class distribution. The hierarchical clustering method is used at the stage of numerical data discretization aimed to obtain a balanced data cluster, compared to binary splits. There are two stages in this research, first the numeric type data will be discretized using hierarchical clustering with 3 methods, single link, complete link, and average link. Second, the discretized data will be merged again then the formation of a tree with splitting attributes using dispersion ratio and evaluated with 7-fold cross validation. The data used in this study were taken from a 7 dataset UCI machine learning repository, ecoli, adult, dermatology, bank marketing, zoo, credit approval, and satlog heart. The results obtained show that data discretization by hierarchical clustering can increase predictions by 14.6% compared to data without discretization. The splitting attribute process with the dispersion ratio of the data resulting from hierarchical clustering discretization can increase the prediction by 6.15%.

***Keywords**: Decision tree, dispersion ratio, hierarchical clustering.*

I. PENDAHULUAN

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 [1]. Klasifikasi merupakan salah satu jenis dari *machine learning* dengan mererapkan metode *supervised learning*. Klasifikasi yang ada yaitu *decision tree algorithm, naïve bayes classifier, neural network, k-nearest neighbour*, *support vector machine*, dan lain sebagainya [2]. Menurut [3] *decision tree* merupakan metode klasifikasi yang memiliki proses seleksi fitur. *Model*

DOI: http://dx.doi.org/10.12962/j24068535.v18i2.a1005

tree yang terbentuk biasanya menggunakan fungsi *Information Gain* atau IG untuk memisahkan antar fitur [4]. *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 [5]. IG juga bergantung pada distribusi kelas sehingga jika kelas *imbalance* maka nilai *true positive* dan *false positive* sama [2].

Kekurangan pada *decision tree* dengan IG dapat diatasi dengan konsep seleksi fitur yang signifikan. Hal ini telah dilakukan oleh [5] dengan menggunkan *Correlation Ratio* pada proses pemisahan fitur pada data kesehatan. Penggunaan correlation ratio masih memiliki kekurangan untuk data yang *non-liner* [6]. Pada penelitian lainnya [2] 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 dibutuhkan diskritisasi untuk memperoleh interval data [7]. Selain itu, di dunia nyata data sering kali bertipe numerik sehingga untuk merubah fitur menjadi diskrit menggunakan prosedur preprosesing data yaitu diskritisas. Menurut [8] diskritisasi juga mengurangi kebutuhan *storage* pada sistem, mempercepat proses *mining* data dan meningkatkan akurasi dari klasifikasi.

Pada penelitian [5] dan [2] 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 [9]. Diskritisasi dengan *k-means* mampu menangani nilai batas yang lebih baik dibandingkan dengan *equal interval* biasa.Diskritisasi dengan k-*means* sangat tergantung dengan nilai *k* dan inisalisasi *centroid* awal. 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 [2], 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 [10]. *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* [3]. Sehingga menghasilkan *cluster* yang lebih konstan dan tidak dipengaruhi oleh inisialiasasi awal seperti *k-means* [11].

Penelitian ini menggunakan *hierarchical clustering* untuk diskritisasi data yang beratribut numerik. Bertujuan untuk mengurangi kelemahan *cluster* yang selalu mengangap ukuran dan densitas sama pada diskritisasi *k-means*. Hasil *clustering* dengan *hierarchical clustering* membentuk *cluster* lebih fleksibel bisa memanjang. Penggunaan *Information Gain* atau *Gini Index* menggasilkan model yang memiliki bias tertentu karena tergantung pada distribusi kelas, sehingga mengakibatkan fitur yang digunakan tidak terlalu signifikan. *Dispersion ratio* yang tergantung pada distribusi frekuensi dapat melakukan pemiliihan fitur yang signifikan terhadap model yang akan dibentuk.

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 dan dibentuk dengan metode *dispersion ratio*.

II. KAJIAN PUSTAKA

A. Decision Tree

Decision tree merupakan salah satu pembelajaran yang merepresentasikan pengentahuan dalam bentuk aturan (*rule*) klasifikasi [4][12]. 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 didalam proses pembentukannya. Menurut [13] algoritma *decision tree* 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* menggunkan konsep pembentukan pohon keputusan yang berulang, dimana *parent* dari pohon akan melukukan *splitting root* dan membentuk *leaf tree* sampai *leaf* tidak mempunyai cabang [14]. Atribut yang sudah dipilih sebagai *leaf* atau *parent* tidak akan dicoba pada proses

percabangan di cabang tertentu. Seperti contoh pada Gambar 1 dan Tabel I, *decision tree* yang dibuat dengan algortima C4.5 yang menggunkan perhitungan *information gain* menghasilkan *decision tree play tennis* seperti Gambar 1. *Association rule pada play tennis* menggambarkan salah satu keunggulan dari penggunaan klasifikasi dengan DT. Dimana fitur *temperature* tidak diperhitungkan dalam kemungkinan *play tennis* yang berarti sudah ada proses seleksi fitur yang terjadi ketika pembentukan *rule*. Dari *rule* tersebut dapat diketehui bahwa kemungkinan permainan *tennis* bisa terjadi ketika:

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

Pada umumnya metode *splitting* pada *decision tree* menggunakan algoritma *Information Gain* (IG) dimana atribut yang memiliki niali IG tertinggi akan dipilih. Perhitungan *splitting* atribut *decision tree* dengan IG menggunkan perhitungan *gain entropy* seperti persamaan (1) dan (2). Nilai $p(j|t)$ pada persamaan (1) merupakan frekuensi kelas j didalam atribut t. Pada persamaan (2) nilai $Entropy(p)$ merupakan nilai entropi dari *parent node* dan nilai k merupakan partisi dari *split* atribut p.

$$Entropi\ (t) = -\sum_{t} p(j|t) \log p(j|t) \tag{1}$$

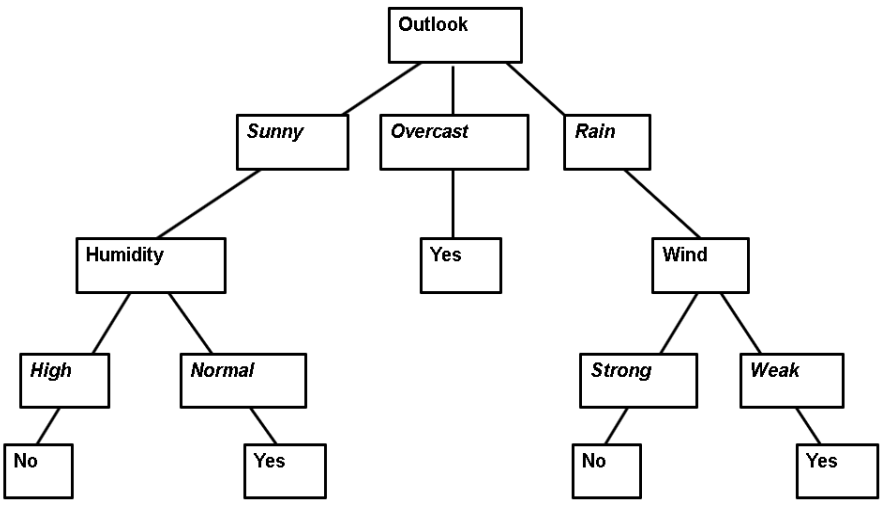
$$GAIN_{split} = Entropi(p) - \sum_{i=1}^{k} \frac{n_i}{n} * Entropi\ (i) \tag{2}$$

B. Hierarchical Clustering

Hierarchical clustering adalah metode pengelompokan data ke dalam bentuk pohon hierarki yang disebut *dendrogram*. *Hierarchical clustering* membentuk *cluster* yang terbawah atau *root node* sebagai *cluster* setiap data. Sedangkan *cluster* diatasnya dibentuk oleh kedekatan antar data dibawahnya. Menurut [9][15], ada dua cara untuk membentuk *hierarchical clustering* yaitu *divisive* dan *agglomerative*.

TABEL I
DATA "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



Gambar 1. *Decision tree* hasil dari perhitungan C4.5 pada data Tabel I.

Divisive atau *top-down* merupakan metode memecah data dimulai dengan satu *cluster* yang berisi semua data kemudian dipecah menjadi beberapa *cluster-cluster* lainnya [16]. *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).

$$single\ d_{uv} = \max\{d_{uv}\}, d_{uv} \in D \quad (3)$$

$$complate\ d_{uv} = \min\{d_{uv}\}, d_{uv} \in D \quad (4)$$

$$average\ d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D \quad (5)$$

$$D(a,b) = \sqrt{\sum_{i=0}^{n} (b_i - a_i)^2} \quad (6)$$

Perhitungan tingkat kemiripan dengan *single link* dihitung berdasarkan nilai kemiripan terbesar $\max\{d_{uv}\}$ diantara anggota *cluster* seperti persamaan (3). *complete link* merupakan perhitungan tingkat kemiripan data berdasarkan nilai terkecil $\min\{d_{uv}\}$ diantara anggota cluster seperti persamaan (4). Sedangkan *average link* merupakan perhitungan tingkat kemiripan data berdasarkan jarak rata-rata $\text{average}\{d_{uv}\}$ diantara anggota cluster seperti persamaan (5). Menghitung jarak antar data pada metode *hierarchical clustering* menggunakan algoritma *euclidean distance* (6) dimana nilai jarak *D* merupakan akar kuadrat dari jumlah selisih $(b_i - a_i)^2$.

C. Dispersion Ratio

Penggunaan *dispersion ratio* merupakan perbaikan metode dari *correlation ratio* pada penelitian [5][2], 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.

Persamaan (7) 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.

$$DR_i = \sqrt{\frac{\sum_{y \in Y} n_y \left(\bar{m}_y^{(i)} - \bar{m}^{(i)}\right)^2}{\sum_{y \in Y} \sum_{j=1}^{y} \left(\bar{v}_{jy}^{(i)} - \bar{m}^{(i)}\right)^2}} \quad (7)$$

D. Dataset

Data pada penelitian ini menggunakan dataset yang telah disediakan oleh UCI *machine learning repository.* Data yang digunakan merupakan data yang sama seperti pada penelitian [2] 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 II merupakan daftar data yang digunakan pada penelitian ini. Jenis data yang digunakan beragam seperti data *real, categorical*, dan *integer.*

Pada penelitian sebelumnya [2] 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 numeric semua, sedangkan data yang lain atributnya bertipe gabungan antara numerik dan nominal. Data *dermatology* dan *credit approval* memiliki beberapa *missing value* pada data.

III. METODE YANG DIAJUKAN

Diagram alir perancangan metode pada klasifikasi dengan *decision tree* terdiri dari beberapa tahap seperti Gambar 2. Tahap pertama, data training dipisah menjadi 2 bagian yaitu data yang bertipe numerik dan data yang bertipe

nominal. Tahap kedua, proses diskritisasi data yang bertipe numerik dengan metode *hierarchical clustering*. Tahap ketiga, merupakan data hasil *clustering* dengan 3 metode *single link, complete link*, dan *average link.* Tahap keempat, melakukan *merge* data numerik hasil *clustering* dengan data nominal. Tahap kelima, melakukan *splitting* data hasil *merger* dengan metode *dispersion ratio*. Tahap terakhir, model *decision tree* yang terbentuk dilakukan evaluasi dan dibentuk *rule base* dari klasifikasi.

A. Diskritisasi

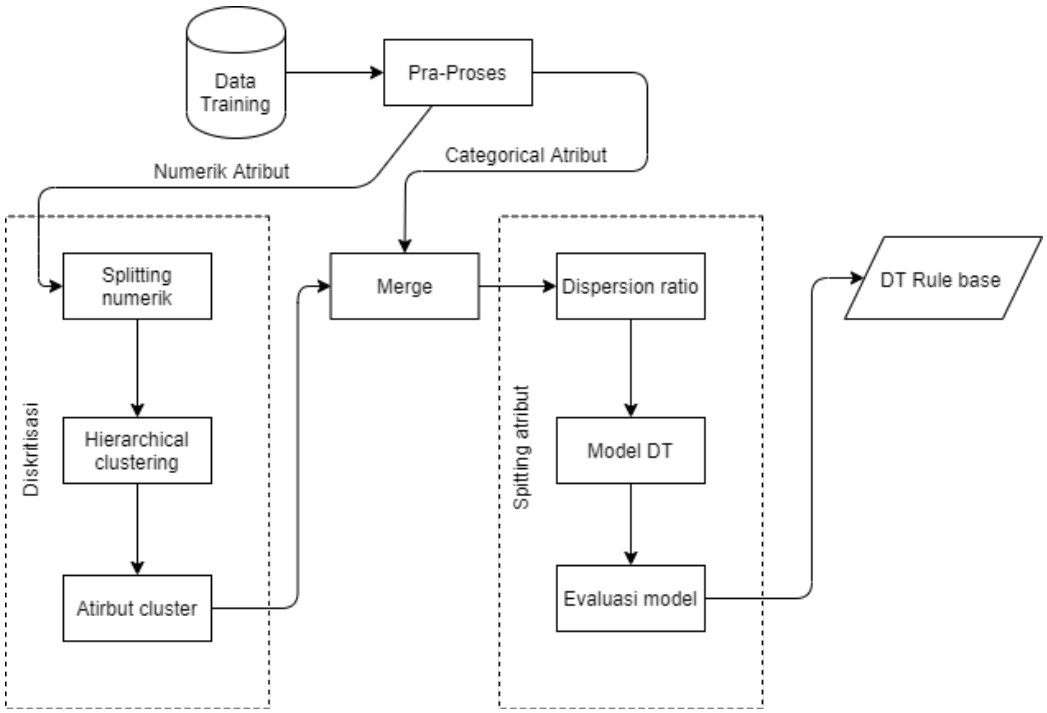
Tahap diskritisasi dilakukan untuk data yang bertipe numerik yang dijelaskan pada Gambar 3. Proses disktritisasi pada penelitian ini menggunakan metode *hierarchical clustering*. Proses *hierarchical clustering* dilakukan dengan cara *agglomerative (down-top)* jadi *cluster* akan dibentuk dari semua 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.

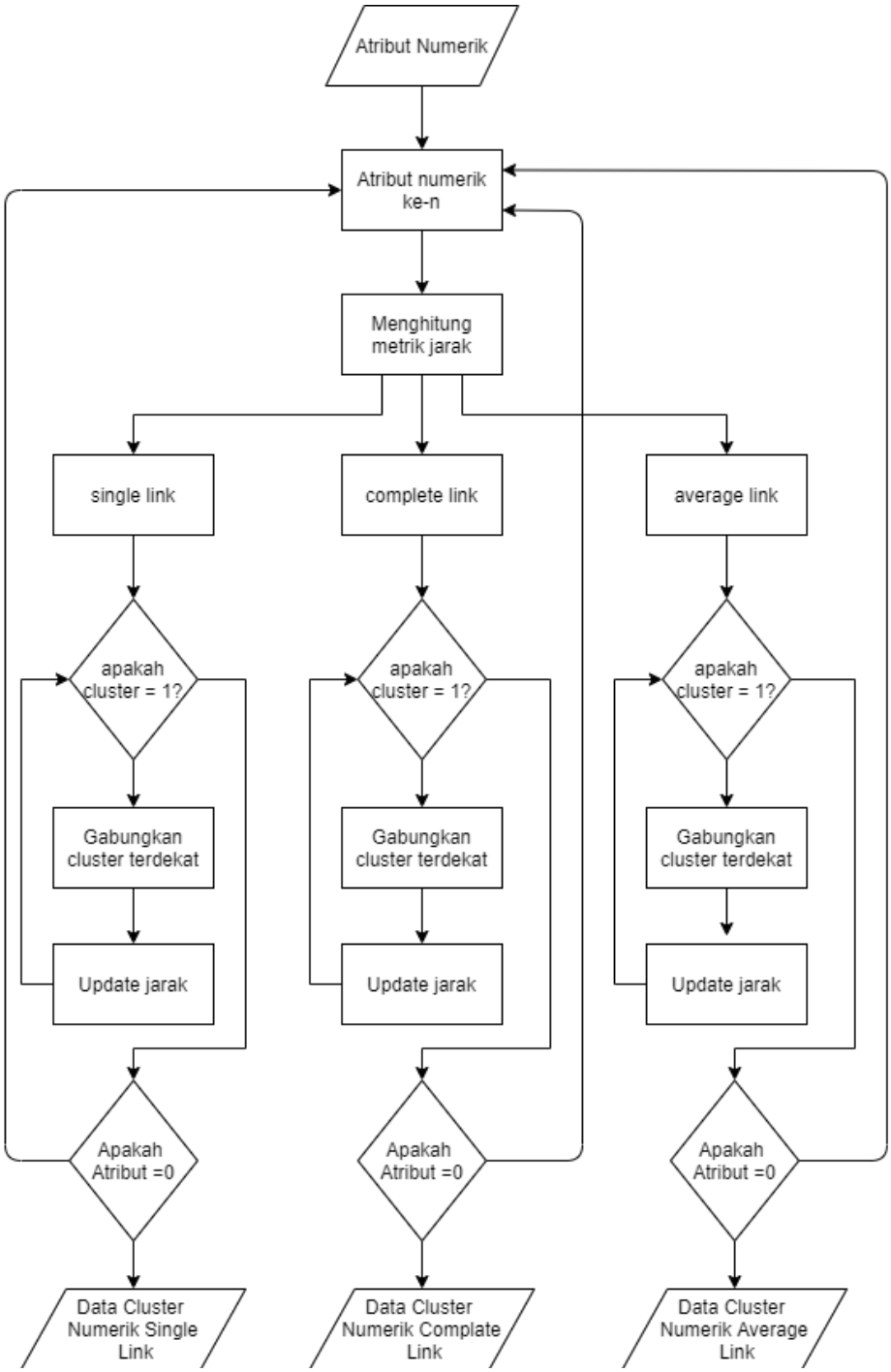
Setiap hasil *cluster* yang terbentuk dari masing-masing metode dijadikan *input* data pada tahap *merge*. Pada data *single link* dilakukan tahap *merge* dengan data nominal kemudian di hitung nilai *dispersion ratio* untuk membentuk *decision tree* dan dihitung akurasinya. Tahap tersebut juga dilakukan 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*.

TABEL II
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)



Gambar 2. Diagram alir metode penelitian.



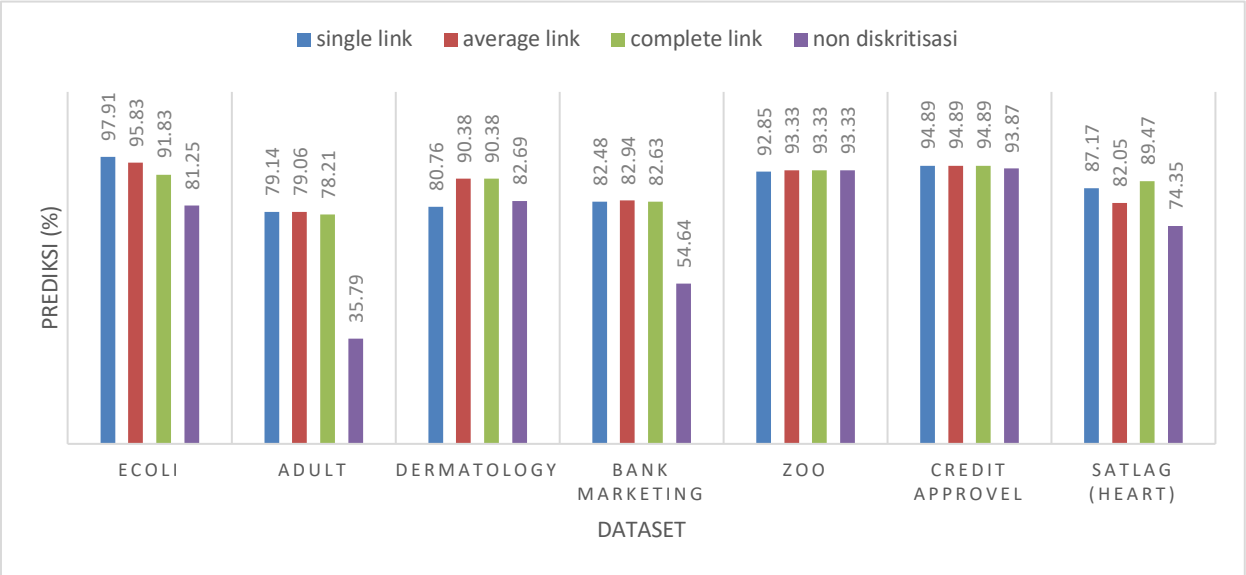
Gambar 3. Diskritisasi dengan *hierarchical clustering.*

B. Splitting 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 III yang diambil dari penelitian [2] telah menunjukkan 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 III
ALGORITMA SPLITTING DECISION TREE.

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**



Gambar 4. Diagram prediksi *decision tree* dengan disktritisasi dan tanpa diskritisasi.

IV. HASIL DAN PEMBAHASAN

Pada bab ini akan memapaparkan hasil beserta pembahasaan pada ujicoba dataset, dengan beberapa kriteria yang berbeda.

A. Hasil Prediksi dengan Diskritisasi dan Tanpa Diskritisasi

Prediksi klasifikasi model *decision tree* seperti Gambar 4 menjelaskan bahwa data hasil diskritisasi dengan *hierarchical clustering* nilainya lebih baik ketimbang data yang tidak dilakukan proses diskritisasi *hierarchical clustering*. Hal ini terjadi karena dataset dengan atribut yang bertipe numerik telah dilakukan proses *cluster* data sehingga lebih baik dalam perhitungan prediksi DT dengan *dispersion ratio.* 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 ujicoba 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*.

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.

B. Hasil Prediksi dengan Dispersion Ratio dan Information Gain

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 pada Tabel IV. 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 4.77, *complete link* dengan nilai 1.7 dan *average link*

dengan 0,29. Jadi prediksi *dispersion ratio* dataset *single link* unggul 5 data yaitu: *ecoli*, *dermatology*, *satlog heart* dan *credit approval*. Prediksi *dispersion ratio* dataset *average link* unggul 2 data, bank marketing dan credit approval, *dispersion ratio* dataset *complete link* unggul 2 data juga, *zoo* dan *credit approval*. Sedangkan prediski *information gain* hanya unggul pada dataset *adult* dengan *cluster average link* dan dataset *zoo* dengan cluster *single link.*

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 pada Tabel V. 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 4 dataset juga yaitu *ecoli, dermatology, bank marketing*, 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*, serta pada dataset *adult* dengan cluster *single link.*

V. KESIMPULAN

Prediksi dengan *dispersion ratio* dan diskritisasi *hierarchical clustering* sangat cocok untuk dataset yang memiliki atribut numerik dengan interval data yang tinggi seperti dataset *bank marketing* dan *adult*. Selain itu, prediksi *dispersion ratio* untuk dataset dengan *cluster* = 2 dan *cluster* = 3, juga lebih tinggi daripada menggunkan *information gain* dengan selisih rata-rata prediksi tertinggi adalah 6.51 % .

TABEL IV
HASIL PREDIKSI KLASIFIKASI DENGAN NILAI $K = 2$.

Dataset	*Dispersion*			*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	82.32	82.19	82.32
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

TABEL V
HASIL PREDIKSI KLASIFIKASI DENGAN NILAI $K = 3$.

Dataset	Dispersion			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	81.88	81.57	81.73
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

DAFTAR PUSTAKA

[1] V. Herrera Semenets, O. A. P. Garcıa, R. H. Leon, J. van den Berg, dan C. Doerr, "A Data Reduction Strategy and its Application on Scan and Backscatter Detection Using Rule-based Classifier," *Expert Syst. Appl.*, 2017, doi: 10.1016/j.eswa.2017.11.041

[2] S. Roy, S. Mondal, A. Ekbal, M. Sankar, dan D. Felix, "Dispersion Ratio based Decision Tree Model for Classification," *Expert Syst. Appl.*, vol. 116, hal. 1–9, 2019, doi: 10.1016/j.eswa.2018.08.039

[3] J. Wang, S. Zhou, Y. Yi, dan J. Kong, "An Improved Feature Selection Based on Effective Range for Classification," *Recent Advances in Information Technology*, vol. 2014, 2014.

[4] L. Rutkowski, L. Pietruczuk, P. Duda, dan M. Jaworski, "Decision Trees for Mining Data Streams Based on the McDiarmid's Bound," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, hal. 1272–1279, 2013.

[5] S. Roy, "CRDT : Correlation Ratio Based Decision Tree Model for Healthcare Data Mining," dalam *Proc. IEEE Int. Conf. Bioinforma. Bioeng.*, 2016, doi: 10.1109/BIBE.2016.21

[6] A. Li, A. Kumar, Y. Ha, dan H. Corporaal, "Microprocessors and Microsystems Correlation ratio based volume image registration on GPUs,"

Microprocess. Microsyst., vol. 39, no. 8, hal. 998–1011, 2015, doi: 10.1016/j.micpro.2015.04.002
[7] D. M. Maslove, T. Podchiyska, dan H. J. Lowe, "Discretization of continuous features in clinical datasets," *Journal of the Americal Medical Informatics Association*, hal. 544–553, 2013, doi: 10.1136/amiajnl-2012-000929
[8] E. Xu, S. Liangshan, R. Yongchang, W. Hao, dan Q. Feng, "A New Discretization Approach of Continuous Attributes," dalam *Proc. Asia-Pacific Conference on Wearable Computing Systems*, hal. 141–143, 2010, doi: 10.1109/APWCS.2010.40
[9] R. Dash, R. L. Paramguru, dan R. Dash, "Comparative Analysis of Supervised and Unsupervised Discretization Techniques," *Int. J. Adv. Sci. Technol.*, 2011.
[10] B. Al Kindhi, T. A. Sardjono, M. H. Purnomo, dan G. J. Verkerke, "Hybrid K-Means, Fuzzy C-Means, and Hierarchical Clustering for DNA Hepatitis C Virus Trend Mutation Analysis," *Expert Syst. Appl.*, 2018, doi: 10.1016/j.eswa.2018.12.019
[11] S. Horng, F. Yang, dan S. Lin, "Expert Systems with Applications Hierarchical fuzzy clustering decision tree for classifying recipes of ion implanter," *Expert Syst. Appl.*, vol. 38, no. 1, hal. 933–940, 2011, doi: 10.1016/j.eswa.2010.07.076
[12] M. K. Mouthami, "Sentiment Analysis and Classification Based On Textual Reviews," dalam *Proc. Int. Conf. Inf. Commun. Embed. Syst.*, 2013.
[13] R. Pandya, "C5 . 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *International Journal of Computer Applications*, vol. 117, hal. 16, pp. 18–21, 2015.
[14] S. Cheng dan M. Pecht, "Using cross-validation for model parameter selection of sequential probability ratio test," *Expert Syst. Appl.*, vol. 39, no. 9, hal. 8467–8473, 2012, doi: 10.1016/j.eswa.2012.01.172
[15] M. Jafarzadegan, F. Safi-esfahani, dan Z. Beheshti, "Combining hierarchical clustering approaches using the PCA method," *Expert Syst. Appl.*, vol. 137, hal. 1–10, 2019, doi: 10.1016/j.eswa.2019.06.064
[16] F. Ros hal S. Guillaume, "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Systems with Applications*, vol. 128, hal. 96–108, 2019, doi: 10.1016/j.eswa.2019.03.031