Klasifikasi Penentuan Jenis Obat Menggunakan Algoritma *Decision Tree*

Rika Nursyahfitri[1], Alfanda Novebrian Maharadja[2], Riva Arsyad Farissa[3], Yuyun Umaidah[4]

[1,2,3,4] Teknik Informatika, Ilmu Komputer, Universitas Singaperbangsa Karawang
[1]rika.nursyahfitri17181@student.unsika.ac.id, [2]alfanda.maharadja17005@student.unsika.ac.id,
[3]riva.arsyad17184@student.unsika.ac.id, [4]yuyun.umaidah@staff.unsika.ac.id

Abstrak

Data Mining merupakan suatu proses untuk mengidentifikasi informasi dan pengetahuan yang bermanfaat. Klasifikasi merupakan salah satu teknik *data mining* yang dapat digunakan untuk prediksi, dimana nilai yang diprediksi berupa label. Klasifikasi penentuan jenis obat bertujuan untuk memprediksi jenis obat yang tepat untuk pasien dengan menganalisis *dataset* yang telah diperoleh. Data yang digunakan pada penelitian ini adalah data hasil catatan medis pasien berdasarkan gejala penyakit yang diderita namun belum diketahui jenis obatnya. *Dataset* yang digunakan merupakan data sekunder, yang berasal dari kaggle.com. Data terdiri dari 200 *record* dengan 6 variabel (Usia, Jenis Kelamin, Tingkat Tekanan Darah, Tingkat Kolesterol, Na to K dan Jenis Obat) dimana 5 variabel sebagai *predictor* dan 1 variabel sebagai *class* target. Data kemudian dipresentasikan kedalam bentuk pohon keputusan dengan suatu model matematis menggunakan bahasa pemograman R. Untuk menyelesaikan permasalahan, maka digunakan sebuah metode klasifikasi dalam *data mining* yaitu *decision tree* C4.5. Algoritma C4.5 digunakan untuk menemukan hubungan antara calon sejumlah variabel, sehingga menjadi sebuah variabel target klasifikasi dengan pembagian data menjadi 2 yaitu 70% data *training* dan 30% data *testing.* Hasil pengujian yang diperoleh pada penelitian ini berupa aturan dan tingkat nilai *accuracy* sebesar 100%, sehingga dapat dsimpulkan kinerja algoritma C4.5 dinilai sangat baik dalam memprediksi jenis obat.

Kata kunci : klasifikasi, jenis obat, *decision tree, confusion matrix,* C4.5

1. Pendahuluan

Data mining merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Bramer, 2007). *Data mining* ditujukan untuk mengekstrak pengetahuan yang berguna dengan berfokus pada algoritma, sehingga menemukan pola data baru dalam database (Ente, D.R , et al., 2020).

Klasifikasi merupakan salah satu teknik *data mining* yang dapat digunakan untuk memprediksi, dimana nilai yang diprediksi berupa label (variabel target). Menurut Supriyanti, W, Kusrini, Amborowati, A (2016) bahwa klasifikasi merupakan suatu teknik untuk menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. *Decision tree* merupakan salah satu teknik klasifikasi yang paling populer karena dapat menghasilkan prediksi yang sangat kuat (Kusrini & Luthfi, 2009). *Decision tree* menggunakan representasi struktur pohon, dimana setiap node merepresentasikan atribut, cabang merepresentasikan nilai atribut, dan node merepresentasikan kelas. Node di bagian atas *decision tree* disebut root. *Decision Tree* adalah metode klasifikasi yang paling populer. Selain perkembangan yang relatif cepat, hasil model yang dibangun juga mudah dipahami. *Decision Tree* adalah metode klasifikasi yang paling populer. Selain perkembangan yang relatif cepat, hasil model yang dibangun juga mudah dipahami. Didalam *decision tree* terdapat beberapa algoritma yang dapat digunakan dalam pembentukan pohon keputusan, antara lain ID3 (*Iterative Dichotomiser 3),* CART (*Classification and Regression Tree*) dan C4.5.

Penelitian terkait implementasi *decision tree* telah banyak dilakukan sebelumnya, seperti yang telah dilakukan oleh Noviandi (2018) untuk memprediksi terhadap wanita yang telah melahirkan dengan melihat beberapa faktor lainnya, apakah pasien menderita penyakit diabetes atau tidak. Hasil eksperimen menujukkan bahwa algoritma *decision tree* C4.5 memiliki tingkat akurasi sebesar 70.32% dengan menghasilkan 9 *rule*. Penelitian lain telah dilakukan oleh Santosa, I, Rosiyah, H & Rahmanita, E (2018) untuk mendiagnosa penyakit Tuberkulosis (TB). Hasil dari penelitian ini berupa sistem yang dapat membantu masyarakat dalam mendiagnosa penyakit TB dan memperoleh hasil tingkat nilai akurasi sebesar 90% dengan menggunakan evaluasi *confusion matrix.* Pada penelitian lain juga telah dilakukan oleh Raharjo, M et al., (2019) untuk memprediksi peminatan jurusan robotika menggunakan metode *decision tree* menghasilkan nilai akurasi algoritma klasifikasi C4.5 sebesar 84.14% dengan evaluasi menggunakan nilai AUC sebesar 0.887 dengan tingkat klasifikasi baik

Sedangkan penelitian terkait perbandingan algoritma klasifikasi telah dilakukan oleh Bahri, S, Midyanti, D. M & Hidayati, R (2018) mengenai klasifikasi penyakit anak untuk mendiagnosa penyakit pada anak menggunakan algortima C4.5 dan *Naive Bayes*. Hasil peneliatian menjelaskan bahwa tingkat nilai akurasi algoritma C4.5 lebih unggul sebesar 90.00% dibanding dengan algoritma *Naive Bayes* sebesar 89.58%. Selanjutnya penelitian yang telah dilakukan oleh Anam, C & Santoso, H. B (2018) dalam klasifikasi penerima beasiswa menggunakan algoritma C4.5 dan *Naive Bayes* dengan metode evaluasi 10-*fold cross validation*, menghasilkan bahwa tingkat nilai akurasi algoritma C4.5 sebesar 96.40% lebih baik dibanding *Naive Bayes* sebesar 95.11% dengan *time taken* kedua algoritma yaitu 0 s.

Berdasarkan penelitian-penelitian sebelumnya terkait metode *decision tree*, algoritma C4.5 memiliki *performance* dan hasil nilai akurasi yang cukup baik dalam klasifikasi dan prediksi data. Sehingga pada penelitian ini bertujuan untuk mengklasifikasikan jenis obat yang akurat untuk pasien dengan menggunakan metode *decision tree.*

2. Tinjauan Pustaka

2.1 Klasifikasi

Klasifikasi merupakan salah satu teknik pada *machine learning* yang digunakan pada proses *data mining*. Klasifikasi merupakan salah satu teknik penambangan data pembelajaran mesin klasik. Menurut Supriyanti, W, Kusrini, Amborowati, A (2016) bahwa klasifikasi merupakan suatu teknik menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.

Proses Klasifikasi didasarkan pada 4 komponen sebagai berikut (Gorunescu, 2011):

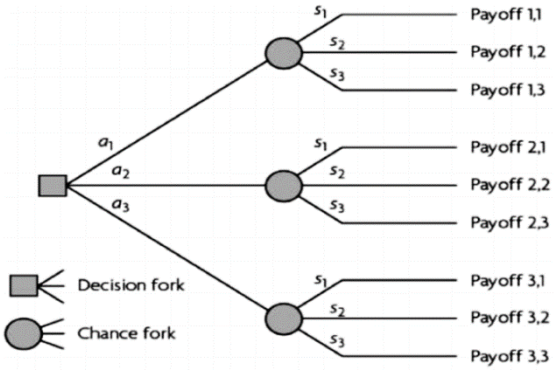
1. *Class*
   Class merupakan variabel tidak bebas yang berupa kategorial yang mempresentasikan label yang terdapat pada objek.
2. *Predictor*
   Predictor merupakan variabel bebas suatu model berdasarkan karakteristik atribut data kategorial.
3. *Training Dataset*
   Training dataset disebut juga sebagai data latih, merupakan *dataset* yang berisi nilai dari *class* dan *predictor* untuk dilatih agar model dapat dikelompokkan ke kelas yang benar.
4. *Testing Dataset*
   Testing dataset disebut juga sebagai data uji, merupakan data baru yang digunakan untuk mengklasifikasikan model yang dibuat dan mengevaluasi akurasi klasifikasi.

Pada teknik *data mining* dan *machine learning* terdapat beberapa model yang telah dikembangkan dalam metode klasifikasi untuk menyelesaikan permasalahan klasifikasi antara lain: (Mardi, Y, 2017)

1. Pohon Keputusan (*Decison Tree)*
2. Pengklasifikasi *Bayes* / *Naive Bayes*
3. Jaringan Saraf Tiruan
4. Analisis Statistik
5. Algoritma Genetik
6. *Rough Sets*
7. Pengklasifikasi *K-Nearest Neighbour*
8. Metode Berbasis Aturan
9. *Memory Based Reasoning*
10. *Support Vector Machine*

2.2 *Decision Tree*

Decision tree merupakan algoritma *supervised machine learning* yang digunakan untuk memecahkan masalah klasifikasi. Tujuan utama dari algoritma *decision tree*, karena mampu menghasilkan model prediksi secara spesifik dalam bentuk aturan yang mudah untuk diimplementasikan (Noviandi, 2018). Berikut pada Gambar 1 merupakan bentuk



decision tree secara umum.

Gambar 1 Bentuk *Decision Tree* secara umum
(Sumber: Kasih, P, 2019)

Berdasarkan gambar diatas, *decision tree* merupakan struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap simpul internal menandakan suatu pengujian pada atribut, setiap cabang merepresentasikan *output*, dan simpul daun merepresentasikan kelas atau distribusi kelas. Simpul yang paling atas disebut sebagai *Root node* yang memiliki beberapa *output* tetapi tidak memiliki *input*. Sedangkan internal *node* memiliki satu *input* dan beberapa *output*, dan leaf node hanya memiliki satu *input* tanpa memiliki *output. Leaf node* merupakan hasil akhir yang mewakili label kelas dari kombinasi atribut yang terbentuk menjadi *rule*. (Kasih, P, 2019).

2.3 Algoritma C4.5

Algoritma C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi

data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan – aturan dapat digunakan untuk memprediksi nilai atribut bertipe diskret dari record yang baru. Algortima C4.5 sendiri merupakan pengembangan dari algortima *decision tree*, dimana pengembangan dilakukan dalam hal, bisa mengatasi *missing* data, bisa mengatasi data kontinu dan pruning (Elisa, E, 2017)

Ada beberapa tahapan dalam membuat *decision tree* (pohon keputusan) C4.5 antara lain:

1. Input data
2. Tentukan atribut yang akan dijadikan akar dengan menentukan nilai *entropy* terendah dan nilai gain tertinggi. Untuk menentukan nilai gain, dapat dilihat pada bersamaan berikut.

$$Gain\ (S,A) = Entropy(S) - \sum_{i=0}^{n} \frac{|Si|}{|S|} * Entropy(S) \quad (1)$$

Keterangan:
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

Dimana, perhitungan nilai *entropy* dapat dilihat pada persamaan berikut.

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi \quad (2)$$

Keterangan:
S : himpunan kasus
n : jumlah partisi S
Pi : proporsi dari Si tehadap

Dari atribut sebagai akar yang didapat pada langkah awal dibuat cabang untuk tiap-tiap nilai.

3. Selanjutnya membagi kasus dalam tiap cabang.
4. Pada setiap cabang yang belum menunjukan pada suatu kelas tertentu, maka ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama dan proses selesai.

2.4 *Confusion Matrix*

Confusion Matrix merupakan salah satu alat ukur pada pembelajaran *supervised learning* berbentuk matrik yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi *dataset* terhadap kelas tepat dan tidak tepat pada algoritma yang digunakan (Santosa. I, Rosiyah. H & Rahmanita. E, 2018).

Confusion matrix terdiri dari dua jenis, *confusion matrix binary* dan *multiclass. Confusion matrix binary,* dimana terdapat klasifikasi dengan dua *output*, sedangkan *multiclass* terdapat klasifikasi dengan jumlah *output class* yang lebih dari dua (*multiple classes*). Untuk menghitung akurasi, *precision* dan *recall* dapat dilakukan dengan menghitung rata-rata pada setiap kelas. Berikut pada Tabel 1 terdapat contoh tabel *confusion matrix multiclass.*

Tabel 1 *Confusion Matrix Multiclass*

		True Class/Actual				
		A	B	C	X	Y
Predicted Class	A	TP_A	AB	AC	AX	AY
	B	BA	TP_B	BC	BX	BY
	C	CA	CB	TP_C	CX	CY
	X	XA	XB	XC	TP_X	XY
	Y	YA	YB	YC	YX	TP_Y

Dari tabel diatas *confusion matrix* dapat diperoleh nilai *Recall, Precision* dan *Acuuracy.* Untuk mencari nilai akurasi dapat dihitung dengan menggunakan rumus berikut:

$$Acuuracy = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l} * 100\% \quad (3)$$

Accuracy merupakan proporsi kasus yang diidentifikasi benar terhadap jumlah semua kasus.

$$Recall = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FN_i)} * 100\% \quad (4)$$

Recall merupakan proporsi kasus positif yang diidentifikasi dengan benar.

$$Precision = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (FP_i + TN_i)} * 100\% \quad (5)$$

Precision merupakan proporsi kasus dengan hasil positif yang benar.

Keterangan:
TP_i = *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.
TN_i = *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.
FN_i = *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem untuk kelas ke-i.
FP_i = *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem untuk kelas ke-i
l = jumlah kelas.

2.5 *Cross Validation*

Cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan melakukan perulangan dalam mengacak atribut masukan. *K-fold Cross Validation* merupakan salah satu metode validasi algoritma dengan membagi data menjadi *k-fold*, dimana *k*-1 buah *fold* digunakan sebagai data *testing* dan 1 buah *fold* digunakan sebagai data *training*. *K-fold Cross Validation* digunakan karena dapat mengurangi waktu komputasi dengan tetap menjaga keakuratan estimasi (Widaningsih, S, 2019).

2.6 *Dataset*

Dataset yang digunakan terdiri dari data numerik dan kategori. Data terdiri dari 200 *record* dan 6 variabel antara lain Usia, Jenis Kelamin, Tingkat Tekanan Darah (BP), Tingkat Kolesterol dan Na_to_K (perbandingan jumlah Natrium dan Kalium yang ditemukan didalam darah) dan DRUG (jenis obat). Natrium dan Kalium adalah elektrolit yang dibutuhkan tubuh yang berfungsi secara normal untuk membantu menjaga volume cairan dan darah dalam tubuh. Namun, apabila mengonsumsi terlalu banyak Natrium dan tidak cukup Kalium seseorang dapat terkena tekanan darah tinggi.

Dataset jenis obat dibagi menjadi data *training* sebesar 70% dan data *testing* sebesar 30%. Data *training* digunakan untuk menghasilkan model prediksi dengan menggunakan algoritma *decision tree* dan data *testing* digunakan untuk melihat performa model prediksi yang dihasilkan.

2.7 RStudio

Gambar 2 Logo RStudio
(Sumber: Medium.com)

Rstudio merupakan sebuah perangkat lunak *open source* atau gratis dengan pemograman R yang digunakan untuk mengolah data statistik untuk Windows, Macintosh, Linux dan UNIX. R pertama kali diciptakan oleh Ross Ihaka dan Robert Gentkeman, dimana nama R diambil dari nama depan kedua penciptanya. Paket (Rstudio merupakan suatu *Integrated Development Environment* (IDE) yang dikenal sebagai salah satu *powerful software* atau perangkat lunak pemrograman terintegrasi untuk analisis data, yaitu simulasi data, perhitungan dan tampilan grafik. Rstudio dapat menganalisis data dengan sangat efektif dan dilengkapi dengan operator pemrosesan array dan matriks, dengan fungsi tampilan grafik dan fungsi pemodelan data yang sangat baik (Januarsjaf, A, 2017).

3. Metodologi Penelitian

3.1 Studi Literatur

Studi literatur digunakan untuk mengumpulkan, mempelajari dan memahami informasi serta teori-teori yang berkaitan dengan penelitian melalui sumber studi literatur yang digunakan sebagai studi pustaka. Hal-hal yang diperlukan dalam penelitian meliputi metode Klasifikasi menggunakan algoritma *Decision Tree.*

3.2 Metode Pengumpulan Data

Dataset yang digunakan merupakan data sekunder dari *database* yang dapat diakses melalui https://www.kaggle.com/prathamtripathi/drug-classification. Data terdiri dari 200 *record* dengan beberapa variabel prediktor medis (Usia, Jenis Kelamin, Tingkat Tekanan Darah, Tingkat Kolesterol, Na to K dan Jenis Obat). Kategori usia berkisar antara 15 – 74 tahun. Kategori tekanan darah terdiri dari low, normal dan high. Kategori kolesterol terdiri dari high dan normal. Kategori Na to K berkisar antara 6,27 – 38,2 dan kategori jenis obat terdiri dari Drug A, Drug B,Drug C, Drug X, Drug Y.

3.3 Metode Analisis Data

Metode analisis data yang digunakan yaitu *Cross-Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM merupakan standar yang telah dikembangkan pada tahun 1996 yang ditunjukan untuk melakukan proses analisis dari suatu industri sebagai strategi pemecahan masalah dari bisnis atau unit penelitian. Berikut penjelasan dari tahapan-tahapan metode CRISP-DM: (Chapman, 2000)

a. *Business Understanding*
Tahapan ini merupakan fase awal untuk mengetahui masalah yang terjadi terhadap data jenis obat dan melakukan solusi yang tepat untuk permasalahan yang ada.

b. *Data Understanding*
Pada tahap ini dilakukan pengumpulan data yang dibutuhkan untuk dilakukan pemahaman terhadap tiap atribut yang terdapat dalam data yang sudah diperoleh.

c. *Data Preparation*
Tahap ini mencakup semua aktivitas untuk membuat kumpulan data akhir. Dimana dilakukan pembersihan data dan pemilihan atribut yang akan digunakan untuk selanjutnya pemodelan.

d. *Modelling*
Tahapan ini meliputi pemilihan teknik *data mining* dengan menentukan algoritma yang akan digunakan. Dalam tahap ini, berbagai macam teknik pemodelan dipilih dan diterapkan ke *dataset* yang sudah disiapkan untuk mengatasi kebutuhan bisnis tertentu. Tahap pembuatan model juga mencakup penilaian dan

analisa komparatif dari berbagai model yang dibangun.

e. Evaluation
Dalam tahapan ini akan dilakukan evaluasi serta pengukuran keakuratan hasil yang dicapai oleh model yang telah dibuat. Untuk mengetahui hubungan antar faktor atribut digunakan *Correlation Matrix* yang dapat mendeskripsikan bentuk dan kekuatan hubungan antar faktor tersebut.

f. Deployment
Pada tahap terakhir, penelitian yang telah dilakukan akan dipresentasikan dalam bentuk laporan akhir berisi grafik atau deksripsi yang mudah dipahami.

4. Hasil dan Pembahasan

Hasil penelitian yang dilakukan berupa hasil prediksi melalui pengolahan data dengan menggunakan C4.5 dan diimplementasikan menggunakan *software* bahasa pemograman R atau biasa disebut RStudio.

4.1 *Bussiness Understanding*

Pada tahap ini, berfokus pada pemahaman tujuan kebutuhan berdasarkan penilaian bisnis. Selanjutnya pemahaman tersebut diubah menjadi sebuah rencana awal *data mining*, dengan tujuan dapat mengklasifikasikan jenis obat yang akurat untuk pasien dengan memprediksi berdasarkan data yang digunakan. Dengan menentukan dan memahami label target terlebih dahulu.

4.2 *Data Understanding*

Pada tahap ini, data yang digunakan bersifat sekunder yang diperoleh dari kaggle.com. Kemudian diidentifikasi dan dilakukan pemahaman terhadap data dengan mendeksripsikan agar dapat memberikan gambaran data. Berikut pada Tabel 2 terdapat *type* data dan keterangan atribut yang digunakan.

. Tabel 2 Keterangan Data Atribut

Atribut	*Type* data	Keterangan
Usia	Numerik	*Predictor*
Jenis Kelamin	Kategorikal	*Predictor*
Tingkat Tekanan Darah	Kategorikal	*Predictor*
Tingkat Kolesterol	Kategorikal	*Predictor*
Na_to_K	Numerik	*Predictor*
Jenis Obat	Kategorikal	*Class target* (Drug A, Drug B, Drug C, Drug X dan Drug Y)

Pada tabel diatas, terdiri dari data numerik dan kategorikal dan terdapat 6 atribut yang digunakan, dimana 5 atribut sebagai *predictor* dan 1 atribut sebagai *class target.*

4.3 *Data Preparation*

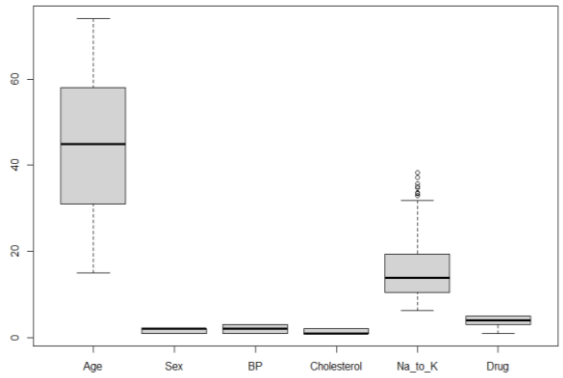
Pada tahapan selanjutnya, menyiapkan *dataset* yang akan diolah untuk selanjutnya pada tahap pemodelan menggunakan RStudio. Pada tahap ini, dilakukan pemeriksaan *missing value* dan data *outlier* agar data dapat di olah dengan baik.

```
> summary(data)
      Age            Sex                BP
 Min.   :15.00   Length:200         Length:200
 1st Qu.:31.00   Class :character   Class :character
 Median :45.00   Mode  :character   Mode  :character
 Mean   :44.31
 3rd Qu.:58.00
 Max.   :74.00

 Cholesterol          Na_to_K          Drug
 Length:200         Min.   : 6.269   Length:200
 Class :character   1st Qu.:10.445   Class :character
 Mode  :character   Median :13.937   Mode  :character
                    Mean   :16.084
                    3rd Qu.:19.380
                    Max.   :38.247
```

Gambar 3 Hasil *Summary*

Pada Gambar 3, dilakukan pengecekan terhadap *missing value* dengan menggunakan perintah *summary()*. Diketahui bahwa tidak terdapat *missing values* pada masing-masing atribut. Kemudian dilakukan pengecekan data *outlier.*



Gambar 4 *Visualisasi* Hasil *Data Outlier*

Berdasarkan hasil pada Gambar 4, diketahui bahwa terdapat data *outlier* pada atribut Na_to_K. Selanjutnya data *outlier* tersebut dilakukan dengan pembersihan (*cleaning)* agar data selanjutnya dapat di *modelling* dengan baik.

Kemudian dilakukan pemilihan data untuk selanjutnya dijadikan atribut. Dengan hasil dimana terdapat 6 atribut yang digunakan diantaranya usia, jenis kelamin, tingkat tekanan darah (BP), tingkat kolesterol, Na_to_K (perbandingan jumlah Natrium dan Kalium yang ditemukan didalam darah) dan jenis obat.

4.4 *Modelling*

```
#pembagian data [70% train, 30% test]
set.seed(1234)
ml <- sample(2, nrow(data), replace = TRUE, prob = c(0.7,0.3))

train <- data[ml == 1,]
test <- data[ml == 2,]

#penggunaan k-fold cross valid.
set.seed(1234)
folds<-cut(seq(1,nrow(data_new)),breaks = 10,labels = FALSE)
for(i in 1:10){
  testIndexes <- which(folds==i, arr.ind = TRUE)
  test <- data_new[testIndexes,]
  train <- data_new[-testIndexes,]
}

#modeling data
library(party)
predictor <- Drug~Age+Sex+BP+Cholesterol+Na_to_K
tree <- ctree(predictor, data = train)
tree
plot(tree)
```

Gambar 5 Script pemodelan pada RStudio

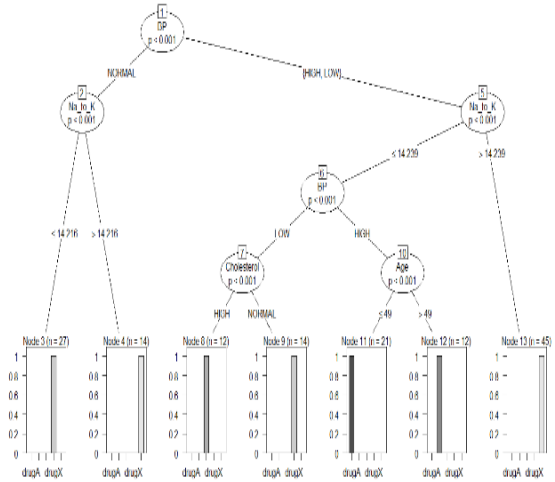
Pada Gambar 5 menunjukkan proses pengujian C4.5 menggunakan RStudio. Proses dilakukan dari pembagian *dataset* menjadi 70% data *training* dan 30% data *testing*. Selanjutnya digunakan *k-fold Cross Validation* dengan *k = 10*, karena memiliki kemampuan estimasi kinerja algoritma yang lebih akurat (Widaningsih, S, 2019).

Setelah dilakukan pemodelan, maka selanjutnya akan menghasilkan dari algoritma C4.5 yaitu berupa hasil prediksi dan aturan-aturan (*rule)* dan pohon keputusan (tree).

```
> print(comparation_result)
      prediction actual
 [1,] "DrugY"    "DrugY"
 [2,] "drugX"    "drugX"
 [3,] "drugX"    "drugX"
 [4,] "DrugY"    "DrugY"
 [5,] "DrugY"    "DrugY"
 [6,] "DrugY"    "DrugY"
 [7,] "drugB"    "drugB"
 [8,] "drugA"    "drugA"
 [9,] "DrugY"    "DrugY"
[10,] "DrugY"    "DrugY"
```

Gambar 6 Hasil Prediksi

Pada Gambar 6, menunjukkan hasil prediksi yang dihasilkan dari algoritma C4.5. Dari hasil prediksi yang dihasilkan berdasarkan data yang digunakan dapat menentukan bahwa *prediction* dan *actual* menghasilkan nilai yang sama, artinya algoritma C4.5 dapat memprediksi kelas untuk semua *instance* secara tepat.



Gambar 7 Hasil pohon keputusan (tree)

Dari Gambar diatas, terlihat bahwa semua atribut berkontribusi dalam pohon keputusan dengan beberapa aturan yang dihasilkan sebagai berikut:
R1= Jika tekanan darah NORMAL dan Na_to_K kurang dari 14.216, maka Drug X.
R2= Jika tekanan darah NORMAL dan Na_to_K lebih dari 14.216, maka Drug Y.
R3= Jika tekanan darah HIGH dan Na_to_K lebih besar dari 14.239, maka Drug Y.
R4= Jika tekanan darah HIGH, Na_to_K kurang dari 14.239, tekanan darah HIGH dan umur lebih dari 49, maka Drug B.
R5= Jika tekanan darah HIGH, Na_to_K kurang dari 14.239, tekanan darah HIGH dan umur kurang dari 49, maka Drug A.
R6= Jika tekanan darah HIGH, Na_to_K kurang dari 14.239, tekanan darah LOW dan tingkat kolesterol NORMAL, maka Drug X.
R7= Jika tekanan darah HIGH, Na_to_K kurang dari 14.239, tekanan darah LOW dan tingkat kolesterol HIGH, maka Drug C.
R8= Jika tekanan darah LOW dan Na_to_K lebih besar dari 14.239, maka Drug Y.
R9= Jika tekanan darah LOW, Na_to_K kurang dari 14.239, tekanan darah HIGH dan umur lebih dari 49, maka Drug B.
R10= Jika tekanan darah LOW, Na_to_K kurang dari 14.239, tekanan darah HIGH dan umur kurang dari 49, maka Drug A.
R11= Jika tekanan darah LOW, Na_to_K kurang dari 14.239, tekanan darah LOW dan tingkat kolesterol NORMAL, maka Drug X.
R12= Jika tekanan darah LOW, Na_to_K kurang dari 14.239, tekanan darah LOW dan tingkat kolesterol HIGH, maka Drug C.

4.5 *Evaluation* Data

Pada tahap evaluasi, dilakukan dengan menggunakan metode *confusion matrix* untuk mengetahui hasil tingkat nilai *accuracy, recall* dan *precision.* Karena pada *class* target atau label klasifikasi terdapat lebih dari 2 *class* maka

menggunakan *confusion matrix multiclass.* Berikut pada Gambar 8 merupakan hasil dari *confusion matrix multiclass.*

```
> confusionMatrix(table(data = testPred, reference = test$Drug))
Confusion Matrix and Statistics

          reference
data       drugA drugB drugC drugX DrugY
  drugA        2     0     0     0     0
  drugB        0     1     0     0     0
  drugC        0     0     3     0     0
  drugX        0     0     0     5     0
  DrugY        0     0     0     0     9
```

Gambar 8 Hasil *confusion matrix*

Berdasarkan hasil *confusion matrix* diatas, dapat diketahui tingkat nilai *accuracy, precision* dan *recall* sebagai berikut:

```
Overall Statistics

               Accuracy : 1
                 95% CI : (0.8316, 1)
    No Information Rate : 0.45
    P-Value [Acc > NIR] : 1.159e-07

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: drugA Class: drugB Class: drugC Class: drugX Class: DrugY
Sensitivity                 1.0         1.00         1.00         1.00         1.00
Specificity                 1.0         1.00         1.00         1.00         1.00
Pos Pred Value              1.0         1.00         1.00         1.00         1.00
Neg Pred Value              1.0         1.00         1.00         1.00         1.00
Prevalence                  0.1         0.05         0.15         0.25         0.45
Detection Rate              0.1         0.05         0.15         0.25         0.45
Detection Prevalence        0.1         0.05         0.15         0.25         0.45
Balanced Accuracy           1.0         1.00         1.00         1.00         1.00
```

Gambar 9 Hasil evaluasi menggunakan Rstudio

5. Kesimpulan dan Saran

5.1 Kesimpulan

Dalam klasifikasi penentuan jenis obat ini bertujuan untuk memprediksi jenis obat yang akurat untuk pasien dengan menggunakan metode *decision tree* C4.5, pembagian data dilakukan menjadi 2 dimana 70% *data training* dan 30% *data testing*. Berdasarkan pengujian, menghasilkan 12 aturan (*rule*) yang dihasilkan oleh pohon keputusan (*tree)* dengan semua atribut dapat berkontribusi dalam pohon keputusan. Secara umum, kinerja algoritma C4.5 sangat baik dalam memprediksi jenis obat, hal tersebut dibuktikan dari hasil prediksi yang sesuai dengan data aktual dan nilai akurasi yang dihasilkan sebesar 100%.

Dalam penelitian ini hanya menggunakan satu metode saja yaitu algoritma *decision tree* C.45. Untuk pengembangan lebih lanjut disarankan dengan melakukan komparasi dengan metode klasifikasi lainnya, agar hasil dari beberapa metode tersebut dapat dibandingkan keakuratannya.

Daftar Pustaka:

Anam, C., & Santoso, H. B. (2018): *Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa*. *Jurnal Ilmiah Ilmu-Ilmu Teknik*, *8*(1), 13–19. https://ejournal.upm.ac.id/index.php/energy/article/view/111/449

Bahri, S., Marisa Midyanti, D., Hidayati, R., Sistem Komputer, J., & Mipa, F. (2018): *Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak*. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, 24–31.

Bramer, M. (2007): *Principles of Data Mining*. London, Spinger.

Chapman, Peter, dkk. (2000): *CRISP-DM v.1.0 Step-by-step data mining guide*, SPSS Inc.

Elisa, E. (2017): *Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti*. *Jurnal Online Informatika*, *2*(1), 36. https://doi.org/10.15575/join.v2i1.71

Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. (2020): *Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5*. *Indonesian Journal of Statistics and Its Applications*, *4*(1), 80–88. https://doi.org/10.29244/ijsa.v4i1.330

Gorunescu, F. (2011): *Data Mining Concepts, Models and Technique*. Berlin: Spinger.

Januarsjaf, A. (2017, Januari 14): *Apakah R itu?* Dipetik Februari 20, 2021, dari https://rstudio-pubs-static.s3.amazonaws.com/241862_81533a8076474817a5aeb40c3f1f9406.html

Kasih, P. (2019): *Pemodelan Data Mining Decision Tree Dengan Classification Error Untuk Seleksi Calon Anggota Tim Paduan Suara*. *Jurnal Innovation in Research of Informatics (INNOVATICS)*, *1*(2), 63-69.

Kurniawan, Y. I. (2018): *Perbandingan Algoritma Naive Bayes dan C. 45 Dalam Klasifikasi Data Mining*. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, *5*(4), 455-464.

Mardi, Y. (2017): *Data Mining : Klasifikasi Menggunakan Algoritma C4.5*. *Jurnal Edik Informatika*, *2*(2), 213–219.

Noviandi. (2018): *Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes*. *Inohim*, *6*(1), 1–5.

Raharjo, M., Putra, J. L., & Sandi, T. A. A. (2019): *Implementasi Metode Decision Tree Klasifikasi Data Mining Untuk Prediksi Peminatan Jurusan Robotika oleh Mahasiswa*. *Jurnal Teknik Komputer*, *5*(2), 161-166.

Santosa, I., Rosiyah, H., & Rahmanita, E. (2018): *Implementasi Algoritma Decision Tree C. 45 Untuk Diagnosa Penyakit Tubercolusis (Tb)*. *Jurnal Ilmiah NERO*, *3*(3), 169–176.

Supriyanti, W., Kusrini, & Amborowati, A. (2016): *Perbandingan Kinerja Algoritma c4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa*. *Jurnal INFORMA Politeknik Indonusa*, *1*(3), 61–67.

Widaningsih, S. (2019). *Perbandingan Metode Data Mining untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika dengan Algoritma C4.5, Naive Bayes, Knn Dan Svm*. *Jurnal Tekno Insentif*, 16-25.