TUGAS AKHIR – SS141501

KLASIFIKASI SENYAWA OBAT KANKER UNTUK OPTIMASI PROTEKSI RADIASI MENGGUNAKAN PENDEKATAN *MACHINE LEARNING*

RIZKY MUBAROK
NRP 062114 4000 0074

Dosen Pembimbing
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

TUGAS AKHIR – SS141501

KLASIFIKASI SENYAWA OBAT KANKER UNTUK OPTIMASI PROTEKSI RADIASI MENGGUNAKAN PENDEKATAN *MACHINE LEARNING*

RIZKY MUBAROK
NRP 062114 4000 0074

Dosen Pembimbing
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA, KOMPUTASI, DAN SAINS DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

FINAL PROJECT – SS141501

CLASSIFICATION OF CANCER DRUG COMPOUND FOR OPTIMIZING RADIATION PROTECTION USING MACHINE LEARNING APPROACH

RIZKY MUBAROK
SN 062114 4000 0074

Supervisor
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCE
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2018

LEMBAR PENGESAHAN

KLASIFIKASI SENYAWA OBAT KANKER UNTUK OPTIMASI PROTEKSI RADIASI MENGGUNAKAN PENDEKATAN *MACHINE LEARNING*

TUGAS AKHIR

Diajukan Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Sains
pada
Program Studi Sarjana Departemen Statistika
Fakultas Matematika, Komputasi, dan Sains Data
Institut Teknologi Sepuluh Nopember

Oleh :

Rizky Mubarok
NRP. 062114 4000 0074

Disetujui oleh Pembimbing:
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si. (.....)
NIP. 19820326 200312 1 004

Mengetahui,
Kepala Departemen

Dr. Suhartono
NIP. 19710929 199512 1 001

SURABAYA, JULI 2018

(*Halaman ini sengaja dikosongkan*)

KLASIFIKASI SENYAWA OBAT KANKER UNTUK OPTIMASI PROTEKSI RADIASI MENGGUNAKAN PENDEKATAN *MACHINE LEARNING*

Nama Mahasiswa : Rizky Mubarok
NRP : 062114 4000 0074
Departemen : Statistika
Dosen Pembimbing : Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

Abstrak

Salah satu pengobatan penyakit kanker yang banyak digunakan adalah terapi radiasi atau radioterapi dengan menggunakan senyawa yang mematikan sel kanker. Senyawa untuk optimasi proteksi radiasi akan diklasifikasikan secara biner yaitu proteksi radiasi tinggi dan rendah. Senyawa tersebut berjumlah 84 senyawa dengan masing-masing senyawa disusun oleh 217 prediktor, sehingga data yang digunakan tergolong high dimensional data. Oleh karena itu, dilakukan feature selection berdasarkan nilai mean decrease gini (MDG). Metode yang digunakan adalah naïve bayes classifier (NBC) dan classification and regression tree (CART). Pada klasifikasi berdasarkan kelas awal yaitu dua kelas proteksi radiasi, nilai AUC dari data testing dengan NBC dan CART berturut-turut adalah 0,549 yang didapatkan dari 5% prediktor dan 0,663 yang didapatkan dari 10% prediktor. Selain menggunakan kelas awal, juga dilakukan pembentukan dua kelas baru dengan pendekatan mixture normal. Hasil nilai AUC data testing dengan menggunakan kelas baru tersebut adalah pada NBC meningkat menjadi 0,592 dengan menggunakan 20% dan 25% prediktor, serta CART turun menjadi 0,617 dengan 5%, 35%, dan 100% prediktor.

Kata Kunci : Classification and Regression Tree (CART), Feature Importance, High Dimensional Data, Mixture Distribution, Naïve Bayes.

(*Halaman ini sengaja dikosongkan*)

CLASSIFICATION OF CANCER DRUG COMPOUND FOR OPTIMIZING RADIATION PROTECTION USING MACHINE LEARNING APPROACH

Name : Rizky Mubarok
Student Number : 062114 4000 0074
Department : Statistics
Supervisor : Dr. rer. pol. Heri Kuswanto, S.Si., M.Si

Abstract

The leading cause of the world's mortality and morbidity is cancer. One of the cancer treatments is radiation therapy or radiotherapy. The therapy is using compounds that can increase the death rate of a cancer cell. Compounds that used for optimizing the radiation protection will be binary classified in high and low radiation protection. The total compounds are 84 compounds with each compound composed of 217 predictors so that the data is classified as high-dimensional data. Therefore, feature selection is needed and performed based on the mean decrease gini (MDG). The method used for classification is naïve bayes classifier (NBC) and classification and regression tree (CART). In the classification based on initial class which is two classes of radiation protection, testing data have the AUC value from NBC is 0.549 using 5% predictors and CART is 0.663 using 10% predictors. In this research, also conducted the formation of a new class with mixture normal distribution approach. The AUC value of testing data using new class is on NBC increased to 0,592 by using 20% and 25% predictors, while CART decreased to 0,617 with 5%, 35%, 100% predictors.

Keywords: Classification and Regression Tree (CART), Feature Importance, High Dimensional Data, Mixture Distribution, Naïve Bayes.

(*Halaman ini sengaja dikosongkan*)

KATA PENGANTAR

Puji syukur penulis panjatkan atas rahmat dan hidayah yang diberikan Allah SWT sehingga penulis dapat menyelesaikan laporan Tugas Akhir yang berjudul **"Klasifikasi Senyawa Obat Kanker Untuk Optimasi Proteksi Radiasi Menggunakan Pendekatan *Machine Learning*"** dengan lancar.

Penulis menyadari bahwa Tugas Akhir ini dapat terselesaikan tidak terlepas dari bantuan dan dukungan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada :

1. Kedua orang tua, atas segala do'a, nasehat, kasih sayang, dan dukungan yang diberikan kepada penulis demi kesuksesan dan kebahagiaan penulis.
2. Dr. Suhartono selaku dosen wali dan Ketua Departemen Statistika, yang telah banyak memberikan saran dan arahan dalam proses belajar di Departemen Statistika.
3. Dr. Sutikno, M.Si. selaku Ketua Program Studi Sarjana yang telah memberikan fasilitas, sarana, dan prasarana dalam proses belajar di Departemen Statistika.
4. Dr. rer. pol. Heri Kuswanto, S.Si., M.Si. selaku dosen pembimbing yang telah meluangkan waktu dan dengan sangat sabar memberikan bimbingan, saran, dukungan serta motivasi selama penyusunan Tugas Akhir.
5. Dr. rer. pol. Dedy Dwi Prastyo, S.Si., M.Si. dan Santi Puteri Rahayu, Ph.D. selaku dosen penguji yang telah banyak memberikan masukan kepada penulis.
6. Sahabat-sahabat penulis, Veronica C. Maghfiroh, Siti Aisyah, Izzan Rasyadi, Taufik Afiif Maldini, Syahrul Eka Adi Laksana, Dedi Setiawan, Bayu Samudra, Dhamai Brillianggara, dan Rizky Nanda Noverianto yang selama ini telah membantu, mendukung, dan mendengarkan keluh kesah penulis selama masa perkuliahan berlangsung.
7. Teman-teman seperjuangan Tugas Akhir, khususnya Dedi Setiawan, Taufik Afiif Maldini, Erlin Sukmaputri, dan Kiki Noor Aisyah yang selama ini telah berjuang bersama dan saling memberikan semangat.

8. Teman-teman Statistika ITS angkatan 2014, Respect, yang selalu memberikan dukungan kepada penulis selama ini.
9. Teman-teman HIMASTA-ITS 2016/2017 khususnya Pengurus Harian, yang selama perkuliahan ini memberikan banyak pembelajaran dan mendukung penulis dalam mengembangkan *softskill* penulis.
10. Semua pihak yang turut membantu dalam pelaksanaan Tugas Akhir yang tidak bisa penulis sebutkan satu persatu.

Besar harapan penulis untuk mendapatkan kritik dan saran yang membangun sehingga Tugas Akhir ini dapat memberikan manfaat bagi semua pihak yang terkait.

Surabaya, Juli 2018

Penulis

DAFTAR ISI

Halaman

(*Halaman ini sengaja dikosongkan*)

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

(*Halaman ini sengaja dikosongkan*)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kanker merupakan istilah umum untuk sekelompok besar penyakit yang ditandai dengan pertumbuhan sel abnormal yang kemudian dapat menyerang bagian tubuh di sekitar sel abnormal tersebut dan / atau menyebar ke organ lain (World Health Organization, 2009). Kanker menjadi penyebab utama morbiditas dan mortalitas di dunia, hal tersebut menjadi permasalahan kesehatan serius yang terjadi baik di negara maju maupun berkembang. Terdapat 14 juta kasus baru dan 8 juta kematian kanker di dunia pada tahun 2012, angka tersebut sesuai dengan tingkat insiden dan tingkat kematian yaitu 182 dan 102 per 100.000 jiwa. Jenis kanker di dunia yang banyak menyerang laki-laki adalah kanker paru-paru (16,7%), kanker prostat (15%), kanker kolorektal (10%), kanker perut (8,5%), dan kanker hati (7,5%). Sedangkan jenis kanker yang banyak menyerang perempuan adalah kanker payudara (25,2%), kanker kolorektal (9,2%), kanker paru-paru (8,7%), kanker serviks (7,9%), dan kanker perut (4,8%) (International Agency for Research on Cancer, 2014). Berdasarkan proyeksi World Health Organization (WHO), kematian akibat kanker akan meningkat di tahun 2020 yaitu sebesar 10,1 juta jiwa (5,8 laki-laki dan 4,3 perempuan) dan meningkat lagi di tahun 2025 yaitu sebesar 10,6 juta jiwa (6,7 laki-laki dan 4,9 perempuan). Prediksi peningkatan kematian akibat kanker juga terjadi di Indonesia dimana pada tahun 2020 diprediksikan terjadi 239.030 kematian (128.167 laki-laki dan 110.863 perempuan) dan di tahun 2025 sebanyak 283.776 (153.566 laki-laki dan 130.210 perempuan) kematian (International Agency for Research on Cancer, 2013).

Kanker dapat mempengaruhi hampir semua bagian tubuh dan memiliki banyak subtipe molekuler dan anatomi yang masing-masing memerlukan strategi pengelolaan yang spesifik. Sel kanker dapat ditemukan pada tumor ganas yang memiliki ciri-

ciri sel tumbuh secara tidak terbatas, memiliki selubung, dan menyebar atau dapat menyusup ke jaringan di sekitarnya. Kementerian Kesehatan Republik Indonesia (2015) menerangkan bahwa faktor-faktor yang memengaruhi timbulnya penyakit kanker diantaranya adalah faktor genetik; faktor karsinogen yaitu zat kimia, radiasi, virus, dan iritasi kronis; serta faktor perilaku atau gaya hidup yaitu pola makan yang tidak sehat, konsumsi alkohol, dan kurangnya aktivitas fisik.

Penanganan pada penyakit kanker akan lebih baik jika berhasil diketahui atau dideteksi sejak dini. Sehingga, diperlukan upaya pencegahan dengan meningkatkan kesadaran masyarakat dalam mengenali gejala dan resiko penyakit kanker agar dapat menentukan langkah-langkah pencegahan yang tepat. Berbagai penelitian terkait dengan pengobatan kanker telah dilakukan oleh peneliti di dunia, diantaranya adalah Diamantis dan Banerji (2016) yang meneliti tentang *Antibody-Drug Conjugates* (ACDs) sebagai kelas baru yang muncul dari pengobatan kanker yang menggabungkan selektivitas dari pengobatan yang ditargetkan dengan *cytotoxic potency* dari obat kemoterapi, serta Sharma, *et al.* (2016) yang meneliti tentang kombinasi obat dan radioterapi untuk memperbaiki hasil klinis pasien penderita kanker. Kemudian penelitian terkait pengobatan kanker juga telah dilakukan oleh Yap, *et al*. (2016) tentang identifikasi target kanker prostat untuk penemuan obat baru, serta Santos, *et al*. (2016) tentang pembuatan peta komprehensif dari target molekuler obat kanker. Selain itu jugam dilakukan penelitian tentang penerapan *Structure-Based Drug Design* (SBDD) untuk target protein membran integral oleh Montfort dan Workman (2017).

Dalam beberapa tahun terakhir, *machine learning* menjadi sorotan yang menarik di bidang kesehatan dan penemuan obat, salah satunya pada kasus kanker. *Machine learning* merupakan metode analisis data yang membangun model analitik secara otomatis (SAS, 2018). *Machine learning* merupakan cabang dari *artificial intelligence* (AI) yang didasarkan pada gagasan bahwa

mesin harus bisa belajar dan beradaptasi melalui pengalaman dalam menganalisis data. Di bidang kesehatan dan penemuan obat, tujuan dari penggunaan *machine learning* diantaranya adalah mecari senyawa untuk obat baru secara efisien berdasarkan hasil prediksi variabel, seleksi variabel, klasifikasi, dan lain sebagainya. Berbagai penelitian terkait dengan pemanfaatan *machine learning* dalam bidang kesehatan dan penemuan obat terutama pada kasus penyakit kanker telah dilakukan oleh para peneliti di dunia, diantaranya adalah penelitian yang dilakukan oleh Soria, *et al*. (2008) yang membandingkan metode C4.5, *multulayer perceptron classifier* (MLP), dan *naïve bayes* pada kasus kanker payudara dimana MLP menghasilkan akurasi paling tinggi yaitu 94,9% dibandingkan *naïve bayes* dan C4.5 dengan akurasi 93,1% dan 87,6%. Kemudian dilakukan juga perbandingan metode C4.5 dan *naïve bayes* untuk mengklasifikasikan survivabilitas penderita kanker jantung oleh Dimitoglou, *et al*. (2012) dimana C4.5 dengan akurasi 94,44% lebih baik daripada *naïve bayes* dengan akurasi 92,38%. Kathija dan Nisha (2016) juga melakukan penelitian klasifikasi data kanker payudara menggunakan SVM dan *naïve bayes* yang menghasilkan akurasi *naïve bayes* lebih besar yaitu 97% dibandingkan dengan SVM dengan akurasi 93%. Selain itu, metode *naïve bayes*, regresi logistik, dan *decision tree* juga dibandingkan oleh Mandal (2017) pada kasus deteksi sel kanker payudara dengan hasil akurasi regresi logistik paling tinggi yaitu 97,9% dibandingkan dengan *decision tree* dan *naïve bayes* dengan akurasi 96,5% dan 94,4%. Penelitian juga dilakukan oleh Elsayad dan Elsalamony (2013) yang membandingkan performa beberapa model *decision tree* (C&R, CHAID, QUEST, C5.0) dan SVM untuk mendiagnosis kanker payudara, hasilnya menunjukkan bahwa SVM memiliki akurasi paling tinggi yaitu 96,6%, sedangkan akurasi C&R sebesar 95,6%, CHAID sebesar 96,2%, QUEST sebesar 95,7%, dan C5.0 sebesar 96,6%. Selain itu, Aruna, *et al.* (2011) juga melakukan penelitian untuk mendeteksi kanker payudara menggunakan

metode *naïve bayes*, RBF *neural* networks, J48, CART, dan SVM-RBF Kernel dengan hasil akurasi SVM-RBF Kernel paling tinggi diantara metode yang lain yaitu 99%, sedangkan akurasi CART sebesar 96,22%, J48 sebesar 95,28%, *naïve bayes* sebesar 94,33%, dan RBF NN sebesar 92,45%. Berdasarkan penelitian-penelitian tersebut, metode *naïve bayes* dan CART tergolong metode klasifikasi yang baik dengan tingkat akurasi yang tinggi.

Terapi radiasi atau yang biasa dikenal dengan radioterapi adalah salah satu pengobatan untuk penyakit kanker. Terapi ini tetap digunakan walaupun memiliki efek samping yang merugikan yaitu kematian sel normal dan jaringan normal di sekitar sel kanker yang diinduksi P53 (Morita, *et al*., 2014). P53 merupakan gen supresor tumor yang bertindak menghentikan perkembangan tumor. Hal ini dilakukan dengan mengaktifkan beberapa protein yang memicu kematian sel-sel yang rusak sehingga sel-sel tersebut tidak mereplikasi dan membelah tak terkendali (Kamus Kesehatan, 2018). Hal tersebut menjadi latar belakang Ariyasu, *et al.* (2014) yang melakukan penelitian dengan membuat 84 senyawa yang kemudian senyawa-senyawa tersebut dicobakan pada sel normal dan sel yang terkena radiasi sinar gamma. Percobaan pertama dilakukan dengan pemberian senyawa pada sel normal, percobaan ini mampu mengukur toksisitas dari senyawa tersebut. Sedangkan percobaan kedua dilakukan pemberian senyawa pada sel yang telah terkena radiasi sinar gamma (10 Gy), percobaan ini mampu mengukur proteksi radiasi. Tingkat kematian sel digunakan sebagai indikator di kedua percobaan tersebut. Jika tingkat kematian sel di percobaan pertama rendah dan di percobaan kedua tinggi, maka senyawa tersebut bisa digunakan sebagai radioprotektor. Data yang diperoleh dari penelitian tersebut kemudian digunakan untuk penelitian lanjutan yang dilakukan oleh Matsumoto, *et al.* (2016) yang membandingkan hasil ketepatan klasifikasi senyawa untuk optimasi proteksi radiasi dan toksisitas dengan menggunakan *random forest* (RF), *support vector machine* (SVM), *extreme gradient boosting* (XGB), dan K-*nearest neighbor* (KNN). Dari

hasil penelitian tersebut, didapatkan hasil bahwa untuk memprediksi senyawa proteksi radiasi, KNN memiliki nilai akurasi lebih tinggi yaitu 75,7% dibandingkan dengan RF, XGB, dan SVM yaitu masing-masing sebesar 68,8%, 63,5%, dan 62,8%. Penggunaan pendekatan *machine learning* tersebut dilakukan dengan melakukan *feature selection* dengan memilih *feature* terpenting yang dilihat dari indeks Gini, namun terdapat metode yang lebih akurat dan stabil untuk menghitung *feature importance* yaitu dengan indeks *mean decrease gini* (MDG) (Calle & Urrea, 2010).

Berdasarkan penelitian-penelitian yang telah dilakukan, belum terdapat justifikasi metode yang paling bagus di antara metode-metode yang lain. Selain itu juga nilai ketepatan klasifikasi yang dihasilkan oleh Matsumoto, *et al.* (2016) relatif rendah. Hal tersebut menunjukkan bahwa perlu dilakukan *treatment* untuk membuat nilai ketepatan klasifikasi meningkat. Salah satu hipotesis yang memungkinkan adalah tidak tepat dalam pembentukan kelas, sehingga perlu membuat pengelompokan baru yang kemudian dijadikan kelas untuk klasifikasi. Metode pembentukan kelas baru yang digunakan adalah menggunakan pendekatan *mixture distribution*. Jika diketahui suatu data, tidak selamanya satu distribusi saja dapat merepresentasikan data tersebut. Namun, apabila ada indikasi beberapa komposisi muncul dari data tersebut, maka tidak menutup kemungkinan bahwa distribusi data yang lebih tepat adalah distribusi *mixture* (Dempster, *et al*., 1977).

Oleh karena itu, pada penelitian ini akan dilakukan klasifikasi senyawa untuk optimasi proteksi radiasi dengan beberapa metode. Metode yang digunakan adalah dengan menggunakan metode *naïve bayes* dan *classification and regression tree* (CART). Metode *naïve bayes classifier* digunakan karena metode ini hanya membutuhkan jumlah data *training* yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian (Pattekari & Parveen, 2012). Sedangkan metode CART dipilih karena bisa digunakan untuk

data kontinu maupun kategorik, sederhana, dan mudah diinterpretasikan (Breiman, *et al.*, 1993). Senyawa untuk optimasi proteksi radiasi akan diklasifikasikan secara biner berdasarkan tingkat kematian sel kanker, yaitu proteksi radiasi tinggi dan proteksi radiasi rendah. Pada dasarnya, metode *naïve bayes* dan CART tidak didesain untuk *high dimensional data*, sehingga pada penelitian ini juga dilakukan *feature selection* sebelum melakukan klasifikasi dengan menggunakan *mean decrease gini* (MDG) untuk mengetahui tingkat kepentingan dari masing-masing feature (*feature importance*). Kemudian, nantinya senyawa akan diklasifikasikan dengan menggunakan *features* sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% dari keseluruhan *features*. Persentase *feature* tersebut dipilih karena dengan menggunakan *feature* sejumlah 5%-35% data tidak tergolong *high dimensional data*, sedangkan 100% dipilih untuk membandingkan performa metode *naïve bayes* dan CART dengan metode yang digunakan pada penelitian Matsumoto, *et al.* pada tahun 2016. Selain itu, hasil ketepatan klasifikasi dengan menggunakan dua kelas proteksi radiasi juga dibandingkan dengan kelas baru hasil pengelompokan dengan pendekatan *normal mixture distribution*. Pada penelitian ini, metode evaluasi yang digunakan adalah *total accuracy rate*, *sensitivity*, *specificity*, dan AUC. Berdasarkan hasil klasifikasi tersebut, dapat diketahui senyawa-senyawa yang memiliki proteksi radiasi tinggi, sehingga senyawa tersebut bisa direkomendasikan untuk digunakan sebagai radioprotektor yang baik.

1.2 Rumusan Masalah

Permasalahan yang terjadi pada pengobatan kanker dengan menggunakan radioterapi adalah diperlukan senyawa yang mampu untuk meningkatkan tingkat kematian pada sel kanker dan menekan tingkat kematian pada sel normal atau jaringan normal yang berada di sekitar sel kanker tersebut. Oleh karena itu, pada penelitian ini akan dilakukan klasifikasi senyawa untuk optimasi proteksi radiasi sehingga dapat diketahui senyawa-senyawa yang bisa meningkatkan tingkat kematian pada sel

kanker atau jaringan yang rusak. Metode yang digunakan adalah *naïve bayes* dan *classification and regression tree* (CART), dimana dari kedua metode tersebut akan dibandingkan untuk mengetahui metode yang menghasilkan nilai akurasi lebih tinggi. Selain itu juga dilakukan pembentukan kategori baru untuk digunakan sebagai kelas klasifikasi.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut.

1. Mendapatkan nilai *feature importance* dari masing-masing *features* terhadap proteksi radiasi.
2. Mengevaluasi hasil ketepatan klasifikasi senyawa untuk optimasi proteksi radiasi dengan metode *naïve bayes* dan *classification and regression tree* (CART) menggunakan dua kelas awal proteksi radiasi.
3. Mengelompokkan senyawa untuk optimasi proteksi radiasi dengan menggunakan pendekatan *mixture distribution* yang kemudian digunakan sebagai kelas baru.
4. Mengevaluasi hasil ketepatan klasifikasi senyawa untuk optimasi proteksi radiasi dengan metode *naïve bayes* dan *classification and regression tree* (CART) menggunakan dua kelas baru hasil pengelompokan menggunakan pendekatan *normal mixture distribution*.
5. Membandingkan hasil evaluasi ketepatan klasifikasi menggunakan kelas awal dan kelas baru.

1.4 Manfaat

Penelitian ini diharapkan dapat memberikan beberapa manfaat bagi berbagai pihak, diantaranya sebagai berikut.

1. Memberikan wawasan keilmuan statistika tentang klasifikasi secara umum dengan metode *naïve bayes* dan *classification and regression tree* (CART) menggunakan *feature importance* dari *mean decrease gini* (MDG), serta tentang *mixture distribution.*

2. Memberikan informasi dan rekomendasi untuk profesional di bidang kesehatan terkait dengan senyawa yang memiliki komposisi yang bisa meningkatkan kematian sel kanker yang kemudian dapat digunakan sebagai radioprotektor.

1.5 Batasan Masalah

Batasan masalah yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Metode klasifikasi yang digunakan adalah *naïve bayes classifier* dan *classification and regression tree* (CART) *classifier*.
2. Metode *feature selection* yang digunakan adalah *feature importance* dengan *mean decrease gini* (MDG).

BAB II
TINJAUAN PUSTAKA

Bab ini membahas mengenai statistika deskriptif, *feature selection* menggunakan *mean decrease gini* (MDG), *naïve bayes classifier*, *classification and regression tree* (CART), evaluasi ketepatan klasifikasi, metode validasi, distribusi *mixture*, kanker, dan radioterapi.

2.1 Statistika Deskriptif

Menurut Walpole, *et al*. (2012) statistika deskriptif merupakan metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna, namun teknik statistika ini sama sekali tidak menarik kesimpulan dari data yang akan diteliti. Statistika deskriptif lebih berkenaan dengan pengumpulan dan peringkasan data, serta penyajian hasil ringkasan tersebut. Data-data yang bisa diperoleh merupakan hasil sensus, survei, jejak pendapat, atau pengamatan lainnya yang secara umum masih bersifat acak dan belum terorganisir dengan baik, atau biasa disebut *raw data*. Data-data tersebut diringkas dengan baik dan teratur dalam bentuk tabel atau presentasi grafis agar lebih mudah dipahami oleh pembaca.

2.2 *Feature Selection*

Pemilihan *feature* atau variabel digunakan untuk meningkatkan akurasi. Salah satu cara untuk memilih variabel adalah dengan menghitung tingkat kepentingan variabel atau *feature importance*. *Mean decrease gini* (MDG) merupakan salah satu ukuran tingkat kepentingan *feature* yang dihasilkan dari menggunakan metode *random forest*. Berdasarkan penelitian yang dilakukan oleh Calle & Urrea (2010), metode MDG lebih baik jika dibandingkan dengan metode *mean decrease accuracy* (MDA) dan *gini index* untuk mengukur tingkat kepentingan feature, karena MDG lebih stabil dan menghasilkan hasil yang

lebih *robust*. Berikut merupakan rumus untuk menghitung MDG (Marco & Zuccolotto, 2006).

$$MDG_n = \frac{1}{p}\sum_{t=1}^{p}[d(n,t)\ I(n,t)] \qquad (2.1)$$

Keterangan:

$d(n,t)$ = Besar penurunan indeks Gini untuk variabel x_n pada simpul t.

$I(n,t)$ = Bernilai 1 jika x_n memilah simpul t, bernilai 0 jika selainnya.

p = Banyaknya simpul dalam pohon.

x_n = Variabel prediktor ke-*n*, dimana $n = 1, 2, \dots, N$.

2.3 *Naïve Bayes Classifier*

Naïve bayes classifier merupakan metode klasifikasi popular dan termasuk dalam sepuluh algoritma terbaik dalam *data mining* (Wu & Kumar, 2009). *Naïve bayes classifier* merupakan metode klasifikasi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes atau aturan Bayes dengan asumsi independensi yang kuat (naif). Dalam teorema Bayes, maksud dari independensi yang kuat pada *feature* (variabel) adalah *feature* pada data tidak berkaitan dengan ada atau tidaknya *feature* lain dalam data yang sama. Metode ini menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data yang diberikan (Patil & Sherekar, 2013). Definisi lain dikemukakan oleh ilmuwan Inggris, Thomas Bayes, bahwa *naïve bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2013). Keuntungan menggunakan *naïve bayes classifier* adalah metode ini hanya membutuhkan jumlah data *training* yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian (Pattekari & Parveen, 2012).

Teorema Bayes menggambarkan hubungan antara peluang bersyarat dari dua kejadian. Persamaan dari teorema Bayes dituliskan dalam persamaan (2.2) sebagai berikut.

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)} \quad (2.2)$$

Keterangan:
D = Data dengan kelas yang belum diketahui.
H = Hipotesis data pada suatu kelas spesifik.
$P(H|D)$ = Probabilitas hipotesis *H* berdasarkan kondisi *D* (*posterior probability*).
$P(H)$ = Probabilitas hipotesis *H* (*prior probability*).
$P(D|H)$ = Probabilitas *D* berdasarkan kondisi pada hipotesis *H*.
$P(D)$ = Probabilitas *D*.

Klasifikasi menggunakan metode *naïve bayes* memerlukan petunjuk untuk menentukan kelas yang sesuai bagi data yang akan diklasifikasikan. Oleh karena itu, teorema Bayes pada persamaan (2.2) disesuaikan menjadi sebagai berikut.

$$P(c_i|x_1, x_2, \dots, x_N) = \frac{P(x_1,x_2,\dots,x_N|c_i)\,P(c_i)}{P(x_1,x_2,\dots,x_N)} \quad (2.3)$$

Pada persamaan (2.3), c_i mempresentasikan kelas klasifikasi, sedangkan variabel $x_1 \dots x_N$ merepresentasikan *features* yang dibutuhkan untuk melakukan klasifikasi. Persamaan tersebut menjelaskan bahwa peluang masuknya data tertentu dalam kelas c_i (*posterior*) adalah peluang munculnya kelas c_i (*prior*) dikali dengan peluang kemunculan *features* pada kelas c_i (*likelihood*), kemudian dibagi dengan peluang kemunculan *features* secara global (*evidence*). Oleh karena itu, persamaan (2.3) dapat juga dituliskan sebagai berikut.

$$Posterior = \frac{Likelihood \times Prior}{Evidence} \quad (2.4)$$

Nilai *posterior* yang dihasilkan akan dibandingkan dengan nilai *posterior* lain untuk menentukan kelas yang sesuai. Sedangkan nilai *evidence* selalu tetap untuk setiap kelas pada suatu sampel. Persamaan 2.3 tersebut dapat dijabarkan lebih lanjut dengan menjabarkan $(c_i|x_1, x_2, \dots, x_N)$ menggunakan aturan perkalian sebagai berikut.

$$P(c_i|x_1, x_2, \dots, x_N)$$
$$= P(c_i)\ P(x_1, x_2, \dots, x_N|c_i)$$
$$= P(c_i)\ P(x_1|c_i)\ P(x_2, x_3, \dots, x_N|c_i, x_1)$$
$$= P(c_i)\ P(x_1|c_i)\ P(x_2|c_i, x_1) \dots\ P(x_N|c_i, x, x_2, \dots, x_{N-1}) \quad (2.5)$$

Berdasarkan hasil penjabaran di persamaan (2.5), dapat dilihat bahwa semakin banyak atau semakin kompleks *features* yang mempengaruhi nilai probabilitas, maka semakin kompleks untuk dianalisa. Oleh karena itu digunakan asumsi independensi yang membuat masing-masing *features* $(x_1, x_2, \dots, x_N)$ saling bebas atau independen. Persamaan (2.6) merupakan persamaan yang berlaku ketika digunakan asumsi independen.

$$P(x_u|x_v) = \frac{P(x_u \cap x_v)}{P(x_v)} = \frac{P(x_u)P(x_v)}{P(x_v)} = P(x_u) \qquad (2.6)$$

untuk $u \neq v$, sehingga,

$$P(x_u|c_i, x_v) = P(x_u|c_i) \qquad (2.7)$$

Persamaan (2.7) merupakan persamaan dari *naïve bayes* yang selanjutnya akan digunakan dalam proses klasifikasi. *Naïve bayes classifier* dapat lebih mudah digunakan pada data *feature* bertipe kategorik seperti pada kasus klasifikasi dengan *feature* jenis kelamin dengan nilai {pria, wanita} atau tingkat pendidikan {SD, SMP, SMA, Sarjana}. *Naïve bayes classifier* juga bisa digunakan untuk data *feature* bertipe numerik dengan beberapa perlakuan sebagai berikut (Gorunescu, 2011).

1. Melakukan diskretisasi pada setiap *features* dan mengganti dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasikan data *feature* bertipe nominal menjadi ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk *feature* bertipe data kontinu dan memperkirakan parameter distribusi dengan *training data*. Pada klasifikasi data kontinyu biasanya digunakan distribusi Gaussian untuk merepresentasikan probabilitas bersyarat dari *feature* bertipe data kontinu pada $P(x_n|c_i)$. Sedangkan distribusi Gaussian

dikarakteristikkan dengan dua parameter, yaitu μ (rata-rata) dan σ^2 (varians). Untuk setiap kelas c_i, probabilitas bersyarat kelas c_i untuk *feature* x_n dituliskan pada persamaan (2.8) sebagai berikut.

$$P(x_n|c_i) = \frac{1}{\sigma_{ni}\sqrt{2\pi}} \exp\left(-\frac{(x_n-\mu_{ni})^2}{2{\sigma_{ni}}^2}\right) \qquad (2.8)$$

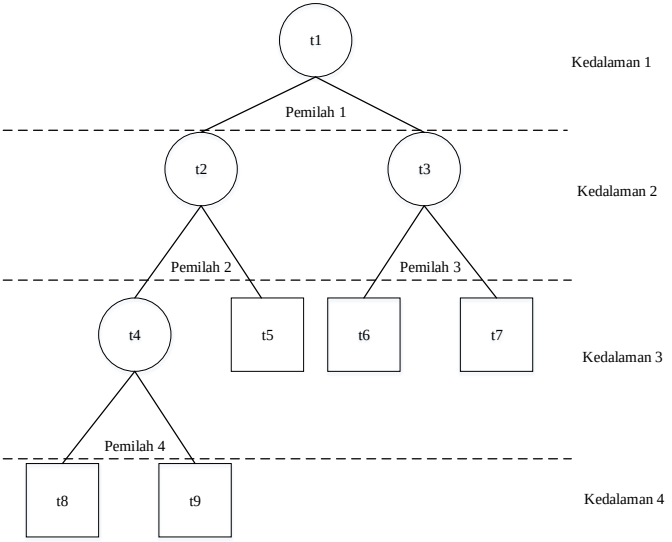
Keterangan:
P = Peluang.
x_n = Nilai *feature* ke-*n*.
c_i = Sub kelas *C* yang dicari.
μ = Rata-rata dari seluruh *feature*.
σ = Varians dari seluruh *feature*.
Estimasi peluang $P(x_n|c_i)$ dapat dihitung untuk setiap variabel x_n dan kelas c_i sehingga data baru akan dapat diklasifikasikan ke dalam kelas c_i apabila peluangnya lebih besar dibandingkan yang lainnya.

2.4 *Classification and Regression Tree* (CART) *Classifier*

Salah satu algoritma yang digunakan untuk klasifikasi menggunakan metode pohon keputusan atau *decision tree* adalah CART. Metode ini dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J., dimana konsep dari metode ini adalah dengan menggunakan algoritma penyekatan rekursif secara biner (*binary recursive partitioning*) (Brieman, *et al*., 1993). *Binary* berarti pemilahan dilakukan pada sekelompok data yang terkumpul pada suatu simpul (*node*) menjadi dua kelompok yang selanjutnya disebut simpul anak (*child node*). *Recursive* berarti pemilahan tersebut dapat terjadi berulang-ulang sampai memenuhi suatu kriteria tertentu. Setiap simpul dapat dipilah menjadi dua simpul anak, kemudian setiap simpul anak tersebut dapat memilah untuk membentuk simpul anak yang lain, dan seterusnya. Jika masih bisa dilakukan pemilahan pada simpul anak tersebut maka disebut simpul dalam (*internal node*), namun jika tidak ada lagi pemilahan setelahnya maka disebut simpul akhir (*terminal node*). Sedangkan *partitioning* memiliki arti

bahwa proses klasifikasi dapat dilakkan dengan cara memilah kumpulan data menjadi beberapa bagian atau partisi.

Menurut Brieman, *et. al.* (1993), CART akan menghasilkan pohon klasifikasi jika data variabel respon berskala kategorik dan akan menghasilkan pohon regresi jika data variabel respon berskala kontinu. Metode CART memiliki beberapa kelebihan. Pertama, metode ini bersifat non parametrik sehingga tidak memerlukan asumsi yang mengikat. Kedua, mampu mengeksplorasi data berdimensi tinggi atau pada data yang kompleks dan multivariabel. Ketiga, kombinasi data kontinu maupun kategorik dapat digunakan pada metode ini. Keempat, tidak hanya memberikan klasifiasi, tetapi juga memberikan estimasi probabilitas kesalahan pengklasifikasian. Kelima, hasil klasifikasi akhir berbentuk sederhana dan bisa mengklasifikasikan data baru secara efisien.



Gambar 2.1 Ilustrasi Struktur Pohon Klasifikasi
(Sumber: Breiman, *et al.*, 1993)

Ilustrasi struktur pohon klasifikasi yang ditunjukkan pada Gambar 2.1. Simpul awal yang mengandung seluruh data dengan notasi t1. Pada Gambar 2.1 simpul dalam (*internal node*) dinotasikan dengan t2, t3, dan t4, sedangkan simpul akhir (*terminal node*) dinotasikan dengan t5, t6, t7, t8, dan t9 dimana setelahnya tidak ada lagi pemilahan, artinya simpul anak yang dihasilkan telah homogen. Setiap simpul berada pada kedalaman (*depth*) tertentu, dimulai dari simpul awal t1 yang berada pada kedalaman 1, t2 dan t3 berada pada kedalaman 2, dan begitu seterusnya hingga dapat simpul t4, t5, t6, t8, dan t9 yang berada pada kedalaman 4.

Klasifikasi menggunakan metode CART memiliki tiga tahapan. Pertama, metode ini membentuk pohon klasifikasi dengan prosedur pembentukan menggunakan pemilahan simpul secara berulang (*recursive*). Tahap kedua adalah pemangkasan pohon klasifikasi (*pruning*) yang menghasilkan rangkaian pohon klasifikasi yang lebih sederhana. Tahap terakhir, penentuan pohon klasifikasi optimal, dimana pohon klasifikasi tersebut dapat mempresentasikan informasi dari data namun tidak berlebihan (*overfitting*).

2.4.1 Pembentukan Pohon Klasifikasi Optimal

Pembentukan pohon klasifikasi diawali dengan menentukan variabel dan nilai dari variabel tersebut yang layak dijadikan pemilah bagi setiap simpul. Menurut Brieman, *et. al.* (1993), proses pembentukan pohon klasifikasi terdiri dari tiga tahap, yaitu pemilihan pemilah, penentuan simpul terminal, dan penandaan label kelas.

a. Pemilihan Pemilah

Pada tahap pemilahan pemilah dilakukan pemilahan pada sampel data *training* berdasarkan aturan pemilahan dan kriteria *goodness of split,* dimana sampel data *training* yang digunakan masih bersifat heterogen. Pemilihan pemilah tergantung pada jenis pohon atau pada jenis variabel respon. Hasil dari proses pemilahan harus lebih homogen dibandingkan dengan simpul

induknya. Tingkat keheterogenan simpul tersebut dapat diukur menggunakan nilai *impurity* atau *r*(*t*). Aturan pemilahan simpul induk menjadi dua simpul anak berdasarkan pada nilai yang berasal dari satu variabel prediktor. Setiap pemilahan hanya bergantung pada satu variabel prediktor saja. Apabila variabel prediktornya merupakan variabel kontinu, maka pemilahan yang diperbolehkan adalah $x_n \leq y_m$ dan $x_n > y_m$ dengan m = 1,2,3,…,n-1 dengan y_m adalah nilai tengah atau median dari dua nilai amatan sampel yang berbeda dan berurutan. Sehingga jika terdapat sejumlah *m* sampel yang memiliki nilai berbeda pada variabel x_m, maka terdapat pemilahan yang berbeda. Namun jika variabel prediktornya merupakan variabel kategorik, maka pemilahan berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua simpul yang saling lepas (*disjoint*). Fungsi heterogenitas yang sering digunakan adalah indeks gini. Penggunaan indeks gini dalam pemilahan pemilah memiliki kelebihan, yaitu proses perhitungannya sederhana dan relatif cepat, serta mudah untuk diterapkan dalam berbagai kasus (Brieman, *et al*., 1993). Fungsi Indeks Gini dituliskan dalam persamaan berikut.

$$r(t) = \sum_{c_0=1}^{C_0} \sum_{c_1=1}^{C_1} p(c_0|t)p(c_1|t) = 1 - \sum_{i=0}^{1} (p_i)^2 \,,\ c_0 \neq c_1 \quad (2.9)$$

Keterangan:
$r(t)$ = Indeks gini (fungsi heterogenitas) pada simpul *t.*
$p(c_0|t)$ = Proporsi kelas 0 pada simpul *t.*
$p(c_1|t)$ = Proporsi kelas 1 pada simpul *t.*

Langkah selanjutnya yaitu menentukan pemilah terbaik dari setiap variabel prediktor, pemilah terbaik adalah pemilah yang memaksimalkan ukuran homogenitas setiap simpul anak terhadap simpul induknya dan juga memaksimalkan ukuran pemilahan antara dua simpul anak tersebut. Setiap pemilahan akan dilakukan pada setiap simpul sampai diperoleh simpul akhir.

Kemudian, menentukan kriteria *goodness of split* yang merupakan suatu evaluasi pemilahan yang dilakukan oleh pemilah *s* pada suatu simpul *t.* Rumus untuk mencari nilai

goodness of split dituliskan pada persamaan (2.10) sebagai berikut.

$$\phi(s,t) = \Delta_1(s,t) = r(t) - p_L r(t_L) - p_R r(t_R) \quad (2.10)$$

Keterangan:
$\phi(s,t)$ = Nilai *goodness of split.*
$r(t)$ = Fungsi keheterogenan pada simpul *t.*
p_L = Proporsi pengamatan simpul kiri.
p_R = Proporsi pengamatan simpul kanan.
$r(t_L)$ = Fungsi keheterogenan pada simpul kiri.
$r(t_R)$ = Fungsi keheterogenan pada simpul kanan.

Pemilah yang menghasilkan nilai *goodness of split* tertinggi merupakan pemilah terbaik karena dapat menghasilkan heterogenitas lebih tinggi. Setiap variabel akan menghasilkan skor untuk menunjukkan seberapa besar variabel tersebut memberikan kontribusi dalam proses pembentukan pohon. Berikut ini merupakan persamaan untuk menentukan besarnya skor setiap variabel.

$$skor = \sum_{g=1}^{G} \phi(s,tg) \quad (2.11)$$

Dimana $\phi(s,tg)$ merupakan nilai *goodness of split* pada setiap simpul. Nilai skor diperoleh dengan menjumlahkan nilai *goodness of split* (*improvement*) dari masing-masing variabel yang berperan sebagai *surrogate* untuk setiap simpul (CART Reference Guide, 2000).

b. Penentuan Simpul Terminal

Suatu simpul *t* dikatakan sebagai simpul terminal ketika tidak terdapat penurunan heterogenitas yang signifikan, atau hanya terdapat satu pengamatan di setiap simpul anak, atau terdapat batasan minimum *m* pengamatan di setiap simpul anak yang dihasilkan (Breiman, *et al.*, 1993).

c. Penandaan Label Kelas

Penandaan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak seperti yang ditujukan dalam persamaan (2.12) berikut.

$$p(c_i|t) = \max p(c_i|t) = max\,\frac{M_{c_i}(t)}{M(t)} \qquad (2.12)$$

Keterangan:

$p(c_i|t)$ = Proporsi kelas *c* pada simpul *t.*

$M_j(t)$ = Jumlah pengamatan kelas *j* pada simpul *t.*

$M(t)$ = Jumlah total pengamatan pada simpul *t.*

Label kelas untuk simpul terminal *t* adalah c_i yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul *t* yang paling kecil, yaitu sebesar $\mathrm{r}(t) = 1 - max\ p(c_i|t)$.

2.4.2 Pemangkasan Pohon Klasifikasi

Pemangkasan pohon klasifikasi atau bisa disebut *prunning* perlu dilakukan karena semakin banyak pemilahan yang dilakukan mengabkibatkan makin kecilnya tingkat kesalahan prediksi atau dengan kata lain nilai prediksi melebihi nilai yang sebenarnya (*overfitting*). Pemangkasan pohon dilakukan dengan menentukan *cost complexity* minimum (Breiman, *et al*., 1993). Ukuran *cost complexity* adalah sebagai berikut.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \qquad (2.13)$$

Keterangan:

$R_\alpha(T)$ = Ukuran kompleksitas suatu pohon T pada kompleksitas α.

$R(T)$ = Penduga pengganti (*resubstitution estimate*) pohon atau ukuran kesalahan klasifikasi pohon T.

α = Parmeter *cost complexity* bagi penambah satu simpul terminal pada pohon T.

$|\tilde{T}|$ = Banyaknya simpul terminal pada pohon T.

Cost complexity pruning digunakan untuk menentukan pohon bagian $T(\alpha)$ yang dapat meminimumkan $R_\alpha(T)$ pada pohon bagian atau setiap nilai α. Nilai parameter kompleksitas (α) akan secara perlahan meningkat selama proses pemangkasan. Selanjutnya, pencarian pohon bagian $T(\alpha) < T_{maks}$ yang dapat meminimumkan $R_\alpha(T)$. Pemangkasan pohon dimulai dengan mengambil t_R dan t_L dari T_{maks} yang dihasilkan dari simpul induk t. Jika diperoleh dua simpul anak dari proses pemilahan yang dilakukan pada simpul induk yang memenuhi persyaratan $R(t) = R(t_R) + R(t_L)$, maka dua simpul anak akan dipangkas. Sehingga diperoleh pohon T_1 yang memenuhi kriteria $R(T_1) = R(T_{maks})$. Proses ini terus dilakukan secara berulang hingga tidak mungkin lagi dilakukan pemangkasan. Jika $R(T)$ digunakan sebagai kriteria penentuan pohon klasifikasi optimal, maka nilai penduga pengganti akan cenderung memilih pohon besar T_1, karena semakin besar pohon, semakin kecil nilai penduga penggantinya.

Hasil yang diperoleh dari tahap pemangkasan akan berupa urutan pohon yaitu $T_{maks} > T_1 > T_2 > \cdots > \{t_1\}$. Urutan pohon tersebut memiliki nilai α yang semakin menurun, yaitu $\alpha_j < \alpha_{j+1}$ dimana $\alpha_1 = 0$ untuk $j \geq 1$ dan $T(\alpha) = T(\alpha_j) = T_j$.

2.4.3 Penentuan Pohon Klasifikasi Optimal

Ukuran pohon klasifikasi yang sangat besar memberikan nilai penduga yang sangat kecil, sehingga pohon tersebut lebih dipilih untuk menduga nilai respon. Namun, ukuran pohon yang besar bersifat *overfitting* sehingga menyebabkan nilai kompleksitas yang tinggi. Oleh karena itu struktur data yang digambarkan cenderung kompleks. Maka diperlukan pemilihan pohon yang optimum dengan ukuran yang lebih sederhana dan memberikan nilai penduga pengganti kecil.

Penduga pengganti yang sering digunakan jika ukuran pengamatan besar adalah *test sample estimate*. Prosedur ini diterapkan dengan membagi sampel *S* menjadi dua bagian, yaitu B_1 (*training*) dan B_2 (*testing*). Pengamatan B_1 digunakan untuk membentuk pohon *T*, sedangkan pengamatan B_2 digunakan untuk

menduga $R(T)$. M_1 merupakan jumlah pengamatan B_1 dan M_2 merupakan jumlah pengamatan B_2. *X(.)* bernilai 0 jika pernyataan dalam tanda kurung salah dan bernilai 1 jika pernyataan dalam tanda kurung benar. Penduga sampel uji dapat ditunjukkan dalam persamaan (2.14) berikut.

$$R^{ts}(T_t) = \frac{1}{M_2}\sum_{(y_m,c_i)\epsilon S_2}^{M} X(d\,(y_m) \neq c_i) \qquad (2.14)$$

Keterangan:
$R^{ts}(T_t)$ = Total proporsi kesalahan *test sample estimate.*
M_2 = Jumlah pengamatan dari data *testing* (L_2).

Dalam hal ini ingin menduga proporsi kesalahan yang dihasilkan dari proses pembentukan pohon klasifikasi, sehingga pohon klasifikasi optimal yang dipilih adalah pohon T_t yang memiliki nilai penduga sampel uji minimum atau $R^{ts}(T_t) = min_t\, R^{ts}(T_t)$.

2.5 Evaluasi Ketepatan Klasifikasi

Pada penelitian ini evaluasi ketepatan klasifikasi yang digunakan adalah nilai *total accuracy rate*, *sensitivity, specificity*, dan *area under curve* (AUC). *Total accuracy rate* merupakan proporsi observasi yang diprediksi secara benar oleh fungsi klasifikasi (Johnson & Winchern, 2007). Menurut Han dan Kamber (2006), s*ensitivity* mengukur proporsi yang benar-benar positif, sedangkan *specificity* mengukur proporsi yang benar-benar negatif. Metode klasifikasi yang baik harusnya mampu mengukur *sensitivity* dan *specificity* sama baiknya. Sedangkan AUC merupakan *area under curve*, dalam hal ini kurva yang dimaksud adalah kurva ROC (*receiver operating characteristic*). AUC mengakomodasi *sensitivity* dan *specificity*, sehingga AUC memiliki kelebihan lebih baik untuk digunakan pada data *unbalance* daripada *total accuracy rate* (Bekkar, *et al*., 2013). Berikut merupakan *cross tabulation* untuk menghitung ketepatan klasifikasi yang ditunjukkan pada Tabel 2.1.

Tabel 2.1 Tabulasi Silang Ketepatan Klasifikasi

Kelas Pengamatan Y	Kelas Prediksi Y		Total
	1	2	
1	m_{11}	m_{12}	$M_{1.}$
2	m_{21}	m_{22}	$M_{2.}$
Total	$M_{.1}$	$M_{.2}$	M

Keterangan:

m_{11} = jumlah pengamatan dari variabel *Y* kelas 1 yang tepat diprediksi sebagai variabel *Y* kelas 1.

m_{12} = jumlah pengamatan dari variabel *Y* kelas 1 yang salah diprediksi sebagai variabel *Y* kelas 2.

m_{21} = jumlah pengamatan dari variabel *Y* kelas 2 yang salah diprediksi sebagai variabel *Y* kelas 1.

m_{22} = jumlah pengamatan dari variabel *Y* kelas 2 yang tepat diprediksi sebagai variabel *Y* kelas 2.

$M_{1.}$ = jumlah pengamatan dari variabel *Y* kelas 1.

$M_{2.}$ = jumlah pengamatan dari variabel *Y* kelas 2.

$M_{.1}$ = jumlah prediksi dari variabel *Y* kelas 1.

$M_{.2}$ = jumlah prediksi dari variabel *Y* kelas 2.

M = jumlah total pengamatan atau prediksi.

Berikut merupakan persamaan untuk menghitung *total accuracy rate*, *sensitivity*, *specificity*, dan AUC.

$$Total\ Accuracy\ Rate = \frac{m_{11}+m_{22}}{M} \tag{2.15}$$

$$Sensitivity = \frac{m_{11}}{M_{1.}} \tag{2.16}$$

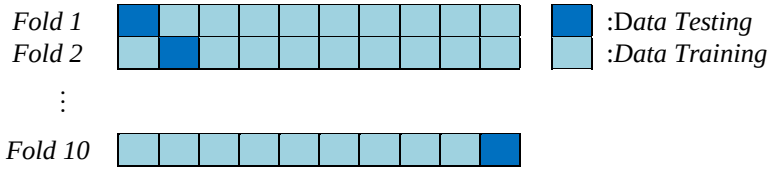
$$Specifity = \frac{m_{22}}{M_{2.}} \tag{2.17}$$

$$AUC = \frac{1}{I}\sum_{c=1}^{C} R_i, \text{ dimana } R_i = \frac{m_{ii}}{\sum_{i=1}^{i} n_{iI}}, i = 1,2, \dots, I \tag{2.18}$$

2.6 Metode Validasi

Metode validasi yang digunakan adalah *k-fold cross validation* yang kemudian disimbolkan dengan $cv(k)$. Dalam $cv(k)$, data sampel dibagi secara random menjadi *k* bagian dengan jumlah kasus pada setiap bagian sama (sedekat mungkin

jumlahnya pada setiap bagian) dan dilakukan pengulangan sebanyak *k* kali. Nilai *k* yang sering digunakan adalah 10 sehingga menjadi 10-*fold cross validation* atau yang kemudian dinotasikan sebagai $cv(10)$. Menurut Witten, *et al.*, (2011), nilai 10 sering digunakan karena menghasilkan estimasi *error* yang paling baik dan membagi data menjadi proporsi yang seimbang. Selain itu, menurut Berthold & Hand (2010) nilai *k* sebesar 10 adalah yang paling memadai untuk mendapatkan perkiraan kesalahan terbaik. Data akan dibagi menjadi 10 bagian, dimana 9 bagian sebagai data *training* dan 1 bagian sebagai data *testing*, kemudian dilakukan pengulangan hingga 10 kali, sehingga setiap data memiliki peluang menjadi data *training* maupun data *testing* (Witten, Frank, & Hall, 2011). Berikut ini merupakan ilustrasi pembagian data pada metode 10-*fold cross validation*.



Gambar 2.2 Ilustrasi Prosedur 10-*fold Cross Validation*
(Sumber: Witten, *et al.*, 2011)

Validasi dengan menggunakan $cv(10)$ diterapkan pada semua metode evaluasi ketepatan klasifikasi yang digunakan yaitu *total accuracy rate*, *sensitivity*, *specificity*, dan AUC. Jika indeks *a* menyatakan data *testing* ke-*a*, maka:

- Persamaan untuk validasi menggunakan $cv(10)$ untuk *total accuracy rate* sebagai berikut.

$$cv(10) = \frac{1}{10}\sum_{a=1}^{10} Total\ Accuracy\ Rate_a \qquad (2.19)$$

- Persamaan validasi menggunakan $cv(10)$ untuk *sensitivity* sebagai berikut.

$$cv(10) = \frac{1}{10}\sum_{a=1}^{10} Sensitivity_a \qquad (2.20)$$

- Persamaan validasi menggunakan $cv(10)$ untuk *specificity* sebagai berikut.

$$cv(10) = \frac{1}{10}\sum_{a=1}^{10} Specifity_a \qquad (2.21)$$

- Persamaan validasi menggunakan $cv(10)$ untuk AUC sebagai berikut.

$$cv(10) = \frac{1}{10}\sum_{a=1}^{10} AUC_a \qquad (2.22)$$

2.7 Distribusi *Mixture*

Jika diketahui suatu data, tidak selamanya satu distribusi saja dapat merepresentasikan data tersebut. Namun, apabila ada indikasi beberapa komposisi muncul dari data tersebut, maka tidak menutup kemungkinan bahwa distribusi data yang lebih tepat adalah distribusi *mixture* (Dempster, Laird, & Rubin, 1977).

Salah satu indikasi adanya sifat *mixture* pada distribusi data univariabel adalah jika hasil *goodness of fit test* pada data tersebut tidak memberi bantuan yang meyakinkan untuk mengidentifikasi distribusi datanya secara tepat (Iriawan, 2001). Misalnya *P-value* dari uji distribusinya cukup rendah walaupun lebih besar dari nilai α (*Type-I error*) yang digunakan. Cara lain untuk melihat indikasi mixture dari distribusi data adalah melihat histogram dari data tersebut. Jika pada histogram muncul beberapa puncak atau modus (biasa disebut multimodal), hal ini mengindikasikan data tidak berasal dari satu populasi yang homogen. Sehingga distribusi mixture bisa jadi lebih sesuai untuk merepresentasikan data sesungguhnya. Persamaan (2.23) merupakan fungsi densitas dari *mixture distribution*.

$$f\left(y \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right) = \sum_{j=1}^{K} \pi_j f\left(y_i \mid \mu_j, \sigma_j^2\right) \qquad (2.23)$$

Pada Persamaan (2.23) π_j menyatakan proporsi tiap komponen distribusi dan $f\left(y_i \middle| \mu_j, \sigma_j^2\right)$ menyatakan pdf dari *mixture* model.

2.8 Algoritma *Expectation Maximization* (EM)

Algoritma *Expectation Maximization* (EM) pertama kali diperkenalkan oleh Dempster, et al (1977). Algoritma EM merupakan suatu metode iteratif yang digunakan untuk estimasi parameter dalam fungsi likelihood dengan pendekatan *Maximum Likelihood Estimation* (MLE) pada data yang tidak lengkap. Menurut McLachlan & Krishnan (1996) Algoritma EM digunakan untuk melakukan estimasi parameter yang tidak dapat diselesaikan dengan MLE dengan menerapkan konsep likelihood *complete-data* berdasarkan likelihood *incomplete-data* yang tidak memiliki penyelesaian secara analitik.

Menurut Benaglia, *et al.*, (2009) algoritma EM menggunakan konsep memaksimalkan secara iterasi, daripada log-likelihood yang diamati. Dalam menyusun algoritma EM, langkah pertama yang dilakukan adalah menurunkan fungsi *likelihood* pada Persamaan (2.24), terhadap mean μ_j dari komponen Gaussian.

$$L(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,|\,\mathbf{y}) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu_j}{\sigma_j} \right)^2 \right) \right) \quad (2.24)$$

$$\frac{\partial L(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,|\,\mathbf{y})}{\partial \mu_j} = -\sum_{i=1}^{N} \frac{\pi_j f\left(y_i \,|\, \mu_j, \sigma_j\right)}{\sum_{l=1}^{K} \pi_l f\left(y_i \,|\, \mu_l, \sigma_l\right)} \frac{y_i - \mu_j}{\sigma_j^2} = 0 \quad (2.25)$$

dimana $\sum_{l=1}^{K} \pi_l f\left(y_i \,|\, \mu_l, \sigma_l\right)$ merupakan variabel laten z_{ij}, dengan ekspektasi probabilitas posterior:

$$z_{ij}^{(t)} = \frac{\pi_j^{(t)} f\left(y_i \,|\, \mu_j^{(t)}, \Sigma_j^{(t)}\right)}{\sum_{l=1}^{K} \pi_i^{(t)} f\left(y_i \,|\, \mu_l^{(t)}, \Sigma_l^{(t)}\right)} \quad (2.26)$$

dengan t mengindikasikan langkah iterasi. Solusi dari $\partial L(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,|\,\mathbf{y})/\partial\mu_j = 0$ menghasilkan nilai minimum dari μ_j pada langkah iterasi ke *t* adalah sebagai berikut.

$$\mu_j^{(t)} = \frac{\sum_{i=1}^{N} z_{ij}^{(t-1)} y_i}{\sum_{i=1}^{N} z_{ij}^{(t-1)}} \tag{2.27}$$

Selanjutnya, jika fungsi *likelihood* para persamaan (2.24) diturunkan terhadap σ_j , maka didapatkan:

$$\sigma_j^{(t)} = \frac{\sum_{i=1}^{N} z_{ij}^{(t-1)} \left(y_i - \mu_j^{(t)}\right)^2}{\sum_{i=1}^{N} z_{ij}^{(t-1)}} \tag{2.28}$$

Kemudian, dapat diturunkan fungsi *likelihood* $L(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,|\,\mathbf{y})$ terhadap distribusi prior π_j. Pada tahap ini, domain $0 \le \pi_j \le 1$ harus dipertimbangkan dan $\sum_{j=1}^{K} \pi_j = 1$. Hal ini dapat dicapai dengan menggunakan sebuah *Lagrange multiplier* η dan memaksimumkan turunan berikut.

$$\frac{\partial}{\partial \pi_j}\left[L(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}\,|\,\mathbf{y}) - \eta\left(\sum_{j=1}^{K} \pi_j - 1\right)\right] = 0 \tag{2.30}$$

dimana selanjutnya didapatkan:

$$\sum_{i=1}^{N} \frac{f\left(y_i \,|\, \mu_j, \sigma_j\right)}{\sum_{l=1}^{K} \pi_l f\left(y_i \,|\, \mu_l, \sigma_l\right)} - \eta = 0 \tag{2.29}$$

Jika kedua sisi dikalikan dengan π_j dan digunakan domain $0 \le \pi_j \le 1$ dan $\sum_{j=1}^{K} \pi_j = 1$, maka diperoleh $\eta = N$. Dengan demikian, untuk mengeliminasi η dan menyusun kembali turunan terhadap π_j didapatkan:

$$\pi_j^{(t)} = \frac{1}{N} \sum_{i=1}^{N} z_{ij}^{(t-1)} \tag{2.30}$$

Algoritma EM adalah sebagai berikut.

1. Inisialisasi parameter $\{\mu_j, \sigma_j^2, \pi_j\}$, mean μ_j; nilai varians σ_j^2 ; dan distribusi prior π_j.
2. E-Step : Evaluasi nilai z_{ij} pada persamaan (2.26) menggunakan nilai parameter awal.
3. M-Step : Estimasi ulang parameter $\{\mu_j, \sigma_j^2, \pi_j\}$.
   a) *Update* mean μ_j dengan menggunakan persamaan (2.27).
   b) *Update* varians σ_j^2 dengan menggunakan persamaan (2.28).
   c) *Update* distribusi prior π_j dengan menggunakan persamaan (2.30).
4. Evaluasi fungsi log-*likelihood* pada persamaan (2.24) dan periksa konvergen, baik dari fungsi log-*likelihood* maupun nilai parameter. Jika kriteria konvergen belum terpenuhi, kembali ke langkah 2.

2.9 Kanker

Kanker merupakan istilah umum untuk sekelompok besar penyakit yang ditandai dengan pertumbuhan sel abnormal yang kemudian dapat menyerang bagian tubuh di sekitar sel tersebut

dan / atau menyebar ke organ lain (World Health Organization, 2009). Definisi lain dipaparkan oleh *National Cancer Institute* (2015) yaitu kanker adalah penyakit yang menyebabkan sel membelah secara abnormal tanpa kontrol dan dapat menyerang jaringan di sekitarnya. Price, *et al.* (2006) mendefinisikan kanker merupakan neoplasma ganas yang menyerang sel. Neoplasma yang bersifat ganas merupakan pertumbuhan sel baru yang tumbuh secara berlebihan dan tidak terkoordinasikan dengan pertumbuhan jaringan normal. Neoplasma terkadang disebut dengan istilah tumor.

2.9.1 Kategori Kanker

Terdapat lebih dari 100 tipe kanker yang biasanya nama dari kanker tersebut berdasarkan organ atau jaringan yang terserang kanker. Tipe-tipe kanker tersebut dikategorikan dalam beberapa kategori sebagai berikut (National Cancer Institute, 2015).

a. *Carcinoma*

Carcinoma adalah jenis kanker yang dibentuk oleh sel epitel, yaitu sel-sel yang menutupi bagian dalam dan luar permukaan tubuh.

b. *Sarcoma*

Sarcoma adalah kanker yang terbentuk pada jaringan tulang dan lunak. Contohnya adalah otot, lemak, pembuluh darah, pembuluh getah bening, dan jaringan fibrosa (tendon dan ligamen).

c. *Leukimia*

Leukimia merupakan kanker yang dimulai pada jaringan pembentuk darah sumsum tulang. Kanker ini tidak membentuk tumor padat. Rendahnya tingkat sel darah normal dapat membuat tubuh lebih sulit mendapatkan oksigen ke jaringannya, mengendalikan pendarahan, atau melawan infeksi.

d. *Lymphoma*

Lymphoma adalah kanker yang dimulai pada limfosit, yaitu sel darah putih yang melawan penyakit. Limfosit abnormal terbentuk di kelenjar getah bening dan pembuluh getah bening serta di organ tubuh lainnya.

e. *Multiple Myeloma*

Multiple myeloma adalah kanker yang dimulai pada sel plasma. Sel *myeloma* (sel plasma abnormal) terbentuk di sumsum tulang dan membentuk tumor di tulang.

f. *Melanoma*

Melanoma adalah kanker yang dimulai pada sel yang menjadi melanosit, yaitu sel khusus yang membuat melanin (pigmen yang memberi warna kulit). Kebanyakan *melanoma* terbentuk pada kulit, namun melanoma juga bisa terbentuk pada jaringan berpigmen lainnya seperti mata.

g. *Brain and Spinal Cord Tumors*

Terdapat berbagai jenis tumor otak dan tulang belakang. Tumor ini dinamai berdasarkan sel dimana tumor tersebut terbentuk pada sistem saraf pusat. Contohnya adalah tumor *astrocytic* dimulai pada sel otak berbentuk bintang.

2.9.2 Faktor Penyebab Kanker dan Radioterapi

Menurut Kementerian Kesehatan RI (2015), faktor-faktor penyebab penyakit kanker antara lain faktor genetik, faktor karsinogen, dan faktor perilaku atau gaya hidup. Faktor genetik merupakan faktor bawaan atau keturunan. Faktor karsinogen meliputi zat kimia, radiasi, virus, hormon, dan iritasi kronis. Sedangkan faktor perilaku atau gaya hidup merupakan kebiasaan yang tidak menyehatkan seperti merokok, pola makan yang tidak sehat, konsumsi alkolol, dan kurang aktivitas fisik.

Terapi radiasi atau radioterapi merupakan salah satu pengobatan kanker. Radioterapi dapat digunakan untuk mengobati kanker primer stadium awal hingga stadium lanjut

dengan menggunakan sinar X atau sinar gamma. Radioterapi dapat diberikan dari luar tubuh (*external beam*) dan dari luar tubuh (*brachytherapy*). *External beam* dilakukan dengan mengarahkan radiasi pada kanker dan jaringan di sekitarnya, sedangkan *brachytherapy* dilakukan dengan memasukkan bahan radioaktif ke dalam tabung tipis dan ditempatkan di dekat sel kanker (Cancer Council Australia, 2018).

(*Halaman ini sengaja dikosongkan*)

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Sumber data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari penelitian Ariyasu, *et al.* (2014) yang berjudul "*Design and Synthesis of 8-Hydroxyquinoline-Based Radioprotective Agents*". Data tersebut merupakan data penyusun senyawa yang berhubungan dengan gen supresor P53. Pada penelitian yang dilakukan oleh Ariyasu, *et al*. (2014) dilakukan dua percobaan, data yang digunakan pada penelitian ini adalah data yang diperoleh dari percobaan kedua. Percobaan tersebut dilakukan dengan memberikan senyawa pada sel yang telah terkena radiasi sinar gamma (10 Gy), percobaan ini mampu mengukur proteksi radiasi. Jika tingkat kematian sel kanker di percobaan tersebut tinggi, maka senyawa tersebut bisa digunakan sebagai radioprotektor.

3.2 Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini terdiri dari satu variabel respon dan 217 variabel prediktor. Variabel-variabel tersebut dituliskan dalam Tabel 3.1 sebagai berikut.

Tabel 3.1 Variabel yang Digunakan dalam Penelitian

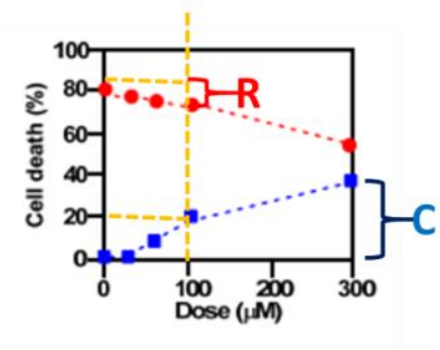
Variabel	Nama Variabel	Skala
Respon (Y)	Target kelas $Y(0)$ = Proteksi radiasi rendah (tingkat kematian sel kanker ≤10%) $Y(1)$ = Proteksi radiasi tinggi (tingkat kematian sel kanker >10%)	Ordinal

Tabel 3.1 Variabel yang Digunakan dalam Penelitian (Lanjutan)

Variabel	Nama Variabel	Skala
Respon (Y)	Tingkat Kematian Sel Kanker	Rasio
Prediktor (X)	x_1 = pKa	Rasio
	x_2 = ALogP98	Rasio
	x_3 = ALogP MR	Rasio
	x_4 = ES Sum aaaC	Rasio
	x_5 = ES Sum aaCH	Rasio
	x_6 = ES Sum aaN	Rasio
	x_7 = ES Sum aaO	Rasio
	x_8 = ES Sum aasC	Rasio
	x_9 = ES Sum ddsN	Rasio
	x_{10} = ES Sum ddssS	Rasio
	x_{11} = ES Sum dO	Rasio
	x_{12} = ES Sum dsN	Rasio
	⋮	
	x_{216} = Wiener	Rasio
	x_{217} = Zagreb	Rasio

Variabel respon yang digunakan untuk klasifikasi adalah variabel respon dengan skala ordinal, sedangkan untuk pengelompokan menggunakan variabel respon berskala rasio. Variabel prediktor yang digunakan dalam penelitian ini terdiri atas lima tipe prediktor, yaitu struktur, ALogP (indikator sulubilitas lemak), ukuran dan berat, energi, serta prediktor lain-lain. Jumlah dari masing-masing variabel prediktor tersebut dihitung dengan menggunakan *Discovery Studio*, yaitu *software*

yang bisa memodelkan *chemical properties* secara 3D. Sedangkan variabel respon merupakan target kelas yang terdiri dari 2 kelas (biner). Klasifikasi didasarkan pada tingkat kematian sel, kelas 0 untuk proteksi radiasi dengan tingkat kematian sel kurang dari 10% dan kelas 1 untuk proteksi radiasi dengan tingkat kematian sel lebih dari 10%. Persentase senyawa yang tergolong kelas 0 adalah 53,57% atau sebanyak 45 dan kelas 1 sebesar 46,43% atau sebesar 39 senyawa. Batas atau *threshold* kematian sel sebesar 10% dipilih berdasarkan titik potong dari tingkat kematian sel normal sebesar 20% dan dosis senyawa sebesar $100\mu M$. Berikut ini adalah gambar pada penelitian Ariyasu *et al* (2014) dalam menentukan *threshold* untuk kelas klasifikasi pada proteksi radiasi.



Gambar 3.1 Penentuan threshold dalam kelas toksisitas
Sumber : Ariyasu, *et al* (2014)

C merupakan *threshold* untuk toksisitas dan R merupakan *threshold* untuk proteksi radiasi. Penentuan *threshold* untuk proteksi radiasi dilakukan dengan perhitungan tingkat kematian sel kanker pada dosis $0\mu M$ dikurangi dengan tingkat kematian sel kanker pada saat tingkat kematian sel normal sebesar 20%,

sehingga didapatkan *threshold* tingkat kematian sel untuk proteksi radiasi yaitu sebesar 10%.

Struktur data yang digunakan dalam penelitian ini ditunjukkan di Tabel 3.2. Observasi yang digunakan dalam penelitian ini adalah data senyawa yang mengandung beberapa jenis zat penyusun. Sedangkan variabel respon (Y) merupakan hasil klasifikasi senyawa dan variabel prediktor (X) merupakan zat penyusun dari senyawa.

Tabel 3.2 Struktur Data Penelitian

Observasi	Variabel Respon (Y)	Variabel Prediktor (X)				
		X_1	X_2	X_3	…	X_{217}
1	Y_1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	…	$X_{217,1}$
2	Y_2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	…	$X_{217,2}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
84	Y_{84}	$X_{1,84}$	$X_{2,84}$	$X_{3,84}$	…	$X_{217,84}$

3.3 Langkah Analisis

Sebelum melakukan analisis dengan metode yang ditentukan, dilakukan eksplorasi data untuk mengetahui karakteristik dari data senyawa obat kanker. Langkah analisis yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Melakukan *feature selection* dengan menghitung *feature importance* dengan langkah sebagai berikut.
   a. Menghitung *feature importance* berdasarkan nilai *mean decrease gini* (MDG).
   b. Memilih *features* yang akan digunakan dengan menggunakan hasil perhitungan *feature importance*, yaitu sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35% dan 100% dari keseluruhan *features* berdasarkan nilai MDG.

2. Melakukan klasifikasi senyawa berdasarkan dua kelas proteksi radiasi menggunakan metode *naïve bayes* dan CART menggunakan dua kelas awal proteksi radiasi.
   I. Melakukan klasifikasi senyawa berdasarkan dua kelas proteksi radiasi menggunakan metode *naïve bayes* dengan langkah sebagai berikut.
      a. Membagi data menjadi 10 bagian (k = 10).
      b. Menghitung nilai probabilitas untuk setiap *features* ke *n*, $n = 1,2,\ldots,N$. Apabila data berupa data numerik (kontinu), maka akan dihitung nilai rata-rata dan deviasi standar dari masing-masing variabel pada setiap kategori. Selanjutnya adalah menentukan nilai probabilitas menggunakan pendekatan distribusi normal.
      c. Menghitung probabilitas posterior dari setiap senyawa terhadap setiap kategori.
      d. Menentukan kelas berdasarkan nilai probabilitas posterior yang tertinggi.
      e. Mengevaluasi ketepatan klasifikasi dengan menggunakan *total accuracy rate*, *sensitivity*, *specificity*, dan AUC.
      f. Melakukan validasi dengan menggunakan $cv(10)$.
   II. Melakukan klasifikasi senyawa berdasarkan dua kelas proteksi radiasi menggunakan metode *classification and regression trees* (CART) dengan langkah sebagai berikut.
      a. Membagi data menjadi 10 bagian (k = 10).
      b. Membentuk pohon klasifikasi optimal dengan beberapa tahapan sebagai berikut.
         i. Melakukan pemilahan berdasarkan aturan pemilahan indeks Gini yang kemudian dilakukan evaluasi dari hasil pemilahan menggunakan *goodness of split*.

ii. Menentukan *terminal nodes* dengan cara menghentikan pembentukan pohon hingga mencapai batasan minimum pengamatan dalam *terminal nodes* sebanyak satu pengamatan.
iii. Melakukan penandaan label kelas pada *terminal nodes*.

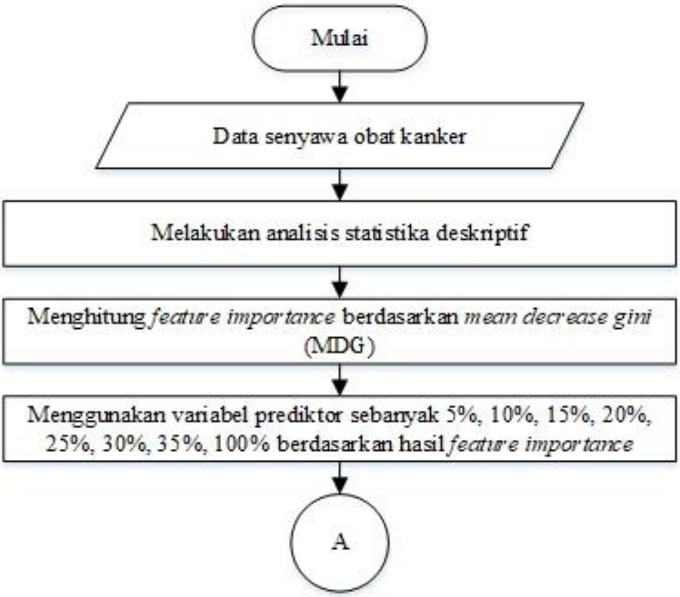
c. Memangkas pohon klasifikasi sampai diperoleh ukuran pohon klasifikasi yang paling kecil dengan kriteria kompleksitas yang minimum.
d. Memilih pohon terbaik.
e. Mengevaluasi ketepatan klasifikasi dengan menggunakan *total accuracy rate, sensitivity, specificity,* dan AUC.
f. Melakukan validasi dengan menggunakan $cv(10)$.

3. Mengelompokkan senyawa obat kanker dengan menggunakan pendekatan *mixture distribution*.
   a. Menentukan jumlah pembagian data berdasarkan distribusi yang ada (*K*).
   b. Memilih jumlah distribusi yang optimal berdasarkan nilai *log-likelihood* terbesar.
   c. Menghitung nilai *posterior probability*.
   d. Mengelompokkan senyawa obat kanker ke dalam kategori yang memiliki *posterior probability* terbesar.
4. Melakukan klasifikasi senyawa berdasarkan kelas baru hasil pengelompokan dengan pendekatan *mixture distribution* menggunakan metode *naïve bayes* dan CART
   I. Melakukan klasifikasi senyawa berdasarkan kelas baru hasil pengelompokan dengan pendekatan *mixture distribution* menggunakan metode *naïve bayes* dengan langkah pada poin (2.I).

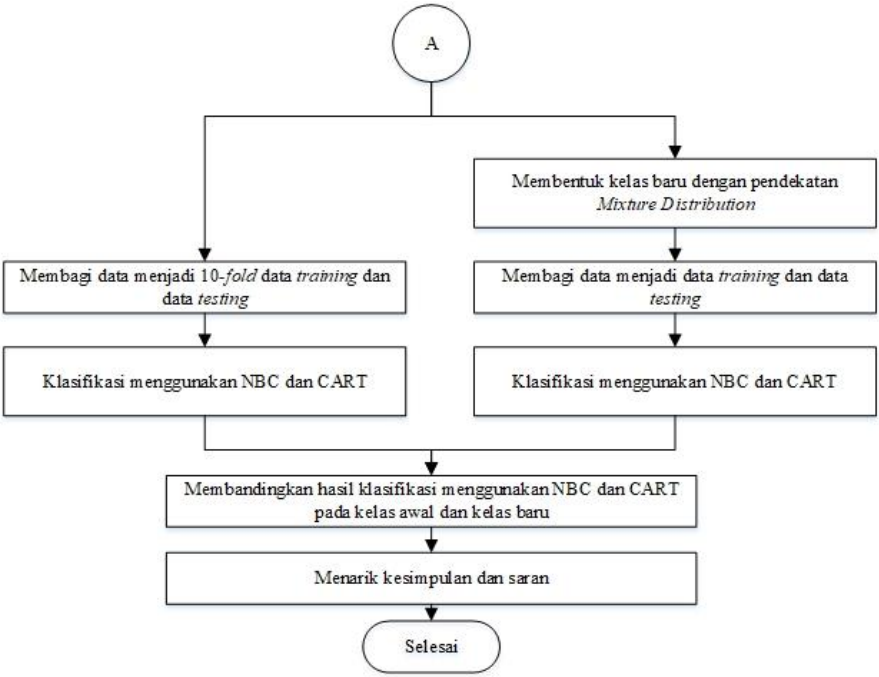
II. Melakukan klasifikasi senyawa berdasarkan kelas baru hasil pengelompokan dengan pendekatan *mixture distribution* menggunakan metode *classification and regression tree* dengan langkah pada poin (2.II).

5. Membandingkan hasil klasifikasi dengan metode *naïve bayes* dan *classification and regression trees* (CART) menggunakan kelas awal dan kelas baru.

Langkah-langkah analisis secara umum dapat digambarkan dalam diagram alir pada Gambar 3.2 sebagai berikut.



Gambar 3.2 Diagram Alir Langkah Analisis Secara Umum



Gambar 3.2 Diagram Alir Langkah Analisis Secara Umum (Lanjutan)

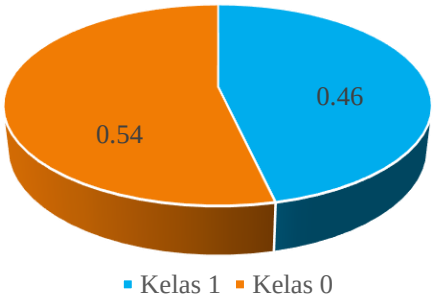
BAB IV

ANALISIS DAN PEMBAHASAN

Pada penelitian ini dilakukan klasifikasi dengan *feature selection* menggunakan *mean decrease gini* (MDG) *index*. Klasifikasi yang dilakukan berdasarkan dua kelas proteksi radiasi dan hasil pengelompokan menggunakan pendekatan *normal mixture distribution*. Metode klasifikasi yang digunakan adalah *naïve bayes classifier* dan *classification and regression tree* (CART). Kemudian akan dilakukan perbandingan nilai ketepatan klasifikasi dari setiap metode yang digunakan.

4.1 Karakteristik Senyawa Obat Kanker

Analisis karakteristik data perlu dilakukan untuk mengetahui karakter data yang akan dianalisis, sehingga *treatment* pada data dan penggunaan metode bisa disesuaikan dengan karakter data yang ada. Pada penelitian ini variabel respon adalah target kelas untuk klasifikasi, kelas 0 untuk proteksi radiasi rendah dan kelas 1 untuk proteksi radiasi tinggi. Dalam mengklasifikasikan senyawa obat kanker, perlu diketahui proporsi klasifikasi aktual per kelas untuk mengetahui keseimbangan (*balance* atau *unbalance*) dari data yang akan digunakan. Berikut merupakan proporsi masing-masing kelas dari 2 kelas proteksi radiasi.



Gambar 4.1 Proporsi Tiap Kelas Proteksi Radiasi

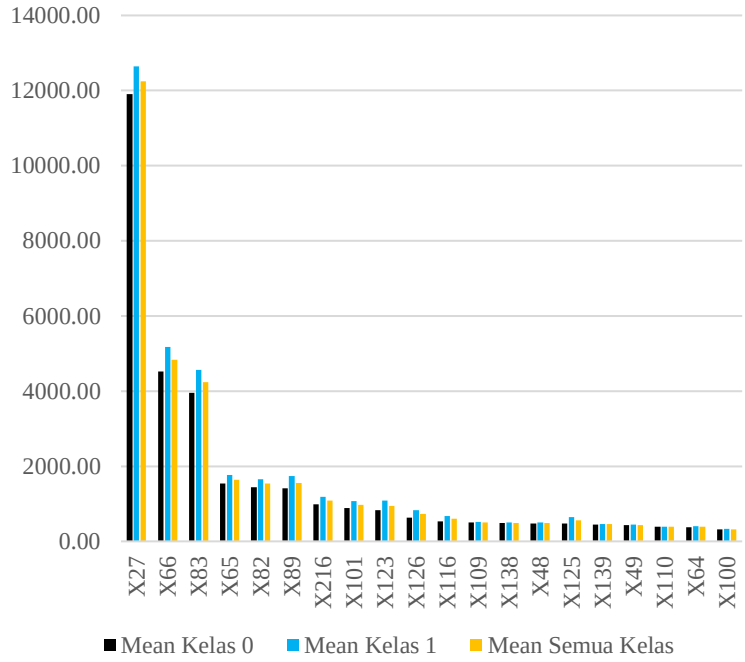
Pada Gambar 4.1 diketahui persentase senyawa obat kanker yang terklasifikasikan di kelas 0 adalah sebesar 54% atau sejumlah 45 dan di kelas 1 sebesar 46% atau sejumlah 39 dari total senyawa obat kanker sebanyak 84. Artinya, proporsi klasifikasi di kedua kelas proteksi radiasi seimbang atau *balance*.

Variabel prediktor yang digunakan pada penelitian ini adalah sebanyak 217 prediktor yang terdiri atas lima tipe prediktor, yaitu struktur, ALogP (indikator sulubilitas lemak), ukuran dan berat, energi, serta prediktor lain-lain. Berikut merupakan rata-rata dari setiap variabel prediktor per kelas klasifikasi proteksi radiasi.

Tabel 4.1 Rata-Rata Setiap Variabel Prediktor per Kelas Klasifikasi Proteksi Radiasi

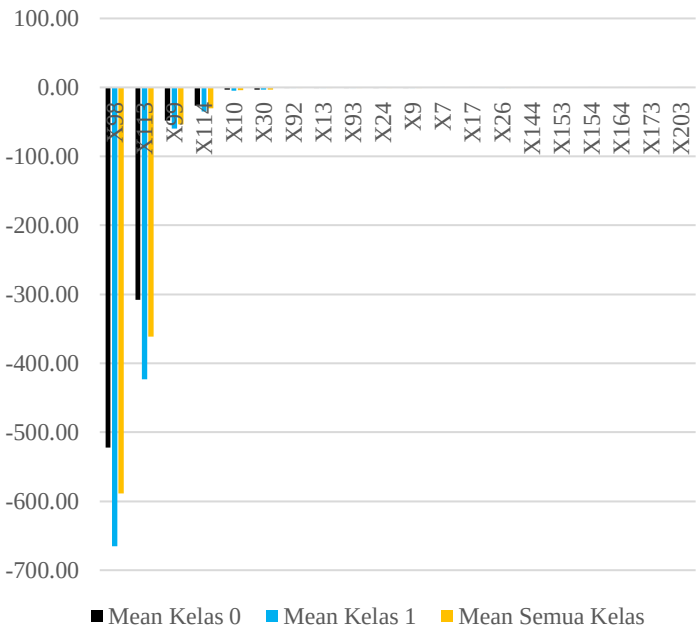
Variabel	Kelas 0	Kelas 1	Semua Kelas
X1	13.63	13.50	13.57
X2	1.54	1.37	1.46
X3	76.58	80.03	78.18
X4	0.74	0.66	0.70
X5	8.35	6.67	7.57
X6	4.21	4.09	4.15
X7	0.00	0.25	0.12
X8	0.57	0.38	0.48
X9	-0.04	-0.02	-0.03
X10	-3.48	-4.80	-4.10
…	…	…	…
X216	989.58	1189.49	1082.39
X217	103.29	111.03	106.88

Berdasarkan hasil perhitungan rata-rata setiap variabel prediktor, dapat diketahui bahwa terdapat 157 variabel yang memiliki rata-rata kelas 0 di bawah rata-rata keseluruhan, sedangkan 60 variabel lainnya memiliki rata-rata kelas 0 di atas rata-rata keseluruhan. Berikut merupakan grafik variabel prediktor dengan rata-rata terbesar dan terkecil.



Gambar 4.2 Variabel Prediktor dengan Rata-Rata Terbesar

Gambar 4.2 menunjukkan bahwa 20 variabel prediktor dengan rata-rata terbesar adalah Apol (X27), E_DIST_mag (X66), V_DIST_mag (X83), E_DIST_equ (X65), V_DIST_equ (X82), Jurs_DPSA_2 (X89), Wiener (X216), Jurs_PPSA_2 (X101), PMI_mag (X123), PMI_Z (X126), Jurs_WPSA_2 (X116), Jurs_SASA (X109), Molecular_3D_SASA (X138), Molecular_SASA (X48), PMI_Y (X125), Molecular_3d_SAVol (X139), Molecular_SAVol (X49), Jurs_TASA (X110), E_ADJ_mag (X64), dan Jurs_PPSA_1 (X100). Sedangkan variabel prediktor dengan rata-rata terkecil ditunjukkan pada Gambar 4.3 sebagai berikut.



Gambar 4.3 Variabel Prediktor dengan Rata-Rata Terkecil

Berdasarkan Gambar 4.3 diketahui bahwa 20 variabel prediktor dengan rata-rata terkecil adalah Jurs_PNSA_2 (X98), Jurs_WNSA_2 (X113), Jurs_PNSA_3 (X99), Jurs_WNSA_3 (X114), ES_Sum_ddssS (X10), Molecular_Solubility (X30), Jurs_FNSA_2 (X92), ES_Sum_dssC (X13), Jurs_FNSA_3 (X93), ES_Sum_sssCH (X24), ES_Sum_ddsN (X9), ES_Sum_aaO (X7), ES_Sum_sF (X17), ES_Sum_ssssC (X26), F_Count (X144), ES_Count_aaNH (X153), ES_Count_aaO (X154), ES_Count_SF (X164), ES_Count_ssssC (X173), Num_Rings7 (X203).

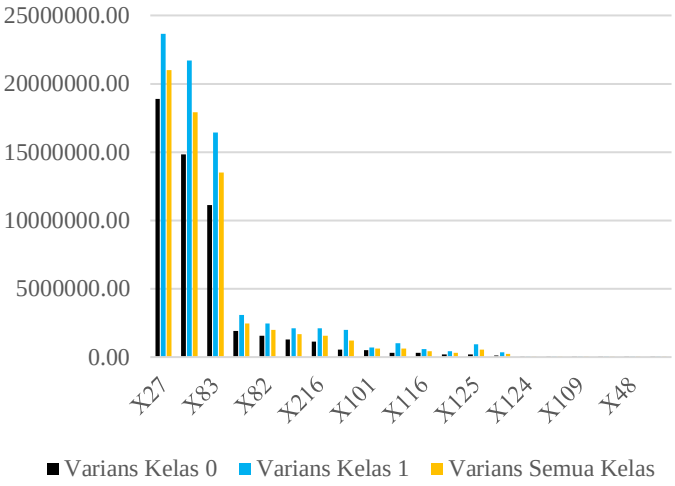
Selanjutnya, untuk mengetahui persebaran data setiap variabel per kelas klasifikasi proteksi radiasi, maka dilakukan penghitungan varians. Varians dari setiap variabel prediktor per kelas klasifikasi proteksi radiasi ditunjukkan pada Tabel 4.2 sebagai berikut.

Tabel 4.2 Varians Setiap Variabel Prediktor per Kelas Klasifikasi Proteksi Radiasi

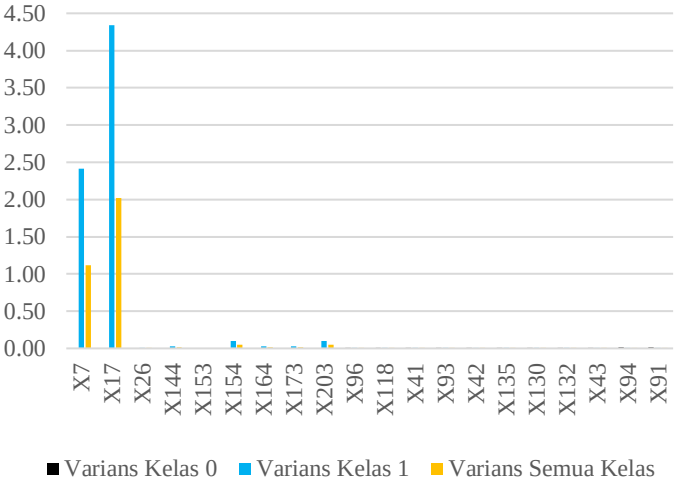
Variabel	Kelas 0	Kelas 1	Semua Kelas
X1	42.53	32.18	37.28
X2	1.78	3.27	2.45
X3	723.38	887.67	792.88
X4	0.52	0.44	0.48
X5	27.26	11.54	20.44
X6	8.27	3.32	5.91
X7	0.00	2.41	1.12
X8	1.57	1.97	1.74
X9	0.06	0.02	0.04
X10	13.94	19.08	16.56
…	…	…	…
X216	1138254.66	2099541.31	1574707.61
X217	1498.21	1837.45	1650.54

Berdasarkan hasil perhitungan varians setiap variabel prediktor, dapat diketahui bahwa terdapat 142 variabel yang memiliki varians kelas 0 di bawah varians keseluruhan, sedangkan 75 variabel lainnya memiliki varians kelas 0 di atas varians keseluruhan. Berikut merupakan grafik variabel prediktor dengan varians terbesar dan terkecil.

Gambar 4.4 menunjukkan bahwa 7 variabel prediktor dengan varians terbesar sama dengan 7 variabel prediktor dengan rata-rata terbesar, kemudian adalah PMI_mag (X123), Jurs_PPSA_2 (X101), PMI_Z (X126), Jurs_PMSA_2 (X116), Jurs_PNSA_2 (X98), PMI_Y (X125), Jurs_WNSA_2 (X113), PMI_X (X124), E_ADJ_mag (X64), Jurs_SASA (X109), Molecular_SASA (X48), dan E_ADJ_equ (X63). Sedangkan variabel prediktor dengan rata-rata terkecil ditunjukkan pada Gambar 4.5 sebagai berikut.



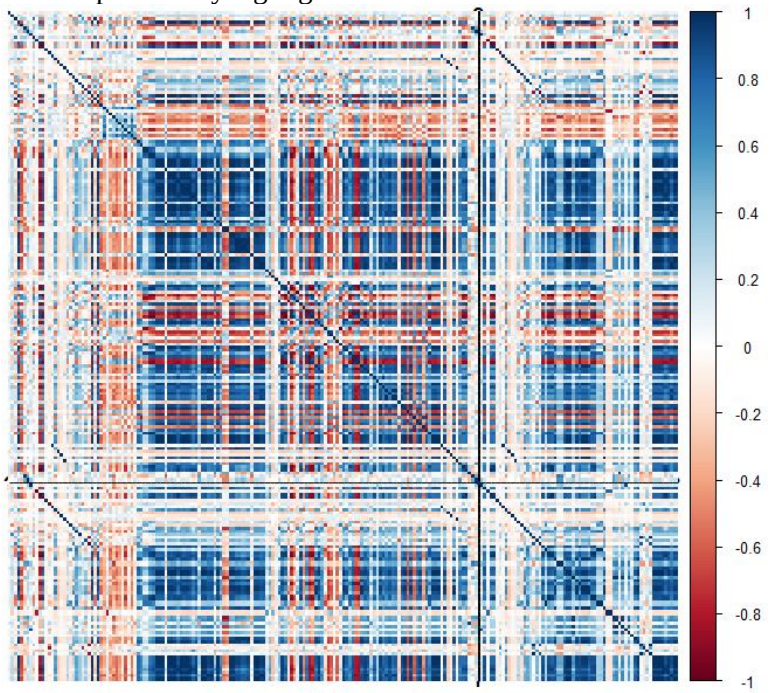
Gambar 4.4 Variabel Prediktor dengan Varians Terbesar



Gambar 4.5 Variabel Prediktor dengan Varians Terkecil

Berdasarkan Gambar 4.5 diketahui bahwa 20 variabel prediktor dengan varians terkecil yaitu Jurs_PNSA_2 (X98), Jurs_WNSA_2 (X113), Jurs_PNSA_3 (X99), Jurs_WNSA_3 (X114), ES_Sum_ddssS (X10), Molecular_Solubility (X30), Jurs_FNSA_2 (X92), ES_Sum_dssC (X13), Jurs_FNSA_3 (X93), ES_Sum_sssCH (X24), ES_Sum_ddsN (X9), ES_Sum_aaO (X7), ES_Sum_SF (X17), ES_Sum_ssssC (X26), F_Count (X144), ES_Count_aaNH (X153), ES_Count_aaO (X154), ES_Count_sF (X164), ES_Count_ssssC (X173), dan Num_Rings7 (X203).

Sebelum mengolah data, perlu diketahui korelasi antar variabel untuk melihat hubungan antara variabel satu dengan variabel lainnya. *Correlation plot* digunakan untuk mengetahui korelasi antar variabel. Berikut merupakan *correlation plot* dari variabel prediktor yang digunakan.



Gambar 4.6 *Correlation Plot* Variabel Prediktor

Berdasarkan Gambar 4.2, dapat dilihat bahwa terdapat beberapa variabel yang memiliki korelasi yang kuat. Variabel yang memiliki korelasi paling tinggi terbanyak terhadap variabel lain yaitu variabel Jurs PPSA 2, SC 2, Apol, CHI 2, E ADJ Equ, E ADJ Mag, IAC total, Kappa 1, dan V ADJ Equ, dimana variabel-variabel tersebut memiliki nilai korelasi kurang dari -0.7 atau lebih dari 0,7 terhadap 104 variabel yang lain. Sedangkan variabel yang tidak memiliki korelasi terhadap variabel lain yaitu pKa, Jurs RPCG, Jurs RPCS, Minimized Energy, ES Count aaNH, Num Terminal Rotomers, QED ALogP, QED ROTB, SA Score, Dipole Mag, Dipole X, Dipole Y, Dipole Z, dan Jurs FPSA 3.

4.2 *Feature Selection*

Proses seleksi variabel atau *feature selection* dilakukan dengan tujuan untuk mengurangi variabel yang digunakan dalam analisis dengan harapan variabel terpilih dapat membangun prediksi lebih baik serta mempercepat proses komputasi. Metode yang digunakan untuk *feature selection* adalah *mean decrease gini* (MDG), dimana MDG merupakan ukuran kepentingan variabel berdasarkan *gini impurity index* yang digunakan untuk perhitungan *splits* selama pelatihan atau *training*. Variabel yang digunakan pada penelitian ini adalah sebesar 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% variabel terpenting dari keseluruhan variabel dengan rincian sebagai berikut.

Tabel 4.3 Jumlah Variabel yang Digunakan

Persentase	Jumlah Variabel
5%	11
10%	22
15%	33
20%	44
25%	55
30%	66
35%	76
100%	217

Mean decrease gini (MDG) merupakan salah satu ukuran tingkat kepentingan variabel yang dihasilkan dari metode *random forest*. Parameter yang digunakan untuk menghitung MDG dari *random forest* ditunjukkan sebagai berikut.

Tabel 4.4 Kandidat Nilai untuk *Parameter Tuning*

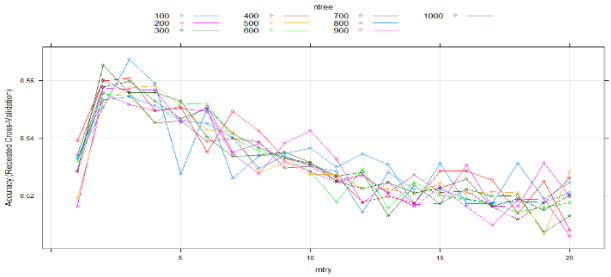
Parameter	Kandidat Nilai
mtry	1, 2, 3, …, 19, 20
ntree	100, 200 300, …, 1000

Berdasarkan kandidat nilai tersebut, selanjutnya adalah menentukan nilai untuk parameter mtry dan ntree yang optimal yang menghasilkan akurasi yang tinggi. Berikut merupakan akurasi untuk setiap kandidat nilai parameter.

Tabel 4.5 Hasil *Parameter Tuning*

mtry	ntree	Akurasi	Kappa
1	100	0,533	0,056
1	200	0,516	0,022
…	…	…	…
3	100	0,567	0,132
…	…	…	…
20	900	0,506	0,009
20	1000	0,520	0,038

Hasil *parameter tuning* yang dituliskan pada Tabel 4.5 digambarkan pada Gambar 4.7 sebagai berikut.



Gambar 4.7 Grafik Hasil *Parameter Tuning*

Hasil *parameter tuning* pada Tabel 4.5 menunjukkan bahwa parameter yang digunakan adalah mtry atau jumlah variabel yang diambil secara acak sebagai kandidat pada setiap pembagian dengan nilai optimal 3 dan ntree atau jumlah pohon yang optimal sebanyak 100 pohon. Dengan menggunakan parameter tersebut, selanjutnya dilakukan perhitungan MDG untuk mengetahui kepentingan dari setiap variabel. Hasil dari perhitungan MDG ditunjukkan pada Tabel 4.6 sebagai berikut, yang telah diurutkan dari nilai MDG terbesar hingga terkecil.

Tabel 4.6 Hasil Perhitungan *Mean Decrease Gini*

Variabel	***Mean Decrease Gini***
X91	0,582
X48	0,573
X120	0,509
X6	0,488
X87	0,479
…	…
X207	0
X209	0

Selanjutnya ada tahap klasifikasi dengan metode *naïve bayes* dan *classification and regression tree* (CART), digunakan variabel yang diambil sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% variabel berdasarkan nilai *mean decrease gini* terbesar sesuai pada Tabel 4.6. Lima variabel dengan nilai MDG terbesar adalah variabel Jurs_FNSA_1 (X91), Molecular_SASA (X48), Minimized_Energy (X120), ES_Sum_aaN (X6), dan Dipole_Z (X87), sedangkan lima variabel dengan nilai MDG terkecil adalah ES_Count_sF (X164), ES_Count_sI (X165), Num_Rings5 (X201), Num_StereoBonds (X207), dan Num_TrueStereoAtoms (X209). Variabel yang digunakan terdapat pada Lampiran 2.

4.3 Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Awal Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dan CART

Pada bagian ini diljelaskan mengenai klasifikasi senyawa obat kanker pada dua kelas awal proteksi radiasi. Metode yang digunakan adalah *naïve bayes classifier* dan *classification and regression tree* (CART). Kemudian dilakukan perbandingan nilai ketepatan klasifikasi dari kedua metode tersebut dan metode-metode yang digunakan pada penelitian sebelumnya yang dilakukan oleh Matsumoto, *et al.*, (2016).

4.3.1 Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Awal Proteksi Radiasi Menggunakan *Naïve Bayes Classifier*

Metode pertama yang akan digunakan dalam klasifikasi senyawa obat kanker adalah *naïve bayes classifier*. Klasifikasi menggunakan metode *naïve bayes classifier* menghasilkan probabilitas suatu senyawa obat kanker terhadap setiap kelas, kemudian dipilih probabilitas terbesar dari kedua probabilitas tersebut untuk mengklasifikasikan senyawa obat kanker. Probabilitas tersebut didapatkan dari persamaan (2.3).

Pada ilustrasi klasifikasi menggunakan *naïve bayes* berikut ini menggunakan 84 observasi atau senyawa dan 5% variabel prediktor hasil *feature selection*. Langkah pertama adalah menghitung probabilitas *prior*. Proteksi radiasi kelas 0 terdiri dari 45 observasi dan kelas 1 terdiri dari 39 observasi. Berikut merupakan probabilitas *prior* dari masing-masing kelas yang didapatkan dari jumlah observasi pada masing-masing kelas dibagi dengan keseluruhan observasi.

Tabel 4. 7 Nilai Probabilitas *Prior*

Kelas	Probabilitas *Prior*
0	0,536
1	0,464

Langkah selanjutnya yaitu menghitung nilai peluang bersyarat. Pada variabel prediktor yang bersifat kontinu, nilai peluang bersyarat dihitung menggunakan pendekatan distribusi normal sehingga diperlukan nilai rata-rata dan standar deviasi dari masing-masing variabel independen pada setiap kelas proteksi radiasi. Pada Tabel 4.8 disajikan rata-rata dan standar deviasi.

Tabel 4. 8 Rata-Rata dan Standar Deviasi Tiap Kelas Proteksi Radiasi

Variabel	Kelas 0		Kelas 1	
	Mean	***St. Deviation***	***Mean***	***St. Deviation***
X_{91}	0.377	0.101	0.376	0.088
X_{48}	481.106	131.990	499.421	147.256
X_{120}	1.715	3.956	5.658	14.880
X_{6}	4.205	2.875	4.090	1.822
X_{87}	0.164	1.026	-0.058	1.311
…	…	…	…	…
X_{70}	2.535	0.428	2.631	0.411

Kemudian, nilai rata-rata dan standar deviasi tersebut digunakan untuk mencari nilai peluang parsial dengan menggunakan pendekatan distribusi normal sesuai dengan Persamaan (2.8). Selanjutnya dilakukan perkalian dari setiap nilai peluang parsial dari setiap observasi per kelas untuk mendapatkan nilai *likelihood*.

Tabel 4. 9 Nilai *Likelihood* dan *Posterior Probability* dari Setiap Observasi per Kelas

Senyawa	Kelas 0		Kelas 1	
	Likelihood	***Posterior Prob.***	***Likelihood***	***Posterior Prob.***
AS-1	$7{,}38\times10^{-7}$	$3{,}95\times10^{-7}$	$2{,}22\times10^{-7}$	$1{,}03\times10^{-7}$
AS-10	$2{,}90\times10^{-7}$	$1{,}55\times10^{-7}$	$5{,}03\times10^{-8}$	$2{,}33\times10^{-8}$
AS-11	$6{,}87\times10^{-7}$	$3{,}68\times10^{-7}$	$1{,}32\times10^{-7}$	$6{,}12\times10^{-8}$
AS-12	$8{,}03\times10^{-7}$	$4{,}30\times10^{-7}$	$8{,}92\times10^{-8}$	$4{,}14\times10^{-8}$
AS-13	$8{,}97\times10^{-9}$	$4{,}80\times10^{-9}$	$1{,}58\times10^{-8}$	$7{,}32\times10^{-9}$
…	…	…	…	…
YT-1	$3{,}47\times10^{-14}$	$1{,}86\times10^{-14}$	$1{,}13\times10^{-9}$	$5{,}24\times10^{-10}$

Nilai *posterior probability* didapatkan dari *likelihood* dikalikan dengan *prior probability*. Nilai *posterior probability* digunakan untuk menentukan pengklasifikasian senyawa atau observasi dengan mengklasifikasikan senyawa ke dalam kelas yang memiliki *posterior probability* terbesar. Sebagai contoh senyawa AS-1 diklasifikasikan ke dalam kelas 0 karena *posterior probability* kelas 0 lebih besar daripada kelas 1 yaitu $3{,}95\times10^{-7}$ lebih besar daripada $1{,}03\times10^{-7}$.

Pada penelitian ini dilakukan pembagian data *training* dan data *testing* menggunakan 10-*fold cross validation*. Variabel yang digunakan sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% variabel berdasarkan nilai *mean decrease gini* terbesar.

1. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 5% Prediktor Terpenting

Data pertama adalah data dengan variabel prediktor sebanyak 11 variabel terpenting (5%) yang dibagi menjadi data *training* dan data *testing*. Hasil klasifikasi didasarkan pada kelas yang memiliki peluang terbesar.

Tabel 4.10 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 5% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	5	0
1	2	1

Tabel 4.10 menunjukkan bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 5 senyawa dan tidak ada yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Selanjutnya dilakukan penghitungan *total accuracy*, *sensitivity*, *specificity*,

dan AUC untuk data *testing fold* 10 sesuai dengan persamaan (2.15), (2.16), (2,17), dan (2.18) sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{5+1}{5+0+2+1} = 0{,}75$$

$$Sensitivity = \frac{5}{5+0} = 1$$

$$Specifity = \frac{1}{2+1} = 0{,}33$$

$$AUC = \frac{1}{2}(1 + 0{,}33) = 0{,}667$$

Pada Tabel 4.11 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 5% variabel terpenting.

Tabel 4.11 Performa *Naive Bayes Classifier* dengan 5% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.693	0.875	0.486	0.680
	Testing	0.444	0.600	0.250	0.425
2	*Training*	0.640	0.850	0.400	0.625
	Testing	0.667	1.000	0.250	0.625
3	*Training*	0.587	0.750	0.400	0.575
	Testing	0.444	0.800	0.000	0.400
4	*Training*	0.600	0.875	0.286	0.580
	Testing	0.667	1.000	0.250	0.625
5	*Training*	0.605	0.927	0.229	0.578
	Testing	0.500	0.500	0.500	0.500
6	*Training*	0.592	0.805	0.343	0.574
	Testing	0.625	0.750	0.500	0.625
7	*Training*	0.618	0.854	0.343	0.598
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.579	0.805	0.314	0.560
	Testing	0.500	0.500	0.500	0.500
9	*Training*	0.618	0.878	0.314	0.596
	Testing	0.625	1.000	0.250	0.625
10	*Training*	0.592	0.875	0.278	0.576
	Testing	0.750	1.000	0.333	0.667

Tabel 4.11 Performa *Naive Bayes Classifier* dengan 5% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
Rata-Rata	*Training*	0.613	0.849	0.339	0,594
	Testing	0.572	0.790	0.308	0,549

Hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold* dirata-rata sesuai dengan persamaan (2.19), (2.20), (2.21), dan (2.22). Berdasarkan Tabel 4.11 diketahui bahwa data *testing* dengan variabel prediktor sebanyak 5% memiliki nilai akurasi total sebesar 0,572, *sensitivity* sebesar 0,790, *specificity* sebesar 0,308, dan AUC sebesar 0,549.

2. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 10% Prediktor Terpenting

Persentase variabel prediktor yang digunakan adalah 10% atau sejumlah 22 variabel. Berikut merupakan *confusion matrix* dari data *testing fold* 10 klasifikasi senyawa obat kanker.

Tabel 4.12 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 10% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	5	0
1	2	1

Tabel 4.12 menunjukkan bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 5 senyawa dan tidak ada yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, dan *specificity* untuk data *testing fold* 10 dengan 10% variabel terpenting sama dengan menggunakan

5% variabel terpenting dikarenakan memiliki *confusion matrix* yang sama. Pada Tabel 4.13 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 10% variabel prediktor terpenting.

Tabel 4.13 Performa *Naive Bayes Classifier* dengan 10% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.600	0.725	0.457	0.591
	Testing	0.555	0.600	0.500	0.550
2	*Training*	0.666	0.750	0.571	0.661
	Testing	0.444	0.600	0.250	0.425
3	*Training*	0.613	0.725	0.485	0.605
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.600	0.725	0.457	0.591
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.631	0.829	0.400	0.615
	Testing	0.750	0.500	1.000	0.750
6	*Training*	0.592	0.731	0.428	0.580
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.618	0.780	0.428	0.605
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.552	0.731	0.342	0.537
	Testing	0.750	0.750	0.750	0.750
9	*Training*	0.644	0.804	0.457	0.631
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.578	0.725	0.416	0.571
	Testing	0.750	1.000	0.333	0.667
Rata-Rata	*Training*	0.610	0.755	0.443	0,599
	Testing	0.552	0.700	0.370	0,537

Tabel 4.13 menunjukkan bahwa data *testing* dengan variabel prediktor sebanyak 10% memiliki nilai akurasi total sebesar 0,552, *sensitivity* sebesar 0,7, *specificity* sebesar 0,37, dan AUC sebesar 0,537.

3. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 15% Prediktor Terpenting

Data yang digunakan adalah sejumlah 33 variabel prediktor yang dibagi menjadi data *training* dan data *testing*. Tabel 4.14 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 15% variabel prediktor terpenting.

Tabel 4.14 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 15% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	4	1
1	2	1

Berdasarkan Tabel 4.14 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+1}{5+2+1+1} = 0{,}625$$

$$Sensitivity = \frac{4}{4+1} = 0{,}8$$

$$Specifity = \frac{1}{2+1} = 0{,}333$$

$$AUC = \frac{1}{2}(0{,}8 + 0{,}333) = 0{,}567$$

Pada Tabel 4.15 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 15% variabel terpenting.

Tabel 4.15 Performa *Naive Bayes Classifier* dengan 15% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.653	0.700	0.600	0.650
	Testing	0.667	0.600	0.750	0.675
2	*Training*	0.680	0.700	0.657	0.679
	Testing	0.444	0.600	0.250	0.425
3	*Training*	0.627	0.750	0.486	0.618
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.613	0.750	0.457	0.604
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.645	0.732	0.543	0.637
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.605	0.756	0.429	0.592
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.671	0.780	0.543	0.662
	Testing	0.375	0.500	0.250	0.375
8	*Training*	0.579	0.805	0.314	0.560
	Testing	0.625	0.500	0.750	0.625
9	*Training*	0.632	0.756	0.486	0.621
	Testing	0.375	0.750	0.000	0.375
10	*Training*	0.618	0.725	0.500	0.613
	Testing	0.625	0.800	0.333	0.567
Rata-Rata	*Training*	0.632	0.745	0.501	0,623
	Testing	0.489	0.570	0.383	0,477

Tabel 4.15 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 15% memiliki nilai akurasi total sebesar 0,489, *sensitivity* sebesar 0,570, *specificity* sebesar 0,383 dan AUC sebesar 0,477.

4. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 20% Prediktor Terpenting

Variabel prediktor yang digunakan adalah sebanyak 44 variabel atau 20% dari 217 variabel. Tabel 4.16 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 20% variabel prediktor terpenting.

Tabel 4.16 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 20% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	4	1
1	2	1

Berdasarkan Tabel 4.16 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 dengan variabel prediktor 20% sama dengan menggunakan variabel prediktor 15%. Pada Tabel 4.17 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 20% variabel terpenting.

Tabel 4.17 Performa *Naive Bayes Classifier* dengan 20% Prediktor Terpenting

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
1	*Training*	0.680	0.750	0.600	0.675
	Testing	0.556	0.600	0.500	0.550
2	*Training*	0.640	0.725	0.543	0.634
	Testing	0.556	0.800	0.250	0.525

Tabel 4.17 Performa *Naive Bayes Classifier* dengan 20% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
3	*Training*	0.653	0.750	0.543	0.646
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.613	0.750	0.457	0.604
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.658	0.732	0.571	0.652
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.605	0.756	0.429	0.592
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.592	0.732	0.429	0.580
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.566	0.756	0.343	0.549
	Testing	0.625	0.500	0.750	0.625
9	*Training*	0.592	0.732	0.429	0.580
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.592	0.725	0.444	0.585
	Testing	0.625	0.800	0.333	0.567
Rata-Rata	*Training*	0.619	0.741	0.479	0,610
	Testing	0.514	0.640	0.358	0,499

Tabel 4.17 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 20% memiliki nilai akurasi total sebesar 0,514, *sensitivity* sebesar 0,640, *specificity* sebesar 0,358, dan AUC sebesar 0,499.

5. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 25% Prediktor Terpenting

Persentase variabel prediktor yang digunakan adalah 25% atau sejumlah 55 variabel. Berikut merupakan *confusion matrix* dari data *testing fold* 10 klasifikasi senyawa obat kanker.

Tabel 4.18 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 25% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	4	1
1	2	1

Berdasarkan Tabel 4.18 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Pada Tabel 4.19 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 25% variabel terpenting.

Tabel 4.19 Performa *Naive Bayes Classifier* dengan 25% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.680	0.750	0.600	0.675
	Testing	0.556	0.600	0.500	0.550
2	*Training*	0.600	0.700	0.486	0.593
	Testing	0.667	0.800	0.500	0.650
3	*Training*	0.653	0.750	0.543	0.646
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.627	0.750	0.486	0.618
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.645	0.756	0.514	0.635
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.618	0.756	0.457	0.607
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.579	0.756	0.371	0.564
	Testing	0.500	0.750	0.250	0.500

Tabel 4.19 Performa *Naive Bayes Classifier* dengan 25% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
8	*Training*	0.566	0.780	0.314	0.547
	Testing	0.625	0.500	0.750	0.625
9	*Training*	0.592	0.732	0.429	0.580
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.605	0.725	0.472	0.599
	Testing	0.625	0.800	0.333	0.567
Rata-Rata	*Training*	0.617	0.746	0.467	0,606
	Testing	0.525	0.640	0.383	0,512

Tabel 4.19 menunjukkan hasil total akurasi, *sensitivity*, dan *specificity* dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 25% memiliki nilai akurasi total sebesar 0,525, *sensitivity* sebesar 0,640, *specificity* sebesar 0,383, dan AUC sebesar 0,512.

6. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 30% Prediktor Terpenting

Variabel prediktor yang digunakan adalah sebanyak 66 variabel atau 30% dari 217 variabel. Tabel 4.20 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 30% variabel prediktor terpenting.

Tabel 4.20 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 30% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	4	1
1	2	1

Berdasarkan Tabel 4.20 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan

ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 dengan variabel prediktor 30% sama dengan menggunakan variabel prediktor 15%, 20%, dan 25%. Pada Tabel 4.21 ditunjukkan performa NBC untuk klasifikasi senyawa obat kanker dengan 30% variabel terpenting.

Tabel 4.21 Performa *Naive Bayes Classifier* dengan 30% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.667	0.750	0.571	0.661
	Testing	0.667	0.800	0.500	0.650
2	*Training*	0.613	0.750	0.457	0.604
	Testing	0.667	0.800	0.500	0.650
3	*Training*	0.653	0.750	0.543	0.646
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.627	0.750	0.486	0.618
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.671	0.780	0.543	0.662
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.618	0.756	0.457	0.607
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.566	0.756	0.343	0.549
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.566	0.780	0.314	0.547
	Testing	0.625	0.500	0.750	0.625
9	*Training*	0.592	0.732	0.429	0.580
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.605	0.725	0.472	0.599
	Testing	0.625	0.800	0.333	0.567
Rata-Rata	*Training*	0.618	0.753	0.462	0,607
	Testing	0.536	0.660	0.383	0,522

Tabel 4.21 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 30% memiliki nilai akurasi total sebesar 0,536, *sensitivity* sebesar 0,660, *specificity* sebesar 0,383, dan AUC sebesar 0,522.

7. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 35% Prediktor Terpenting

Data yang digunakan adalah sejumlah 76 variabel prediktor yang dibagi menjadi data *training* dan data *testing*. Tabel 4.22 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 15% variabel prediktor terpenting.

Tabel 4.22 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 35% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	4	1
1	1	2

Berdasarkan Tabel 4.22 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 2 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+2}{4+1+1+2} = 0{,}75$$

$$Sensitivity = \frac{4}{4+1} = 0{,}8$$

$$Specifity = \frac{2}{1+2} = 0{,}667$$

$$Specifity = \frac{1}{2}(0{,}8 + 0{,}667) = 0{,}733$$

Pada Tabel 4.23 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 35% variabel terpenting.

Tabel 4.23 Performa *Naive Bayes Classifier* dengan 35% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.667	0.725	0.600	0.663
	Testing	0.667	0.800	0.500	0.650
2	*Training*	0.667	0.750	0.571	0.661
	Testing	0.667	0.800	0.500	0.650
3	*Training*	0.680	0.750	0.600	0.675
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.653	0.750	0.543	0.646
	Testing	0.556	0.600	0.500	0.550
5	*Training*	0.645	0.780	0.486	0.633
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.658	0.756	0.543	0.649
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.592	0.732	0.429	0.580
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.592	0.780	0.371	0.576
	Testing	0.625	0.500	0.750	0.625
9	*Training*	0.618	0.732	0.486	0.609
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.618	0.725	0.500	0.613
	Testing	0.750	0.800	0.667	0.733
Rata-Rata	*Training*	0.639	0.748	0.513	0,630
	Testing	0.560	0.660	0.442	0,551

Tabel 4.23 menunjukkan hasil total akurasi, *sensitivity, specificity,* dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 35% memiliki nilai akurasi total sebesar 0,560, *sensitivity* sebesar 0,660, *specificity* sebesar 0,442, dan AUC sebesar 0,551.

8. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Naïve Bayes Classifier* dengan 100% Prediktor

Data yang digunakan adalah semua variabel prediktor atau 217 variabel yang kemudian dibagi menjadi data *training* dan data *testing*. Tabel 4.24 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 100% variabel prediktor.

Tabel 4.24 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan *Naive Bayes Classifier* dengan 100% Prediktor

Aktual	Prediksi	
	0	1
0	5	0
1	2	1

Berdasarkan Tabel 4.24 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 5 senyawa dan tidak terdapat senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{5+1}{5+0+2+1} = 0{,}75$$

$$Sensitivity = \frac{5}{5+0} = 1$$

$$Specifity = \frac{1}{2+1} = 0{,}333$$

$$AUC = \frac{1}{2}(1 + 0{,}333) = 0{,}667$$

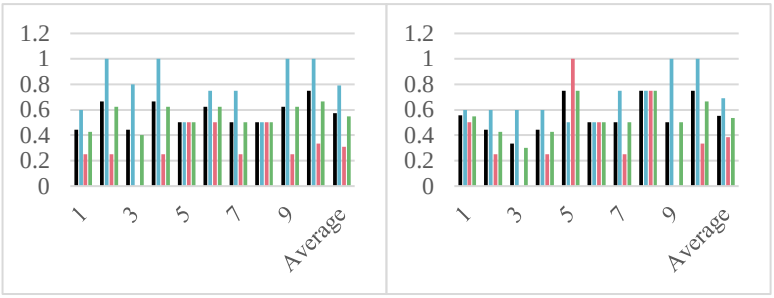
Pada Tabel 4.25 ditunjukkan performa *naïve bayes classifier* untuk klasifikasi senyawa obat kanker dengan 100% variabel.

Tabel 4.25 Performa *Naive Bayes Classifier* dengan 100% Prediktor

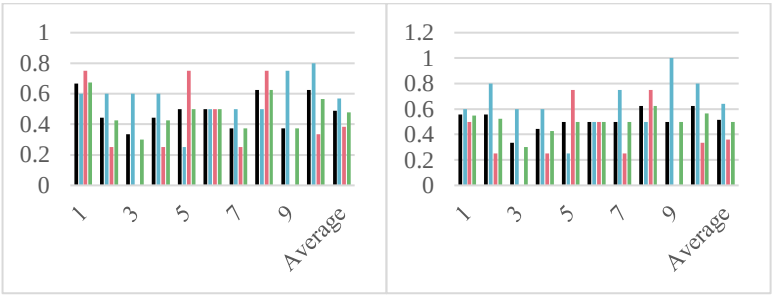
Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.640	0.500	0.800	0.650
	Testing	0.556	0.400	0.750	0.575
2	*Training*	0.587	0.950	0.171	0.561
	Testing	0.556	1.000	0.000	0.500
3	*Training*	0.613	0.875	0.314	0.595
	Testing	0.333	0.600	0.000	0.300
4	*Training*	0.640	0.800	0.457	0.629
	Testing	0.444	0.600	0.250	0.425
5	*Training*	0.566	1.000	0.057	0.529
	Testing	0.500	1.000	0.000	0.500
6	*Training*	0.618	0.756	0.457	0.607
	Testing	0.375	0.500	0.250	0.375
7	*Training*	0.592	0.878	0.257	0.568
	Testing	0.375	0.750	0.000	0.375
8	*Training*	0.566	0.878	0.200	0.539
	Testing	0.500	0.500	0.500	0.500
9	*Training*	0.579	0.854	0.257	0.555
	Testing	0.500	1.000	0.000	0.500
10	*Training*	0.579	0.875	0.250	0.563
	Testing	0.750	1.000	0.333	0.667
Rata-Rata	*Training*	0.598	0.837	0.322	0,579
	Testing	0.489	0.735	0.208	0,472

Tabel 4.25 menunjukkan hasil total akurasi, *sensitivity*, dan *specificity*. Data *testing* dengan variabel prediktor sebanyak 100% memiliki nilai akurasi total sebesar 0,489, *sensitivity* sebesar 0,735, *specificity* sebesar 0,208, dan AUC sebesar 0,472.

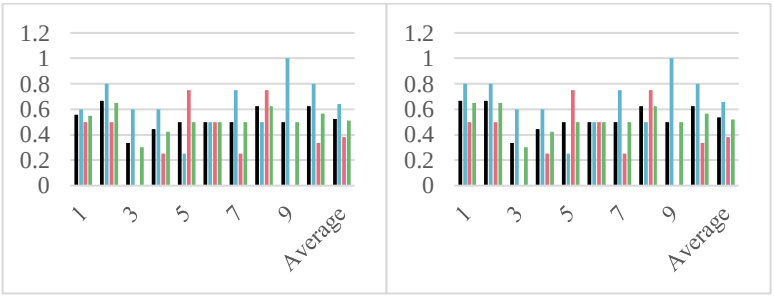
Nilai akurasi total, *sensitivity*, dan *specificity* dari setiap fold untuk jumlah variabel prediktor yang digunakan sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, 100% disajikan dalam Gambar 4.8 sebagai berikut.


1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

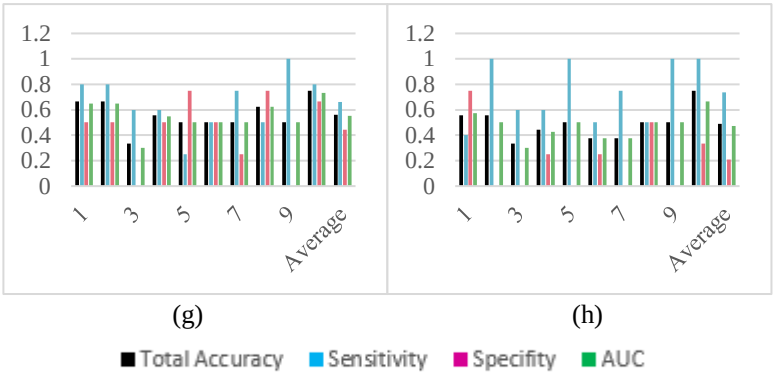
(a) (b)


1
0.8
0.6
0.4
0.2
0
1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

(c) (d)


1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

(e) (f)



Gambar 4.8 Performa *Naïve Bayes Classifier* per Jumlah Prediktor (a) 5% Prediktor (b) 10% Prediktor (c) 15% Prediktor (d) 20% Prediktor (e) 25% Prediktor (f) 30% Prediktor (g) 35% Prediktor (h) 100% Prediktor

Validasi model pada penelitian ini menggunakan 10-*fold cross validation*, yang artinya nilai akurasi total, *sensitivity*, *specificity*, dan AUC setiap *fold* dirata-ratakan untuk setiap jumlah variabel prediktor yang digunakan. Pada Tabel 4.26 disajikan performa metode *naïve bayes classifier* pada setiap jumlah variabel prediktor yang digunakan.

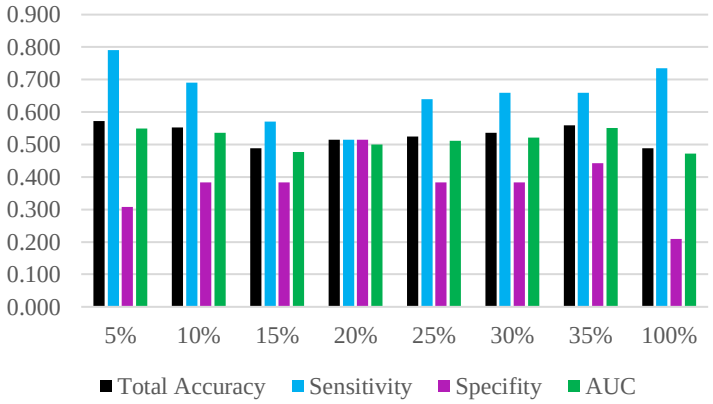
Tabel 4.26 Performa *Naive Bayes Classifier*

Persentase Prediktor	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
5%	*Training*	0.613	0.849	0.339	0,594
	Testing	0.572	0.790	0.308	0.549
10%	*Training*	0.610	0.755	0.443	0,599
	Testing	0.553	0.690	0.383	0.537
15%	*Training*	0.632	0.745	0.501	0,623
	Testing	0.489	0.570	0.383	0.477
20%	*Training*	0.619	0.741	0.479	0,610
	Testing	0.514	0.514	0.514	0.499
25%	*Training*	0.617	0.746	0.467	0,606
	Testing	0.525	0.640	0.383	0.512

Tabel 4.26 Performa *Naive Bayes Classifier* (Lanjutan)

Persentase Prediktor	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
30%	*Training*	0.618	0.753	0.462	0,607
	Testing	0.536	0.660	0.383	0.522
35%	*Training*	0.639	0.748	0.513	0,630
	Testing	0.560	0.660	0.442	0.551
100%	*Training*	0.598	0.837	0.322	0,579
	Testing	0.489	0.735	0.208	0.472

Performa metode *naïve bayes classifier* pada pada data *testing* digambarkan pada Gambar 4.9 sebagai berikut.



Gambar 4.9 Performa *Naïve Bayes Classifier*

Berdasarkan Tabel 4.26 dan Gambar 4.9, dengan menggunakan variabel prediktor sebanyak 5% akan menghasilkan total akurasi yang paling tinggi yaitu 0,572 dan *sensitivity* sebesar 0,79. Sedangkan untuk *specificity* terbesar sebesar 0,383 didapatkan dengan menggunakan variabel prediktor sebanyak 10%, 15%, 25%, dan 30%. Nilai AUC tertinggi didapatkan dengan menggunakan prediktor sebanyak 5% yang menghasilkan nilai AUC sebesar 0,549.

4.3.2 Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan *Classification and Regression Tree* (CART)

Metode kedua yang akan digunakan dalam klasifikasi senyawa obat kanker adalah CART. Secara umum, terdapat 3 tahapan dalam klasifikasi menggunakan CART, yaitu pembentukan pohon klasifikasi optimal, pemangkasan pohon klasifikasi, dan penentuan pohon klasifikasi optimal. Sebelum melakukan klasifikasi, terlebih dahulu dilakukan *parameter tuning* yaitu *complexity parameter* (CP) yang optimal.

Pada ilustrasi berikut, digunakan data 84 observasi atau senyawa dan 5% variabel prediktor hasil *feature selection*. Dalam membentuk pohon klasifikasi diperlukan variabel-variabel yang berperan sebagai pemilah. Jika variabel prediktor berskala kontinu dengan sampel berukuran *n*, maka kemungkinan pemilah yang terbentuk sebanyak (*n-1*) jenis pemilahan yang berbeda. Berikut merupakan banyaknya kemungkinan pemilah untuk membentuk pohon klasifikasi berdasarkan proteksi radiasi sel kanker.

Tabel 4. 27 Banyaknya Kemungkinan Pemilah Variabel Prediktor

Variabel	Skala Data	Jumlah Kategori	Kemungkinan Pemilah
X_{91}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah
X_{48}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah
X_{120}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah
X_{6}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah
X_{87}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah
…	…	…	…
X_{70}	Rasio	2	$2^{2-1} - 1 = 1$ pemilah

Berdasarkan Tabel 4.27 variabel prediktor yang digunakan memiliki kemungkinan pemilah saja karena variabel prediktor berskala data rasio dengan dua kategori. Setelah dilakukan penghitungan kemungkinan pemilah untuk membentuk pohon klasifikasi, tahap selanjutnya yaitu pemilihan pemilah dengan

menggunakan indeks Gini. Berikut merupakan contoh penghitungan indeks Gini. Misalkan akan dihitung indeks Ginipada variabel prediktor X_{134} dengan split 22,29.

Tabel 4. 28 Contoh Hasil Pemilahan Pada Suatu Simpul

Split	**Kelas Proteksi Radiasi**		**Total**
	0	1	
$X_{134} \geq 22{,}29$	42	33	75
$X_{134} < 22{,}29$	3	6	9

Dengan menggunakan Persamaan (2.9) didapatkan nilai indeks Gini pada simpul kiri dan kanan sebagai berikut.

$$r(t_L) = 1 - \left(\left(\frac{42}{75}\right)^2 + \left(\frac{33}{75}\right)^2\right) = 0{,}493$$

$$r(t_R) = 1 - \left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right) = 0{,}444$$

Kemudian, menentukan kriteria *goodness of split* yang merupakan suatu evaluasi pemilahan yang dilakukan oleh pemilah *s* pada suatu simpul *t.* Rumus untuk mencari nilai *goodness of split* dituliskan pada persamaan (2.10). Berikut merupakan nilai *goodness of split* yang didapatkan dari variabel prediktor dengan X_{134} dengan split 22,29.

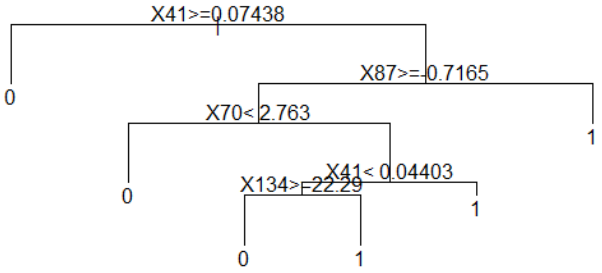
$$\phi(s,t) = 0{,}5 - \left(\frac{75}{84}\right) \times 0{,}493 - \left(\frac{9}{84}\right) \times 0{,}444 = 0{,}0124$$

Nilai *goodness of split* yang didapatkan dari variabel prediktor X_{134} dengan split 22,29 adalah sebesar 0,0124. Nilai *goodness of split* tersebut kemudian akan dibandingkan dengan nilai *goodness of split* dengan menggunakan *split* yang lain. *Split* atau pemilah yang menghasilkan nilai *goodness of split* tertinggi merupakan pemilah terbaik karena dapat menghasilkan heterogenitas lebih tinggi.

Suatu simpul *t* dikatakan sebagai simpul terminal ketika tidak terdapat penurunan heterogenitas yang signifikan, atau hanya terdapat satu pengamatan di setiap simpul anak, atau

terdapat batasan minimum *m* pengamatan di setiap simpul anak yang dihasilkan (Breiman, *et al.*, 1993). Label kelas untuk simpul terminal *t* adalah c_i yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul *t* yang paling kecil, yaitu sebesar $r(t) = 1 - max\, p(c_i|t)$. Maka pada simpul dari variabel prediktor X_{134} dengan split 22,29 adalah 0 ketika $X_{134} \geq 22{,}29$ dan 1 ketika $X_{134} < 22{,}29$.

Selanjutnya dilakukan pemangkasan pohon klasifikasi. Pemangkasan pohon klasifikasi atau bisa disebut *prunning* perlu dilakukan karena semakin banyak pemilahan yang dilakukan mengabkibatkan makin kecilnya tingkat kesalahan prediksi atau dengan kata lain nilai prediksi melebihi nilai yang sebenarnya (*overfitting*). Pemangkasan pohon dilakukan dengan menentukan *cost complexity* minimum (Breiman, *et al*., 1993). Hasil dari *complexity parameter* yang optimal sebesar 0.01994302, hasil tersebut didapatkan dengan 5% variabel prediktor hasil *feature selection* menggunakan MDG. Berikut merupakakan pohon klasifikasi yang dihasilkan.



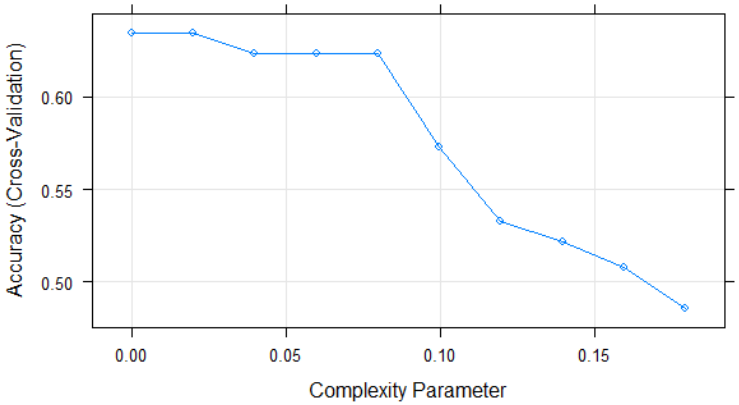
Gambar 4. 10 Contoh Pohon Klasifikasi dengan Menggunakan 5% Prediktor Terpenting

Pohon klasifikasi pada Gambar 4.10 dapat digunakan untuk menentukan kelas dari suatu senyawa. Misalkan senyawa AS-1 yang memiliki prediktor X_{41} sebesar 0,040474, X_{87} sebesar 3,07117, dan X_{70} sebesar 2,83834 akan diklasifikasikan di kelas 1, begitu juga untuk senyawa yang lain.

Sama seperti saat menggunakan metode *naïve bayes*, metode validasi yang digunakan yaitu 10-*fold cross validation*. Variabel yang digunakan juga sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% variabel berdasarkan nilai *mean decrease gini* terbesar.

1. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 5% Prediktor Terpenting

Data pertama adalah data dengan variabel prediktor sebanyak 11 variabel terpenting (5%) yang dibagi menjadi data *training* dan data *testing*. Hasil *parameter tuning* yaitu nilai CP yang optimal sebesar 0.01994302 yang ditunjukkan Gambar 4.11.



Gambar 4.11 Hasil Tuning Complexity Parameter dengan prediktor 5%

Berikut merupakan *confusing matrix* yang dihasilkan dengan menggunakan data tersebut.

Tabel 4.29 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 5% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	4	1
1	0	3

Tabel 4.29 menunjukkan bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 3 senyawa dan tidak ada yang salah diklasifikasikan ke kelas 0. Selanjutnya dilakukan penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 sesuai dengan persamaan (2.15), (2.16), (2.17), dan (2.18) sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+3}{4+1+0+3} = 0{,}875$$

$$Sensitivity = \frac{4}{4+1} = 0{,}8$$

$$Specifity = \frac{3}{0+3} = 1$$

$$AUC = \frac{1}{2}(0{,}8 + 1) = 0{,}9$$

Pada Tabel 4.30 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 5% variabel terpenting.

Tabel 4.30 Performa CART dengan 5% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.813	0.800	0.829	0.814
	Testing	0.667	0.400	1.000	0.700
2	*Training*	0.733	0.600	0.886	0.743
	Testing	0.667	0.600	0.750	0.675
3	*Training*	0.827	0.900	0.743	0.821
	Testing	0.667	0.800	0.500	0.650
4	*Training*	0.800	0.925	0.657	0.791
	Testing	0.667	0.800	0.500	0.650
5	*Training*	0.816	0.927	0.686	0.806
	Testing	0.875	0.750	1.000	0.875
6	*Training*	0.789	0.707	0.886	0.797
	Testing	0.625	0.750	0.500	0.625
7	*Training*	0.803	0.902	0.686	0.794
	Testing	0.250	0.500	0.000	0.250

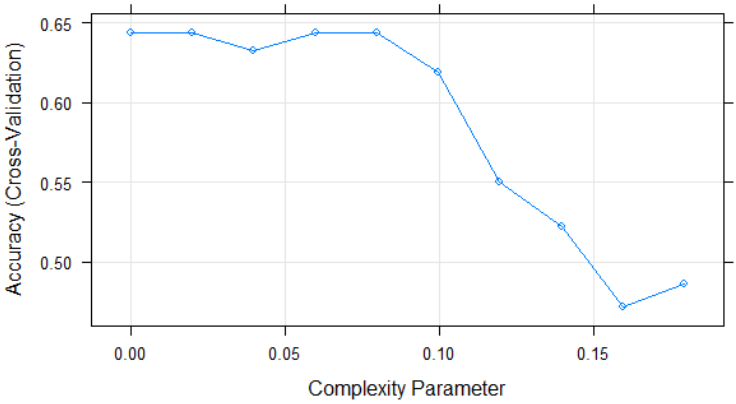
Tabel 4.30 Performa CART dengan 5% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
8	*Training*	0.789	0.854	0.714	0.784
	Testing	0.750	0.750	0.750	0.750
9	*Training*	0.803	0.902	0.686	0.794
	Testing	0.500	0.750	0.250	0.500
10	*Training*	0.816	0.925	0.694	0.810
	Testing	0.875	0.800	1.000	0.900
Rata-Rata	*Training*	0.799	0.844	0.747	0,795
	Testing	0.654	0.690	0.625	0,658

Hasil total akurasi, *sensitivity*, dan *specificity* dari seluruh *fold* dirata-rata sesuai dengan persamaan (2.19), (2.20), (2.21), dan (2.22). Berdasarkan Tabel 4.30 diketahui bahwa data *testing* dengan variabel prediktor sebanyak 5% memiliki nilai akurasi total sebesar 0,654, *sensitivity* sebesar 0,690, *specificity* sebesar 0,625, dan AUC sebesar 0,658.

2. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 10% Prediktor Terpenting

Persentase variabel prediktor yang digunakan adalah 10% atau sejumlah 22 variabel. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.07977208. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.12 Hasil *Tuning Complexity Parameter* dengan prediktor 10%

Berikut merupakan *confusion matrix* dari data *testing fold* 10 klasifikasi senyawa obat kanker.

Tabel 4.31 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 10% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	4	1
1	1	2

Tabel 4.31 menunjukkan bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan ada 1 yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 2 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 0. Selanjutnya dilakukan penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 sesuai dengan persamaan (2.15), (2.16), (2.17), dan (2.18) sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+3}{4+1+1+2} = 0{,}75$$

$Sensitivity = \frac{4}{4+1} = 0{,}8$

$Specifity = \frac{3}{1+2} = 0{,}667$

$AUC = \frac{1}{2}(0{,}8 + 0{,}667) = 0{,}733$

Pada Tabel 4.32 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 10% variabel terpenting.

Tabel 4.32 Performa CART dengan 10% Prediktor Terpenting

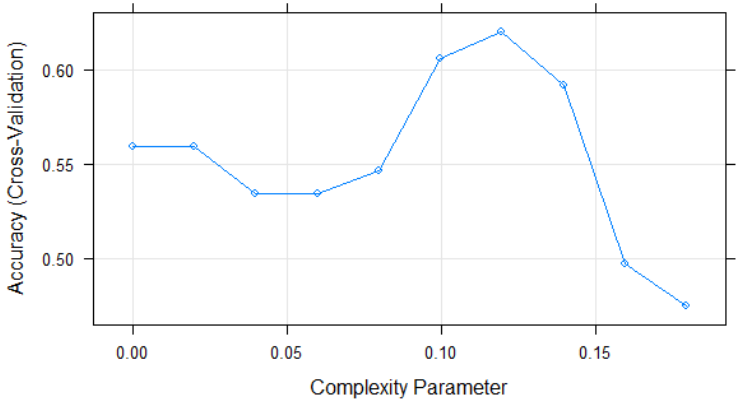
Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.867	0.875	0.857	0.866
	Testing	0.778	0.600	1.000	0.800
2	*Training*	0.773	0.875	0.657	0.766
	Testing	0.444	0.800	0.000	0.400
3	*Training*	0.867	0.850	0.886	0.868
	Testing	0.667	0.800	0.500	0.650
4	*Training*	0.800	0.700	0.914	0.807
	Testing	0.778	0.600	1.000	0.800
5	*Training*	0.855	0.927	0.771	0.849
	Testing	0.750	0.500	1.000	0.750
6	*Training*	0.789	0.707	0.886	0.797
	Testing	0.625	0.750	0.500	0.625
7	*Training*	0.842	0.732	0.971	0.852
	Testing	0.500	0.500	0.500	0.500
8	*Training*	0.803	0.707	0.914	0.811
	Testing	0.750	0.750	0.750	0.750
9	*Training*	0.816	0.878	0.743	0.810
	Testing	0.625	0.750	0.500	0.625
10	*Training*	0.829	0.875	0.778	0.826
	Testing	0.750	0.800	0.667	0.733
Rata-Rata	*Training*	0.824	0.813	0.838	0,825
	Testing	0.667	0.685	0.642	0,663

Tabel 4.32 menunjukkan bahwa data *testing* dengan variabel prediktor sebanyak 10% memiliki nilai akurasi total

sebesar 0,667, *sensitivity* sebesar 0,685, *specificity* sebesar 0,642, dan AUC sebesar 0,663.

3. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 15% Prediktor Terpenting

Data yang digunakan adalah sejumlah 33 variabel prediktor yang dibagi menjadi data *training* dan data *testing*. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.1196581. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.13 Hasil *Tuning Complexity Parameter* dengan prediktor 15%

Tabel 4.33 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 15% variabel prediktor terpenting.

Tabel 4.33 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 15% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	5	0
1	3	0

Berdasarkan Tabel 4.33 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 5 senyawa dan tidak terdapat senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 0 senyawa dan terdapat 3 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{5+0}{5+0+3+0} = 0{,}625$$

$$Sensitivity = \frac{5}{5+0} = 1$$

$$Specifity = \frac{0}{3+0} = 0$$

$$Specifity = \frac{1}{2}(1+0) = 0{,}5$$

Pada Tabel 4.34 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 15% variabel terpenting.

Tabel 4.34 Performa CART dengan 15% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.720	0.575	0.886	0.730
	Testing	0.444	0.200	0.750	0.475
2	*Training*	0.787	0.775	0.800	0.788
	Testing	0.333	0.600	0.000	0.300
3	*Training*	0.787	0.700	0.886	0.793
	Testing	0.444	0.600	0.250	0.425
4	*Training*	0.613	0.325	0.943	0.634
	Testing	0.333	0.000	0.750	0.375
5	*Training*	0.776	0.756	0.800	0.778
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.658	0.707	0.600	0.654
	Testing	0.500	0.750	0.250	0.500
7	*Training*	0.789	0.902	0.657	0.780
	Testing	0.375	0.750	0.000	0.375
8	*Training*	0.763	0.878	0.629	0.753
	Testing	0.625	1.000	0.250	0.625

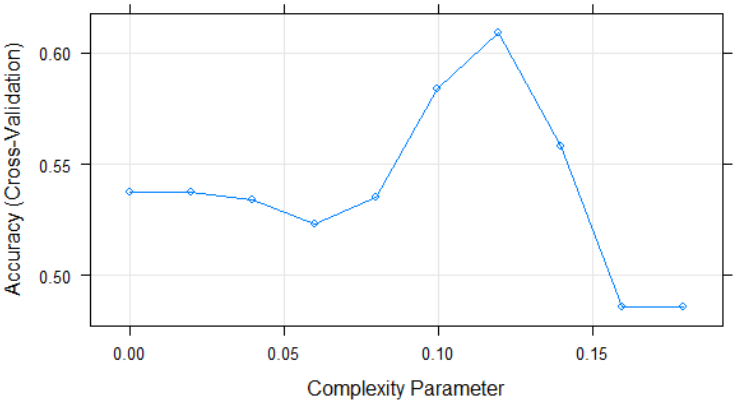
Tabel 4.34 Performa CART dengan 15% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
9	*Training*	0.829	0.927	0.714	0.821
	Testing	0.500	0.750	0.250	0.500
10	*Training*	0.618	1.000	0.194	0.597
	Testing	0.625	1.000	0.000	0.500
Rata-Rata	*Training*	0.734	0.755	0.711	0,733
	Testing	0.468	0.590	0.325	0,458

Tabel 4.34 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 15% memiliki nilai akurasi total sebesar 0,468, *sensitivity* sebesar 0,590, *specificity* sebesar 0,325, dan AUC sebesar 0,458.

4. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 20% Prediktor Terpenting

Variabel prediktor yang digunakan adalah sebanyak 44 variabel atau 20% dari 217 variabel. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.1196581. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.14 Hasil *Tuning Complexity Parameter* dengan prediktor 20%

Tabel 4.35 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 20% variabel prediktor terpenting.

Tabel 4.35 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 20% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	5	0
1	1	2

Berdasarkan Tabel 4.35 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 5 senyawa dan tidak terdapat senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{5+2}{5+0+1+2} = 0{,}825$$

$$Sensitivity = \frac{5}{5+0} = 1$$

$Specifity = \frac{2}{1+2} = 0{,}667$

$AUC = \frac{1}{2}(1 + 0{,}667) = 0{,}833$

Pada Tabel 4.36 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 20% variabel terpenting.

Tabel 4.36 Performa CART dengan 20% Prediktor Terpenting

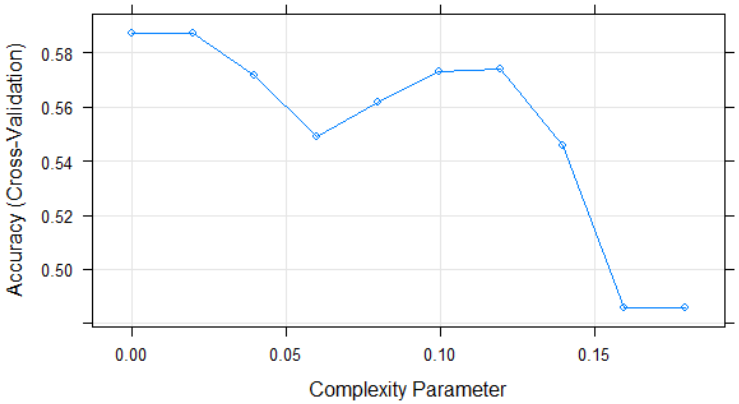
Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.720	0.575	0.886	0.730
	Testing	0.444	0.200	0.750	0.475
2	*Training*	0.787	0.775	0.800	0.788
	Testing	0.333	0.600	0.000	0.300
3	*Training*	0.787	0.700	0.886	0.793
	Testing	0.444	0.600	0.250	0.425
4	*Training*	0.733	0.550	0.943	0.746
	Testing	0.444	0.200	0.750	0.475
5	*Training*	0.776	0.756	0.800	0.778
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.658	0.707	0.600	0.654
	Testing	0.500	0.750	0.250	0.500
7	*Training*	0.789	0.902	0.657	0.780
	Testing	0.375	0.750	0.000	0.375
8	*Training*	0.763	0.878	0.629	0.753
	Testing	0.625	1.000	0.250	0.625
9	*Training*	0.816	0.829	0.800	0.815
	Testing	0.250	0.500	0.000	0.250
10	*Training*	0.868	0.900	0.833	0.867
	Testing	0.875	1.000	0.667	0.833
Rata-Rata	*Training*	0.770	0.757	0.783	0,770
	Testing	0.479	0.585	0.367	0,476

Tabel 4.36 menunjukkan hasil total akurasi, *sensitivity*, *specificity* dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 20% memiliki nilai akurasi total

sebesar 0,479, *sensitivity* sebesar 0,585, *specificity* sebesar 0,367 dan AUC sebesar 0,476.

5. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 25% Prediktor Terpenting

Persentase variabel prediktor yang digunakan adalah 25% atau sejumlah 55 variabel. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.0199430. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.15 Hasil *Tuning Complexity Parameter* dengan prediktor 25%

Berikut merupakan *confusion matrix* dari data *testing fold* 10 klasifikasi senyawa obat kanker.

Tabel 4.37 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 25% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	4	1
1	1	2

Berdasarkan Tabel 4.37 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan

ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 2 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 dengan variabel prediktor 25% sama dengan menggunakan variabel prediktor 10%. Pada Tabel 4.38 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 25% variabel terpenting.

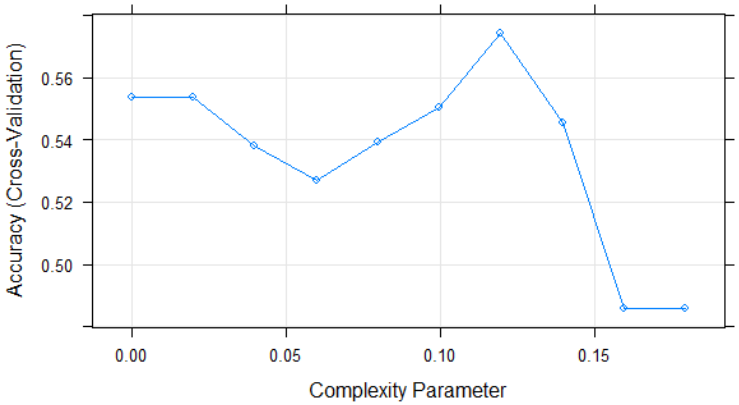
Tabel 4.38 Performa CART dengan 25% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.827	0.975	0.657	0.816
	Testing	0.667	0.800	0.500	0.650
2	*Training*	0.813	0.925	0.686	0.805
	Testing	0.556	1.000	0.000	0.500
3	*Training*	0.813	0.800	0.829	0.814
	Testing	0.556	0.800	0.250	0.525
4	*Training*	0.787	0.700	0.886	0.793
	Testing	0.444	0.400	0.500	0.450
5	*Training*	0.855	0.976	0.714	0.845
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.789	0.902	0.657	0.780
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.921	0.878	0.971	0.925
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.803	0.707	0.914	0.811
	Testing	0.750	0.750	0.750	0.750
9	*Training*	0.842	0.732	0.971	0.852
	Testing	0.375	0.000	0.750	0.375
10	*Training*	0.855	0.875	0.833	0.854
	Testing	0.750	0.800	0.667	0.733
Rata-Rata	*Training*	0.831	0.847	0.812	0,829
	Testing	0.560	0.605	0.492	0,548

Tabel 4.38 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 25% memiliki nilai akurasi total sebesar 0,560, *sensitivity* sebesar 0,605, *specificity* sebesar 0,492, dan AUC sebesar 0,548.

6. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 30% Prediktor Terpenting

Variabel prediktor yang digunakan adalah sebanyak 66 variabel atau 30% dari 217 variabel. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.1196581. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.16 Hasil *Tuning Complexity Parameter* dengan prediktor 30%

Tabel 4.39 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 30% variabel prediktor terpenting.

Tabel 4.39 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 30% Prediktor Terpenting

Aktual	Prediksi	
	0	1
0	4	1
1	2	1

Berdasarkan Tabel 4.39 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 1 senyawa dan terdapat 2 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 dengan variabel prediktor 35% sama dengan menggunakan variabel prediktor 10% dan 25%. Pada Tabel 4.40 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 35% variabel terpenting.

Tabel 4.40 Performa CART dengan 30% Prediktor Terpenting

Fold	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
1	*Training*	0.720	0.575	0.886	0.730
	Testing	0.444	0.200	0.750	0.475
2	*Training*	0.787	0.775	0.800	0.788
	Testing	0.333	0.600	0.000	0.300
3	*Training*	0.680	0.650	0.714	0.682
	Testing	0.222	0.400	0.000	0.200
4	*Training*	0.733	0.550	0.943	0.746
	Testing	0.444	0.200	0.750	0.475
5	*Training*	0.776	0.756	0.800	0.778
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.658	0.707	0.600	0.654
	Testing	0.500	0.750	0.250	0.500
7	*Training*	0.921	0.878	0.971	0.925
	Testing	0.500	0.750	0.250	0.500

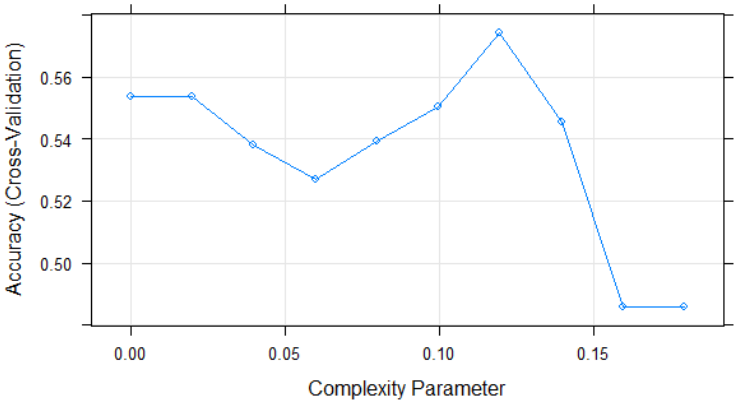
Tabel 4.40 Performa CART dengan 30% Prediktor Terpenting (Lanjutan)

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
8	*Training*	0.763	0.878	0.629	0.753
	Testing	0.625	1.000	0.250	0.625
9	*Training*	0.829	0.805	0.857	0.831
	Testing	0.250	0.000	0.500	0.250
10	*Training*	0.750	0.950	0.528	0.739
	Testing	0.625	0.800	0.333	0.567
Rata-Rata	*Training*	0.762	0.752	0.773	0,763
	Testing	0.444	0.495	0.383	0,439

Tabel 4.40 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 30% memiliki nilai akurasi total sebesar 0,444, *sensitivity* sebesar 0,495, *specificity* sebesar 0,383, dan AUC sebesar 0,439.

7. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 35% Prediktor Terpenting

Data yang digunakan adalah sejumlah 76 variabel prediktor yang dibagi menjadi data *training* dan data *testing*. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.01994302. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.17 Hasil *Tuning Complexity Parameter* dengan prediktor 35%

Tabel 4.41 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 35% variabel prediktor terpenting.

Tabel 4.41 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 35% Prediktor Terpenting

Aktual	**Prediksi**	
	0	1
0	4	1
1	1	2

Berdasarkan Tabel 4.41 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 2 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+2}{4+1+1+2} = 0{,}75$$

$$Sensitivity = \frac{4}{4+1} = 0{,}8$$

$Specifity = \frac{2}{1+2} = 0{,}667$

$AUC = \frac{1}{2}(0{,}8 + 0{,}667) = 0{,}733$

Pada Tabel 4.42 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 35% variabel terpenting.

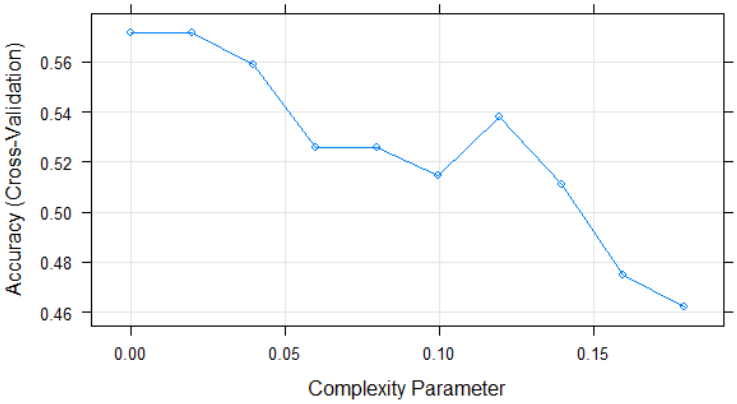
Tabel 4.42 Performa CART dengan 35% Prediktor Terpenting

Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
1	*Training*	0.773	0.775	0.771	0.773
	Testing	0.444	0.200	0.750	0.475
2	*Training*	0.813	0.925	0.686	0.805
	Testing	0.556	1.000	0.000	0.500
3	*Training*	0.840	0.875	0.800	0.838
	Testing	0.444	0.800	0.000	0.400
4	*Training*	0.800	0.700	0.914	0.807
	Testing	0.667	0.600	0.750	0.675
5	*Training*	0.816	0.951	0.657	0.804
	Testing	0.500	0.250	0.750	0.500
6	*Training*	0.789	0.902	0.657	0.780
	Testing	0.500	0.500	0.500	0.500
7	*Training*	0.921	0.878	0.971	0.925
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.803	0.707	0.914	0.811
	Testing	0.750	0.750	0.750	0.750
9	*Training*	0.842	0.732	0.971	0.852
	Testing	0.375	0.000	0.750	0.375
10	*Training*	0.855	0.875	0.833	0.854
	Testing	0.750	0.800	0.667	0.733
Rata-Rata	*Training*	0.825	0.832	0.818	0,825
	Testing	0.549	0.565	0.517	0,541

Tabel 4.42 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 35% memiliki nilai akurasi total sebesar 0,549, *sensitivity* sebesar 0,565, *specificity* sebesar 0,517, dan AUC sebesar 0,541.

8. Klasifikasi Senyawa Obat Kanker dalam Dua Kelas Proteksi Radiasi Menggunakan CART dengan 100% Prediktor

Data yang digunakan adalah semua variabel prediktor atau 217 variabel yang kemudian dibagi menjadi data *training* dan data *testing*. Hasil dari *parameter tuning* yaitu nilai CP yang optimal sebesar 0.01994302. Berikut merupakan grafik dari *tuning* CP.



Gambar 4.18 Hasil *Tuning Complexity Parameter* dengan prediktor 100%

Tabel 4.43 menunjukkan *confusing matrix* data *testing fold* 10 dengan menggunakan 100% variabel prediktor.

Tabel 4.43 *Confusing Matrix* Data *Testing Fold* 10 Menggunakan CART dengan 100% Prediktor

Aktual	Prediksi	
	0	1
0	4	1
1	0	3

Berdasarkan Tabel 4.43 diketahui bahwa pada data *testing fold* 10, senyawa obat kanker kelas 0 yang tepat diklasifikasikan ke kelas 0 ada sebanyak 4 senyawa dan terdapat 1 senyawa yang salah diklasifikasikan ke kelas 1, sedangkan senyawa kelas 1 yang tepat diklasifikasikan ke kelas 1 ada sebanyak 3 senyawa dan tidak terdapat senyawa yang salah diklasifikasikan ke kelas 0. Penghitungan *total accuracy*, *sensitivity*, *specificity*, dan AUC untuk data *testing fold* 10 adalah sebagai berikut.

$$Total\ Accuracy\ Rate = \frac{4+3}{4+1+0+3} = 0{,}875$$

$$Sensitivity = \frac{4}{4+1} = 0{,}8$$

$$Specifity = \frac{3}{0+3} = 1$$

$$AUC = \frac{1}{2}(0{,}8 + 1) = 0{,}9$$

Pada Tabel 4.44 ditunjukkan performa CART untuk klasifikasi senyawa obat kanker dengan 100% variabel.

Tabel 4.44 Performa CART dengan 100% Prediktor

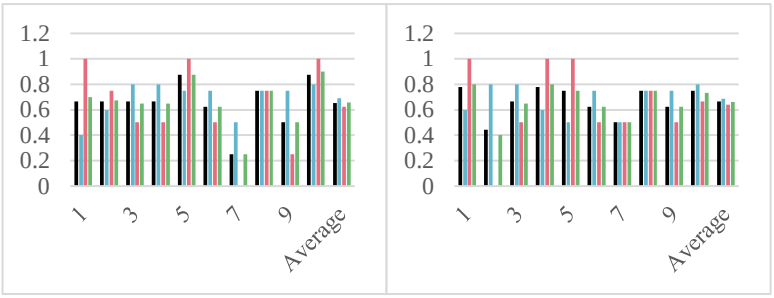
Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
1	*Training*	0.773	0.775	0.771	0.773
	Testing	0.444	0.200	0.750	0.475
2	*Training*	0.867	0.975	0.743	0.859
	Testing	0.444	0.800	0.000	0.400
3	*Training*	0.787	0.700	0.886	0.793
	Testing	0.222	0.400	0.000	0.200

Tabel 4.44 Performa CART dengan 100% Prediktor (Lanjutan)

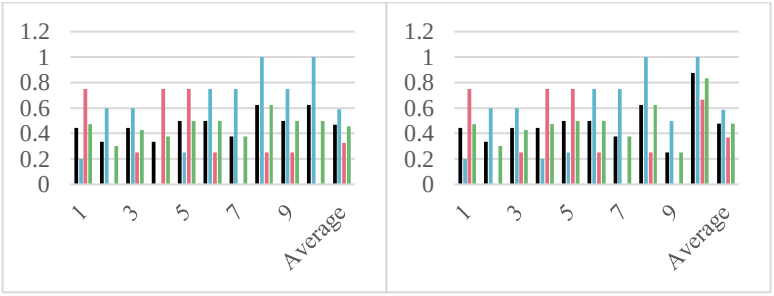
Fold	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
4	*Training*	0.800	0.700	0.914	0.807
	Testing	0.444	0.200	0.750	0.475
5	*Training*	0.829	0.805	0.857	0.831
	Testing	0.500	0.500	0.500	0.500
6	*Training*	0.816	0.902	0.714	0.808
	Testing	0.500	0.750	0.250	0.500
7	*Training*	0.921	0.878	0.971	0.925
	Testing	0.500	0.750	0.250	0.500
8	*Training*	0.803	0.756	0.857	0.807
	Testing	0.875	0.750	1.000	0.875
9	*Training*	0.842	0.878	0.800	0.839
	Testing	0.500	0.500	0.500	0.500
10	*Training*	0.855	0.825	0.889	0.857
	Testing	0.875	0.800	1.000	0.900
Rata-Rata	*Training*	0.829	0.819	0.840	0,830
	Testing	0.531	0.565	0.500	0,533

Tabel 4.44 menunjukkan hasil total akurasi, *sensitivity*, *specificity*, dan AUC dari seluruh *fold*. Data *testing* dengan variabel prediktor sebanyak 100% memiliki nilai akurasi total sebesar 0,531, *sensitivity* sebesar 0,565, *specificity* sebesar 0,500, dan AUC sebesar 0,533.

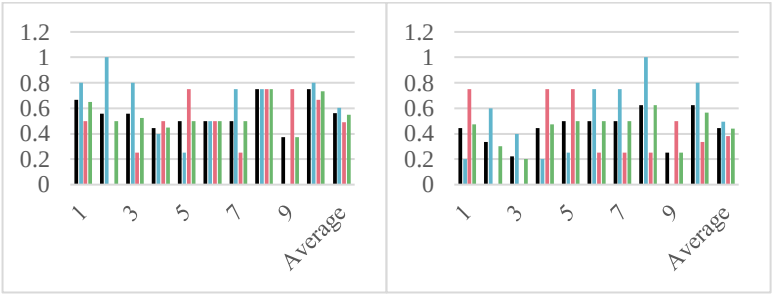
Nilai akurasi total, *sensitivity*, *specificity*, dan AUC dari setiap fold untuk jumlah variabel prediktor yang digunakan sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, 100% disajikan dalam Gambar 4.19 sebagai berikut.


1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

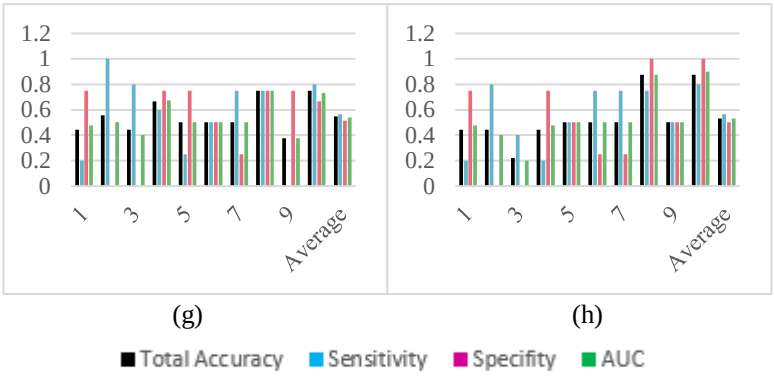
(a) (b)


1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

(c) (d)


1.2
1
0.8
0.6
0.4
0.2
0
1
3
5
7
9
Average

(e) (f)



Gambar 4.19 Performa CART per Jumlah Prediktor
(a) 5% Prediktor (b) 10% Prediktor (c) 15% Prediktor (d) 20% Prediktor (e) 25% Prediktor (f) 30% Prediktor (g) 35% Prediktor (h) 100% Prediktor

Validasi model menggunakan 10-*fold cross validation*, yang artinya nilai akurasi total, *sensitivity*, *specificity*, dan AUC setiap *fold* dirata-ratakan untuk setiap jumlah variabel prediktor yang digunakan. Pada Tabel 4.45 disajikan performa metode CART pada setiap jumlah variabel prediktor yang digunakan.

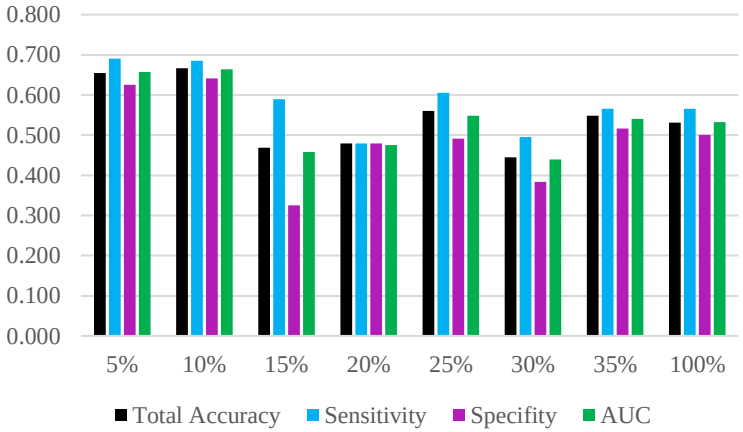
Tabel 4.45 Performa CART

Persentase Prediktor	**Data**	***Total Accuracy Rate***	***Sensitivity***	***Specificity***	**AUC**
5%	*Training*	0.799	0.844	0.747	0,795
	Testing	0.654	0.690	0.625	0.658
10%	*Training*	0.824	0.813	0.838	0,825
	Testing	0.667	0.685	0.642	0.663
15%	*Training*	0.734	0.755	0.711	0,733
	Testing	0.468	0.590	0.325	0.458
20%	*Training*	0.770	0.757	0.783	0,770
	Testing	0.479	0.479	0.479	0.476
25%	*Training*	0.831	0.847	0.812	0,829
	Testing	0.560	0.605	0.492	0.548

Tabel 4.45 Performa CART (Lanjutan)

Persentase Prediktor	Data	*Total Accuracy Rate*	*Sensitivity*	*Specificity*	AUC
30%	*Training*	0.762	0.752	0.773	0,763
	Testing	0.444	0.495	0.383	0.439
35%	*Training*	0.825	0.832	0.818	0,825
	Testing	0.549	0.565	0.517	0.541
100%	*Training*	0.829	0.819	0.840	0,830
	Testing	0.531	0.565	0.500	0.533

Performa metode CART pada Tabel 4.45 digambarkan pada Gambar 4.20 sebagai berikut.



Gambar 4.20 Performa CART

Berdasarkan Tabel 4.45 dan Gambar 4.20, diketahui bahwa prediktor yang optimal untuk digunakan pada metode CART adalah sebanyak 5% prediktor. Hasil total akurasi dengan 5% prediktor yaitu yaitu 0,654, *sensitivity* sebesar 0,690, *specificity* sebesar 0,625, dan AUC sebesar 0,658.

4.3.3 Perbandingan Nilai Ketepatan Klasifikasi dalam Dua Kelas Proteksi Radiasi

Tabel 4.46 menunjukkan perbandingan antara klasifikasi menggunakan *naïve bayes classifier* dan CART per jumlah variabel prediktor yang digunakan pada dua kelas proteksi radiasi. Metode evaluasi ketepatan klasifikasi yang digunakan adalah *total accuracy rate, sensitivity, specificity,* dan AUC.

Tabel 4.46 Perbandingan Nilai Ketepatan Klasifikasi dalam Dua Kelas Proteksi Radiasi

% Pred.	*Total Accuracy Rate*		*Sensitivity*		*Specificity*		AUC	
	NBC	CART	NBC	CART	NBC	CART	NBC	CART
5%	0.572	0.654	0.790	0.690	0.308	0.625	0.549	0.658
10%	0.553	0.667	0.690	0.685	0.383	0.642	0.537	0.663
15%	0.489	0.468	0.570	0.590	0.383	0.325	0.477	0.458
20%	0.514	0.479	0.514	0.479	0.514	0.479	0.499	0.476
25%	0.525	0.560	0.640	0.605	0.383	0.492	0.512	0.548
30%	0.536	0.444	0.660	0.495	0.383	0.383	0.522	0.439
35%	0.560	0.549	0.660	0.565	0.442	0.517	0.551	0.541
100%	0.489	0.531	0.735	0.565	0.208	0.500	0.472	0.533

Tabel 4.46 menunjukkan bahwa nilai total akurasi tertinggi yaitu 0,667 dan AUC tertinggi yaitu 0,663 yang didapatkan menggunakan metode CART dengan 10% prediktor. Selanjutnya dilakukan perbandingan dengan penelitian yang dilakukan oleh Matsumoto, *et al.* (2016) yang membandingkan metode *random forest* (RF), *support vector machine* (SVM), *extreme gradient boosting* (XGB), dan K-*nearest neighbor* (KNN). Berikut merupakan perbandingan nilai AUC yang didapatkan dari penelitian ini dan penelitian yang dilakukan oleh Matsumoto, *et al.* (2016)

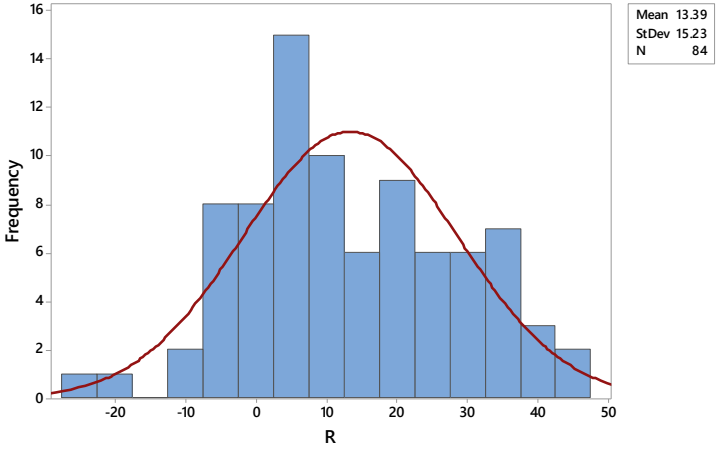
Tabel 4.47 Perbandingan Nilai AUC dengan Penelitian Sebelumnya

Persentase Prediktor	Metode	AUC
100%	NBC	0,472
	CART	0,533
	KNN	0,584
	XGB	0,592
	RF	0,528
	SVM	0,453

.Sesuai dengan hasil nilai AUC yang dituliskan pada Tabel 4.47, kedua metode tersebut memberikan hasil bahwa metode NBC lebih baik daripada metode SVM namun tidak lebih baik dari metode CART, KNN, XGB, dan RF. Sedangkan metode CART lebih baik jika dibandingkan NBC, RF, dan SVM, namun tidak lebih baik jika dibandingkan dengan metode KNN dan XGB.

4.4 Pembentukan Kelas Baru untuk Optimalisasi Klasifikasi Senyawa Obat Kanker menggunakan Pendekatan *Normal Mixture Distribution*

Hasil ketepatan klasifikasi dengan metode NBC dan CART yang relatif kecil menyebabkan munculnya hipotesis bahwa dengan melakukan *data driven exploration* akan menghasilkan nilai ketepatan klasifikasi yang lebih tinggi. Eksplorasi tersebut yaitu dengan membentuk kelas baru. Metode pembentukan kelas baru yang digunakan adalah menggunakan pendekatan *normal mixture distribution* karena data berdistribusi normal. Jika diketahui suatu data, tidak selamanya satu distribusi saja dapat merepresentasikan data tersebut. Namun, apabila ada indikasi beberapa komposisi muncul dari data tersebut, maka tidak menutup kemungkinan bahwa distribusi data yang lebih tepat adalah distribusi *mixture* (Dempster, *et al*., 1977). Berikut merupakan histogram dari tingkat kematian sel kanker.



Gambar 4.21 Histogram Tingkat Kematian Sel Kanker

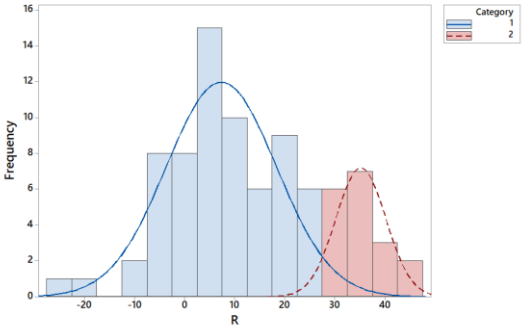
Berdasarkan histogram pada Gambar 4.21, dapat diketahui bahwa secara visual data tingkat kematian sel kanker memiliki lebih dari satu puncak, sehingga diduga data tersebut memiliki lebih dari satu distribusi. Oleh karena itu, pendekatan *normal mixture distribution* tepat untuk digunakan karena dengan menggunakan *normal mixture distribution* dapat menyelesaikan permasalahan data dengan lebih dari satu distribusi.

Nilai *K* atau jumlah komponen / kategori pada penelitian ini telah ditentukan, yaitu $K = 2, 3,$ dan 4. Berdasarkan nilai *K* tersebut, selanjutnya digunakan nilai *log-likelihood* yang diestimasi menggunakan algoritma *expectation-maximization* untuk menentukan model terbaik, dimana *K* dengan nilai *log-likelihood* terbesar adalah jumlah distribusi yang paling optimal. Berikut merupakan hasil perhitungan nilai *log-likelihood* dari setiap *K*.

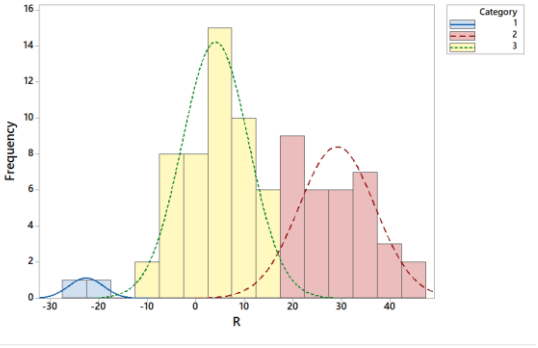
Tabel 4.48 Parameter pada Setiap Nilai *K*

K	Parameter	Komp.1	Komp.2	Komp.3	Komp.4	Log-Likelihood
2	Phi (π) Mu (μ) Sigma (σ)	0.768 7.379 11.382	0.232 33.266 6.655			-344.2608
3	Phi (π) Mu (μ) Sigma (σ)	0.023 -22.55 2.499	0.435 27.254 9.407	0.542 3.792 7.065		-341.8135
4	Phi (π) Mu (μ) Sigma (σ)	0.110 -5.87 2.310	0.359 5.015 4.348	0.507 25.188 10.376	0.024 -22.5 2.500	-340.4461

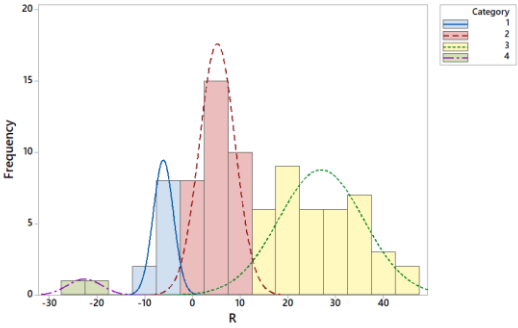
Berdasarkan nilai *log-likelihood* yang ditunjukkan pada Tabel 4.48, model terbaik adalah dengan menggunakan 4 komponen yaitu dengan nilai *log-likelihood* sebesar -340,446. Namun pada penelitian ini akan dilakukan analisis menggunakan *K* = 2, 3, dan 4, karena nilai *log-likelihood* antara ketiga nilai *K* tersebut tidak jauh berbeda. Berikut merupakan *distribution* plot pada data menggunakan *K* = 2, 3, dan 4.



(a)



(b)



(c)

Gambar 4.22 *Distribution Plot*

(a) Dua Komponen (b) Tiga Komponen (c) Empat Komponen

Dalam menentukan keanggotaan masing-masing kategori, maka dilihat nilai *posterior probability*. Suatu observasi dikategorikan dalam kategori yang memiliki nilai *posterior probability* tertinggi. Berikut merupakan nilai *posterior probability* dari setiap observasi dengan K = 2, 3, dan 4.

Tabel 4.49 *Posterior Probability*

K	Senyawa	Komp.1	Komp.2	Komp.3	Komp.3	Kategori
2	AS-1	0.232	0.768			2
	AS-10	0.558	0.442			1
	AS-11	1.000	0.000			1
	AS-12	0.558	0.442			1
	…	…	…			…
	YT-1	0.095	0.905			2
3	AS-1	2E-97	1E+00	2E-03		2
	AS-10	5E-80	1E+00	2E-02		2
	AS-11	3E-19	1E-02	1E+00		3
	AS-12	5E-80	1E+00	2E-02		2
	…	…	…	…		…
	YT-1	2E-116	1E+00	1E-04		2
4	AS-1	4E-53	1E-07	1E+00	4E-97	3
	AS-10	2E-39	4E-05	1E+00	8E-80	3
	AS-11	4E-02	9E-01	5E-02	5E-19	2
	AS-12	2E-39	4E-05	1E+00	8E-80	3
	…	…	…	…	…	…
	YT-1	1E-68	1E-10	1.e-4	2E-117	2

Tabel 4.49 menunjukan nilai *posterior probability* dari setiap kategori. Dengan menggunakan nilai tersebut, maka observasi atau senyawa bisa dikelompokkan berdasarkan nilai *posterior probability* terbesar. Berikut merupakan jumlah senyawa pada setiap kategori.

Tabel 4.50 Jumlah Anggota Setiap Kategori

K	Kategori 1	Kategori 2	Kategori 3	Kategori 4
2	66	18		
3	2	33	49	
4	10	45	27	2

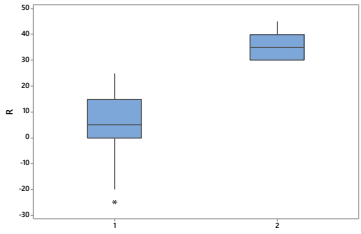
Setelah mendapatkan anggota dari setiap kategori, perlu dilakukan analisis secara deskriptif untuk mengetahui karakter dari setiap kategori, yang kemudian digunakan untuk

mendefinisikan masing-masing kategori. Berikut merupakan perhitungan analisis statistik deskriptif dari setiap kategori.

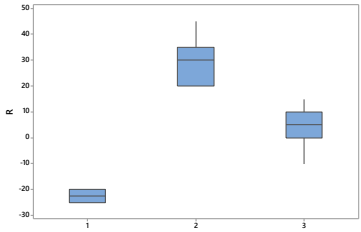
Tabel 4.51 Perhitungan Statistika Deskriptif Setiap Kategori

K	**Kategori**	***Mean***	**Varians**	***Max.***	***Min.***	***Threshold***
2	1	7.42	120.96	-25.00	25.00	$K1 \leq 25$
	2	35.28	24.92	30.00	45.00	$K2 > 25$
3	1	-22.5	12.50	-25.00	-20.00	$K1 \leq -15$
	2	29.24	61.13	20.00	45.00	$K2 > 15$
	3	4.18	47.24	-10.00	15.00	$-15 < K3 \leq 15$
4	1	-6.00	4.44	-10.00	-5.00	$-15 < K1 \leq -5$
	2	5.30	13.97	0.00	10.00	$5 < K2 \leq 15$
	3	27.05	78.58	15.00	45.00	$K3 > 15$
	4	-22.50	12.50	-25.00	-20.00	$K4 \leq -15$

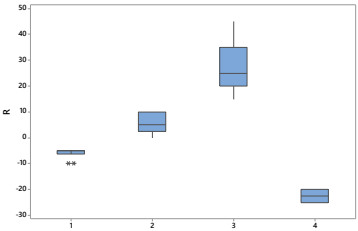
Berikut merupakan *boxplot* dari tingkat kematian sel kanker setiap kategori pada $K = 2, 3$, dan 4



(a)



(b)



(c)

Gambar 4.23 *Boxplot* Tingkat Kematian Sel Kanker Setiap Kategori
(a) K=2 (b) K=3 (c) K=4

Secara visual dan secara perhitungan statistik deskriptif, menunjukkan terdapat perbedaan pada masing-masing kategori. Dengan menggunakan K = 2, kategori 1 memiliki rata-rata kematian sel sebesar 7,42% yang didapatkan dari senyawa yang memiliki tingkat kematian sel kanker di bawah 30%. Sedangkan kategori 2 memiliki nilai rata-rata sebesar 35.28% yang didapatkan dari senyawa yang memiliki tingkat kematian sel kanker minimal 30%.

Pada K = 3, kategori 1 memiliki rata-rata -22.5% yang artinya sel kanker mengalami perkembangan dengan rata-rata 22,5%. Nilai tersebut didapatkan dari senyawa dengan tingkat kematian sel -25% hingga -20%. Pada kategori 3 memiliki rata-rata tingkat kematian sel kanker 4,18% dengan minimal kematian sel kanker sebesar -10% dan maksimal sebesar 15%. Sedangkan pada kategori 2, tingkat kematian sel memiliki rata-rata 29,24% yang didapatkan dari senyawa yang berhasil mematikan sel kanker minimal 20%.

Pengkategorian dengan menggunakan 4 kategori menunjukkan bahwa kategori 4 memiliki rata-rata kematian sel kanker sebesar -22,5% dengan minimal -25% hingga -20%. Kategori 1 mengelompokkan senyawa dengan rata-rata perkembangan sel sebesar 6% dengan minimal tingkat kematian sel kanker -10% hingga -5%. Sedangkan senyawa-senyawa yang tergabung dalam kategori 2 memiliki rata-rata kematian sel

kanker sebesar 5,3% dengan tingkat kematian sel kanker sebesar 0-10%. Kategori paling baik adalah kategori 3 dengan rata-rata tingkat kematian sel kanker sebesar 27,05% yang didapatkan dari senyawa yang berhasil mematikan sel kanker minimal 15%.

Pada 3 dan 4 kategori, terdapat kategori yang bisa mencakup senyawa-senyawa yang memiliki tingkat kematian sel kanker yang negatif. Artinya, dengan penggunaan senyawa tersebut pada radioterapi, justru akan membahayakan pasien karena berakibat menumbuhkan sel-sel kanker, bukan mematikannya. Senyawa-senyawa yang memiliki efek kematian sel kanker negatif adalah sebagai berikut.

Tabel 4.52 Senyawa dengan Tingkat Kematian Sel Kanker Negatif

Tingkat Kematian Sel Kanker	Senyawa
-25%	ST-1
-20%	AS-3
-10%	KH-23, MH-9
-5%	AS-4, KT-1, MH-14, SAr-5, UM-8, YM-14, YN-4, YN-6

Jika dilakukan *ranking* berdasarkan tingkat kematian sel kanker, didapatkan hasil bahwa pada 2 kategori, kategori 2 lebih bagus daripada kategori 1. Pada 3 kategori, kategori yang bagus secara berurutan yaitu kategori 2, 3, dan 1. Sedangkan pada 4 kategori, secara berurutan kategori yang lebih baik adalah 3, 2, 1, dan 4. Berikut merupakan senyawa-senyawa yang tergolong pada setiap kategori dengan K = 2, 3, dan 4.

Tabel 4.53 *Ranking* dan Aggota Setiap Kategori

K	Rank	Kategori	Senyawa Obat Kanker
2	1	2	AS-1, AS-17, AS-2, KH-18, KH-20, KH-22, KH-3, KT-2, MH-1, MH-15, MY-2, SAr-7, UM-10, Vitamine E, YM-13, YN-7, YN-9, YT-1,

Tabel 4.48 *Ranking* dan Anggota Setiap Kategori (Lanjutan)

K	Rank	Kategori	Senyawa Obat Kanker
2	2	1	AS-10, AS-11, AS-12, AS-13, AS-15, AS-16, AS-3, AS-4, AS-5, AS-6, AS-7, AS-8, AS-9, KH-1, KH-10, KH-12, KH-13, KH-16, KH-19, KH-2, KH-21, KH-23, KH-24, KH-25, KH-4, KH-5, KH-6, KT-1, MH-10, MH-11, MH-12, MH-13, MH-14, MH-16, MH-2, MH-3, MH-4, MH-5, MH-6, MH-7, MH-8, MH-9, MY-1, naphthalene, naphthol, phenol, quinoform, quinoline, SAr-1, SAr-2, SAr-3, SAr-4, SAr-5, ST-1,TPEN, UM-7, UM-8, UM-9, Vitamine C, YM-14, YM-16, YN-1, YN-4, YN-5, YN-6, YN-8
3	1	2	AS-1, AS-10, AS-12, AS-15, AS-16, AS-1 7, AS-2, AS-5, AS-6, AS-8, KH-13, KH-18, KH-2, KH-20, KH-22, KH-24, KH-25, KH-3, KT-2, MH-1, MH-15, MY-1, MY-2, naphthol, SAr-7, UM-10, UM-7, UM-9, Vitamine E, YM-13, YN-7, YN-9, YT-1
	2	3	AS-11, AS-13, AS-4, AS-7, AS-9, KH-1, KH-10, KH-12, KH-16, KH-19, KH-21, KH-23, KH-4, KH-5, KH-6, KT-1, MH-10, MH-11, MH-12, MH-13, MH-14, MH-16, MH-2, MH-3, MH-4, MH-5, MH-6, MH-7, MH-8, MH-9, naphthalene, phenol, quinoform, quinoline, SAr-1, SAr-2, SAr-3, SAr-4, SAr-5, TPEN, UM-8, Vitamine C, YM-14, YM-16, YN-1, YN-4, YN-5, YN-6, YN-8,
	3	1	AS-3, ST-1

Tabel 4.48 *Ranking* dan Anggota Setiap Kategori (Lanjutan)

K	Rank	Kategori	Senyawa Obat Kanker
4	1	3	AS-1, AS-10, AS-12, AS-15, AS-16, AS-5, AS-6, AS-7, AS-8, KH-1, KH-13, KH-18, KH-2, KH-24, KH-25, KH-3, KH-4, MH-13, MH-15, MH-16, MY-1, MY-2, naphtol, SAr-2, SAr-7, UM-7, UM-9
	2	2	AS-11, AS-13, AS-17, AS-2, AS-9, KH-10, KH-12, KH-16, KH-19, KH-20, KH-21, KH-22, KH-5, KH-6, KT-2, MH-1, MH-10, MH-11, MH-12, MH-2, MH-3, MH-4, MH-5, MH-6, MH-7, MH-8, naphthalene, phenol, quinoform, quinoline, SAr-1, SAr-3, SAr-4, TPEN, UM-10, Vitamine_C, Vitamine_E, YM-13, YM-16, YN-1, YN-5, YN-7, YN-8, YN-9, YT-1
	3	1	AS-4, KH-23, KT-1, MH-14, MH-9, SAr-5, UM-8, YM-14, YN-4, YN-6
	4	4	AS-3, ST-1

Hasil dari pengkategorian observasi tersebut, selanjutnya digunakan untuk variabel respon pada klasifikasi menggunakan *naïve bayes* dan *classification and regression tree* (CART). Pada 2 dan 3 kategori, terdapat salah satu kategori yang hanya memiliki 2 anggota, sehingga tidak bisa dilakukan 10-*fold cross validation*. Oleh karena itu, dilakukan pembagian data *training* dan data *testing* secara random masing-masing sebesar 70% dan 30%.

4.5 Klasifikasi Senyawa Obat Kanker dengan *Naïve Bayes Classifier* dan CART Menggunakan Kelas dari Hasil Pengelompokan dengan Pendekatan *Normal Mixture Distribution*

Pada bagian ini diljelaskan klasifikasi pada kelas baru hasil pengelompokan menggunakan *normal mixture distribution*.

Metode yang digunakan adalah metode *naïve bayes classifier* dan CART, sedangkan untuk mengevaluasi ketepatan klasifikasi menggunakan AUC dikarenakan jumlah senyawa pada masing-masing kategori tidak seimbang atau *unbalance*.

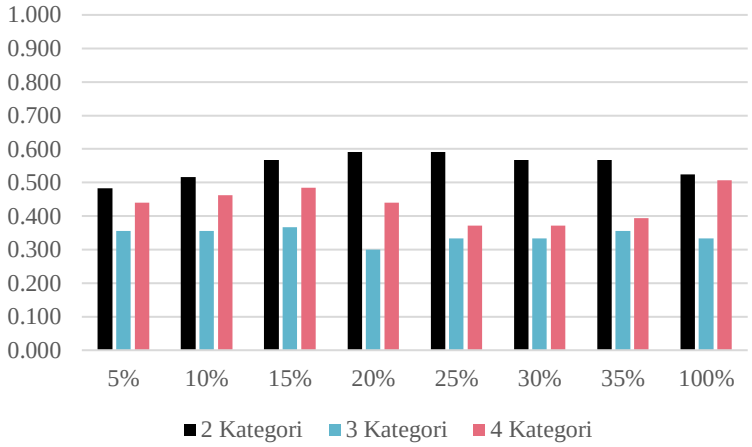
4.5.1 Klasifikasi Senyawa Obat Kanker dengan *Naïve Bayes Classifier* Menggunakan Kelas dari Hasil Pengelompokan dengan Pendekatan *Normal Mixture Distribution*

Metode yang digunakan yaitu *naïve bayes classifier* dengan pembagian data *training* dan data *testing* sebesar 70% dan 30%. Jumlah variabel prediktor yang digunakan adalah sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100%. Berikut merupakan performa *naïve bayes classifier* menggunakan variabel respon hasil pendekatan *mixture distribution*.

Tabel 4.54 Performa *Naive Bayes Classifier* dengan Menggunakan Kelas Baru

Persentase Prediktor	Data	2 Kategori	3 Kategori	4 Kategori
5%	*Training*	0.642	0.536	0.381
	Testing	0.483	0.356	0.439
10%	*Training*	0.643	0.458	0.381
	Testing	0.517	0.356	0.462
15%	*Training*	0.601	0.429	0.378
	Testing	0.567	0.367	0.485
20%	*Training*	0.612	0.487	0.369
	Testing	0.592	0.300	0.439
25%	*Training*	0.591	0.458	0.394
	Testing	0.592	0.333	0.371
30%	*Training*	0.591	0.444	0.386
	Testing	0.567	0.333	0.371
35%	*Training*	0.696	0.444	0.386
	Testing	0.567	0.356	0.394
100%	*Training*	0.745	0.437	0.403
	Testing	0.525	0.333	0.508

Performa metode *naïve bayes classifier* pada data *testing* digambarkan pada Gambar 4.24 sebagai berikut.



Gambar 4.24 Performa *Naïve Bayes Classifier* dengan Menggunakan Kelas Baru

Berdasarkan Tabel 4.54 dan Gambar 4.24, pada kategori sebanyak 2, nilai AUC tertinggi yaitu 0,592 yang didapatkan dengan 20% dan 25% prediktor. Pada kategori sebanyak 3, nilai AUC tertinggi didapatkan dari 15% prediktor yaitu sebesar 0,367. Sedangkan pada 4 kategori, nilai AUC tertinggi didapatkan dari 100% prediktor yaitu sebesar 0,508. Oleh karena itu, dapat disimpulkan nilai AUC paling tinggi yaitu 0,592 yang didapatkan menggunakan 2 kategori dengan 20% dan 25% prediktor.

4.5.2 Klasifikasi Senyawa Obat Kanker dengan *Classification and Regression Tree* (CART) Menggunakan Kelas dari Hasil Pengelompokan dengan Pendekatan *Normal Mixture Distribution*

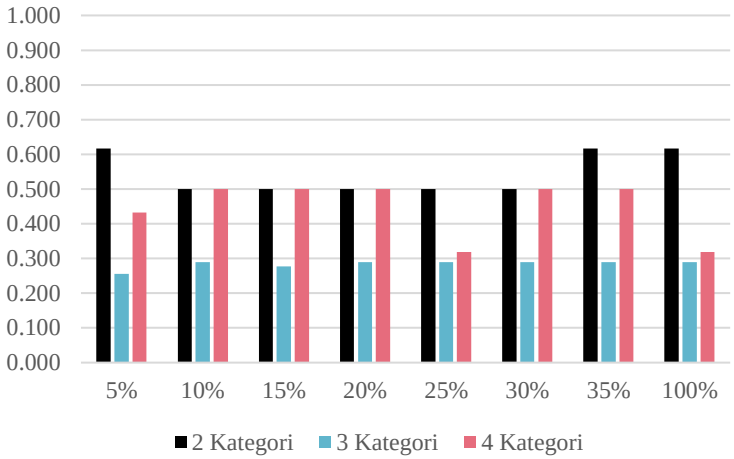
Sama seperti saat menggunakan metode *naïve bayes,* pembagian data yang digunakan yaitu dengan membagi data *training* dan data *testing* sebesar 70% dan 30%. Variabel yang

digunakan juga sebanyak 5%, 10%, 15%, 20%, 25%, 30%, 35%, dan 100% variabel berdasarkan nilai *mean decrease gini* terbesar. Metode evaluasi yang digunakan juga sama yaitu AUC. Sebelum melakukan klasifikasi, terlebih dahulu dilakukan *parameter tuning* yaitu *complexity parameter* (CP) yang optimal. Berikut merupakan performa *classification and regression tree* (CART) menggunakan variabel respon hasil pendekatan *mixture distribution*.

Tabel 4.55 Performa CART dengan Menggunakan Kelas Baru

Persentase Prediktor	Data	2 Kategori	3 Kategori	4 Kategori
5%	*Training*	0.676	0.492	0.347
	Testing	0.617	0.256	0.432
10%	*Training*	0.500	0.439	0.250
	Testing	0.500	0.289	0.500
15%	*Training*	0.500	0.497	0.250
	Testing	0.500	0.278	0.500
20%	*Training*	0.500	0.439	0.250
	Testing	0.500	0.289	0.500
25%	*Training*	0.500	0.439	0.403
	Testing	0.500	0.289	0.318
30%	*Training*	0.500	0.439	0.250
	Testing	0.500	0.289	0.500
35%	*Training*	0.737	0.439	0.250
	Testing	0.617	0.289	0.500
100%	*Training*	0.737	0.439	0.410
	Testing	0.617	0.289	0.318

Tabel 4.55 menunjukkan performa dari metode CART pada masing-masing kategori. Performa metode CART pada data *testing* digambarkan pada Gambar 4.25 sebagai berikut.



Gambar 4.25 Performa CART dengan Menggunakan Kelas Baru

Berdasarkan Tabel 4.55 dan Gambar 4.25, pada kategori sebanyak 2, nilai AUC tertinggi yaitu 0,617 yang didapatkan dengan 5%, 35%, dan 100% prediktor. Pada kategori sebanyak 3, nilai AUC tertinggi didapatkan dari 10%, 20%, 25%, 30%, 35%, dan 100% prediktor yaitu sebesar 0,289. Sedangkan pada 4 kategori, nilai AUC tertinggi didapatkan dari 10%, 15%, 20%, 30%, dan 35% prediktor yaitu sebesar 0,5. Oleh karena itu, dapat disimpulkan nilai AUC paling tinggi yaitu 0,617 yang didapatkan menggunakan 2 kategori dengan 35% dan 100% prediktor.

4.6 Perbandingan Nilai Ketepatan Klasifikasi Menggunakan Kelas Awal dan Kelas Setelah Pengkategorian Menggunakan *Normal Mixture Distribution*

Pada sub bab 4.9 akan dibandingkan nilai AUC klasifikasi dengan menggunakan kelas awal (proteksi radiasi rendah dan tinggi) dengan kelas baru hasil pengkategorian dengan menggunkan pendekatan *normal mixture distribution*. Berikut merupakan nilai AUC yang dihasilkan.

Tabel 4.56 Perbandingan Ketepatan Klasifikasi dengan Kelas Awal dan Kelas Baru

Persentase Prediktor	Data	NBC		CART	
		Awal	**Baru**	**Awal**	**Baru**
5%	*Training*	0,594	0.642	0,795	0.676
	Testing	**0.549**	0.483	0.658	**0.617**
10%	*Training*	0,599	0.643	0,825	0.500
	Testing	0.537	0.517	**0.663**	0.500
15%	*Training*	0,623	0.601	0,733	0.500
	Testing	0.477	0.567	0.458	0.500
20%	*Training*	0,610	0.612	0,770	0.500
	Testing	0.499	**0.592**	0.476	0.500
25%	*Training*	0,606	0.591	0,829	0.500
	Testing	0.512	**0.592**	0.548	0.500
30%	*Training*	0,607	0.591	0,763	0.500
	Testing	0.522	0.567	0.439	0.500
35%	*Training*	0,630	0.696	0,825	0.737
	Testing	0.551	0.567	0.541	**0.617**
100%	*Training*	0,579	0.745	0,830	0.737
	Testing	0.472	0.525	0.533	**0.617**

Berdasarkan Tabel 4.56, dapat diketahui bahwa pada kelas awal, CART menghasilkan nilai AUC sebesar 0,663 dengan 10% prediktor terpenting, sedangkan *naïve bayes* menghasilkan AUC sebesar 0,549 dengan 5% prediktor terpenting. Pada kelas baru, CART menghasilkan nilai AUC sebesar 0,617 dengan 5%, 35%, dan 100% prediktor terpenting, sedangkan *naïve bayes* menghasilkan nilai AUC sebesar 0,592 dengan 20% dan 25% prediktor terpenting. Oleh karena itu, dapat disimpulkan bahwa metode CART menghasilkan nilai AUC lebih tinggi dibandingkan dengan metode *naïve bayes* baik di kelas awal maupun kelas baru.

Selain itu, dapat juga dilihat bahwa dengan menggunakan kelas baru hasil pengelompokan dengan menggunakan pendekatan *normal mixture distribution* dapat menaikkan AUC pada metode NBC yang awalnya sebesar 0,549 dengan

menggunakan 5% prediktor menjadi 0,592 dengan menggunakan 20% dan 25% prediktor. Berbeda dengan metode NBC, metode CART dengan kelas baru menghasilkan nilai AUC yang lebih rendah yaitu 0,617 dengan menggunakan 5% prediktor, yang awalnya sebesar 0,663 dengan menggunakan 10% prediktor.

Kenaikan nilai AUC pada metode NBC relatif tidak signifikan, pada metode CART juga tidak menghasilkan nilai AUC yang lebih tinggi. Oleh karena itu, hipotesis bahwa naiknya akurasi dengan membentuk kategori atau kelas baru dengan menggunakan pendekatan *mixture distribution* belum sepenuhnya terbukti benar. Hal tersebut dikarenakan dengan menggunakan kelas baru tersebut tidak menambah akurasi yang signifikan.

Rendahnya nilai ketepatan klasifikasi menggunakan kelas baru hasil pengelompokan menggunakan pendekatan *normal mixture distribution* diduga karena pengelompokan yang tidak seimbang (*unbalance*). Oleh karena itu diperlukan *treatment* agar pengelompokan tersebut seimbang sehingga metode yg digunakan bisa memberikan performa lebih baik.

(*Halaman ini sengaja dikosongkan*)

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan, diperoleh kesimpulan sebagai berikut.

1 Lima variabel prediktor yang memiliki nilai *mean decrease gini* (MDG) terbesar adalah Jurs FNSA 1, Molecular SASA, Minimized Energy, ES Sum aaN, dan Dipole Z. Selain itu, terdapat 16 variabel prediktor yang memiliki nilai MDG 0 yaitu F Count, ES Sum aaO, ES Sum ddsN, ES Sum dsN, I Count, ES Count aaaC, ES Count aaNH, ES Count aaO, ES Count ddsN, ES Count dsN, ES Count sCI, ES Count sF, ES Count sI, Num Rings 5, Num Stereo Bonds, dan Num True Stereo Atoms.

2 Pada dua kelas awal proteksi radiasi, metode CART menghasilkan nilai ketepatan klasifikasi lebih tinggi daripada *naïve bayes classifier*. Nilai AUC tertinggi yaitu sebesar 0,549 menggunakan metode *naïve bayes classifier* dengan 5% prediktor terpenting, dan pada metode CART menghasilkan nilai AUC tertinggi sebesar 0,663 menggunakan 10% prediktor terpenting.

3 Pengelompokan menggunakan pendekatan *mixture distribution* pada 2 kategori menghasilkan kategori 2 lebih bagus dari kategori 1 jika dilihat dari tingkat kematian sel yang dihasilkan. Pada 3 kategori, secara berurutan kategori yang lebih bagus adalah kategori 2, 3, dan 1. Sedangkan pada 4 kategori, kategori 3 adalah kategori terbaik dan dilanjutkan dengan kategori 2, 1, dan 4.

4 Pada kelas baru hasil pengelompokan dengan pendekatan *normal mixture distribution*, metode CART memiliki performa lebih baik daripada *naïve bayes classifier*. Kelas yang menghasilkan nilai AUC tertinggi adalah dengan menggunakan 2 kelas, dengan Nilai AUC tertinggi metode *naïve bayes classifier* adalah 0,592 menggunakan 20% dan

25% prediktor. Pada metode CART, menghasilkan nilai AUC tertinggi yaitu 0,617 dengan menggunakan 5% prediktor.

5 Metode CART menghasilkan nilai AUC lebih tinggi dibandingkan dengan metode *naïve bayes* baik di kelas awal maupun kelas baru. Perbedaan diantara kedua metode tersebut adalah metode *naïve bayes classifier* memberikan performa lebih bagus dengan menggunakan kelas baru, berbeda dengan metode CART yang jika menggunakan kelas baru, tidak memberikan hasil yang lebih baik.

5.2 Saran

Berdasarkan kesimpulan yang diperoleh, dapat dirumuskan saran sebagai pertimbangan penelitian selanjutnya adalah sebagai berikut.

1. Sebaiknya dilakukan penambahan observasi senyawa obat kanker, sehingga data tidak tergolong *high dimensional data*.
2. Penentuan *threshold* untuk klasifikasi sebaiknya didiskusikan lagi dengan ahli, agar hasil penelitian yang didapatkan lebih akurat baik secara medis maupun statistik.
3. Jika *threshold* atau kategori baru akan digunakan, maka akan lebih baik jika data yang tidak seimbang (*unbalance*) diatasi dengan menggunakan *oversampling* atau *undersampling*.

DAFTAR PUSTAKA

Ariyasu, S., Sawa, A., Morita, A., Hanaya, K., Hoshi, M., Takahashi, I., Wang, B., Aoki, S. (2014). Design and Synthesis of 8-Hydroxyquinoline-Based Radioprotective Agents. *Bioorganic and Medicinal Chemistry, 22*(15), 3891-3905.

Aruna, S., Rajagopalan, S. P., & Nandakishore, L. V. (2011). Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer. *Computer Science and Information Technology, 2*, 37-45.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation Measures for Models Assessment Over Imbalanced Data Sets. *Journal of Information Engineering and Applications. 3*(10). 27-38.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. S. (2009). Mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software. 32*(6). 1-29.

Berthold, M. R., & Hand, D. J. (2010). *Intelligent Data Analysis* (2nd ed.). Berlin: Springer.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1993). *Classification and Regression Trees.* New York: Chapman Hall.

Bustami. (2013). Penerapan Algoritma Naive Bayes untuk Mengklasifikasikan data Nasabah Asuransi. *TECHSI : Jurnal Penelitian Teknik Informatika, 6*(2), 256-261.

Calle, M. L., & Urrea, V. (2010). Stability of Random Forest Importance Measures. *Briefings in Bioinformatics, 12*(1), 86-89.

Cancer Council Australia. (2018). *Radiotherapy*. Diakses pada 25 Februari 2018, dari Cancer Council Australia: https://www.cancer.org.au/about-cancer/treatment/radiotherapy.html

CART Reference Guide. (2000). *CART User's Guide.* San Diego: Salford System.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via The EM Algorithm (with discussion). *Journal of the Royal Statistical Society, B*(39), 1-38.

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of The C4.5 and Naive Bayes Classifier for The Prediction of Lung Cancer Survivability. *Journal of Computing, 4*(8).

Elsayad, A. M., & Elsalamony, H. A. (2013). Diagnosis of Breast Cancer Using Decision Tree Models and SVM. *International Journal of Computer Applications, 83*(5), 19-29.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Austria: Springer Science.

Gorunescu, Florin. (2011). *Data Mining : Concepts, Models, and Techniques*. Berlin: Springer-Verlag Berlin Heidelberg.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques.* San Fransisco: Morgan Kaufmann Publisher.

International Agency for Research on Cancer. (2013). *Prediction*. Diakses pada 25 Januari 2018, dari GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence in 2012: http://globocan.iarc.fr/Pages/burden_sel.aspx

International Agency for Research on Cancer. (2014). *World Cancer Report 2014.* Lyon: International Agency for Research on Cancer.

Iriawan, N. (2001). *Studi Tentang Bayesian Mixture Normal dengan Menggunakan Metode Markov Chain Monte Carlo (MCMC).* Laporan Penelitian Jurusan Statistika ITS, Surabaya.

Johnson, R., & Winchern, D. (2007). *Applied Multivariate Statistical Analysis (6th Edition).* New Jersey: Prentice Hall.

Kamus Kesehatan. (2018). *P53*. Diakses pada 22 Januari 2018, dari Kamus Kesehatan: http://kamuskesehatan.com/arti/p53/

Kathija, & Nisha, S. (2016). Breast Cancer Data Classification Using SVM and Naive Bayes Techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 4*(12), 21167-21175.

Kementrerian Kesehatan Republik Indonesia. (2015). *InfoDatin Situasi Penyakit Kanker*. Jakarta: Kementrerian Kesehatan Republik Indonesia.

Mandal, S. K. (2017). Performance Analysis of Data Mining Algorithms for Breast Cancer Cell Detection Using Naive Bayes, Logistic Regression, and Decision Tree. *International Journal of Engineering and Computer Science, 6*(2), 20388-20391.

Marco, S., & Zuccolotto, P. (2006). Variable Selection Using Random Forests. *Classification and Data Analysis Group (CLADAG)* (pp. 263-270). Berlin Heidelberg: Springer.

Matsumoto, A., Aoki, S., & Ohwada, H. (2016). Comparison of Random Forest and SVM for Raw Data in Drug Discovery: Prediction of Radiation Protection and Toxicity Case Study. *International Journal of Machine Learning and Computing, 6*(2), 145-148.

Morita, A., Ariyasu, S., Wang, B., Asamaru, T., Onoda, T., Sawa, A., Tanaka, K., Takahashi, I., Togami, S., Nenoi, M., Inaba., T., Aoki, S. (2014). AS-2, A Novel Inhibitor of P53-Dependent Apoptosis, Prevents Apoptotic Mitochondrial Dysfunction in A Transcription-Independent Manner and Protects Mice from A Lethal Dose of Ionizing Radiation. *Biochemichal and Biophysical Research Communications, 450*(4), 1498-1504

McLachlan, G. J., Khrisnan, T. (1996). *The EM Algorithm and Extensions.* New Jersey: John Wiley & Sons, Inc.

National Cancer Institute. (2015). *Understanding Cancer*. Diakses pada 24 Februari 2018, dari National Cancer Institute: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Patil, T. R., & Sherekar, M. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal of Computer Science and Applications, 6*(2), 256-261.

Pattekari, S. A., & Parveen, A. (2012). Prediction System for Heart Disease Using Naive Bayes. *International Journal of Advanced Computer and Mathematical Sciences, 3*(3), 290-294.

Price, S. A., & Wilson, L. M. (2006). *Patofisiologi Konsep Klinis Proses-Proses Penyakit* (6th ed.). Jakarta: EGC.

SAS. (2018). *Machine Learning*. Diakses pada 22 Januari 2018, dari SAS : The Power to Know: https://www.sas.com/en_us/insights/analytics/machine-learning.html

Soria, D., Garibaldi, J. M., Biganzoli, E., & Ellis, I. O. (2008). A Comparison of Three DIfferent Methods for Classification of Breast Cancer Data. *Seventh International Conference on Machine Learning and Applications* (pp. 619-624). San Diego: 10.1109/ICMLA.2008.97.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability and Statistics for Engineers and Scientist* (9th ed.). Boston: Prentice Hall.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition)*. United States: Morgan Kaufmann.

World Health Organization. (2009). *Cancer*. Diakses pada 19 Januari 2018, dari World Health Organization: http://www.who.int/cancer/en/

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. New York: CRC Press.

(*Halaman ini sengaja dikosongkan*)

LAMPIRAN

Lampiran 1. Data Penelitian

Senyawa	Target Kelas	Tingkat Kematian Sel	Penyusun Senyawa				
			pKa	ALogP 98	ALogP MR	…	Zagreb
AS-1	1	30	20	1.911	113.928	⋯	150
AS-10	1	25	11.4	-0.404	75.782	⋯	112
AS-11	0	0	7.9	0.685	58.728	⋯	84
AS-12	1	25	8.3	0.967	63.319	⋯	90
AS-13	0	0	20	2.282	51.034	⋯	66
AS-15	1	25	20	1.193	68.088	⋯	94
AS-16	1	25	10.1	0.104	85.143	⋯	122
AS-17	1	45	20	-0.321	93.784	⋯	140
AS-2	1	40	11.5	-0.121	80.374	⋯	118
AS-3	0	-20	20	2.664	51.247	⋯	66
AS-4	0	-5	20	3.013	55.995	⋯	70
AS-5	1	20	20	1.399	72.985	⋯	100
AS-6	1	20	20	1.748	77.733	⋯	104
AS-7	1	15	6.4	1.752	80.624	⋯	102
AS-8	1	20	20	0.865	99.684	⋯	138
AS-9	0	5	20	2.26	118.676	⋯	154
KH-1	1	15	9.8	1.774	41.673	⋯	56
KH-10	0	10	10.1	2.056	46.264	⋯	62
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
YT-1	1	35	9.8	-0.742	121.409	⋯	174

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG)

Variabel	Nama Variabel	MDG
X91	Jurs_FNSA_1	0.582
X48	Molecular_SASA	0.573
X120	Minimized_Energy	0.509
X6	ES_Sum_aaN	0.487
X87	Dipole_Z	0.479
X41	SAscore	0.473
X134	Shadow_YZ	0.460
X45	Molecular_FractionalPolarSurfaceArea	0.459
X73	Kappa_1_AM	0.455
X40	QED_Unweighted	0.449
X70	JX	0.436
X43	SAscore_Fragments	0.434
X110	Jurs_TASA	0.426
X119	Energy	0.422
X107	Jurs_RPCS	0.415
X5	ES_Sum_aaCH	0.414
X114	Jurs_WNSA_3	0.408
X57	CHI_V_0	0.396
X69	IC	0.393
X67	IAC_Mean	0.393
X60	CHI_V_3_C	0.378
X139	Molecular_3D_SAVol	0.373

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X128	Shadow_Xlength	0.372
X108	Jurs_RPSA	0.366
X127	Shadow_nu	0.360
X42	SAscore_Complexity	0.360
X104	Jurs_RNCG	0.357
X8	ES_Sum_aasC	0.355
X136	Shadow_Zlength	0.349
X216	Wiener	0.345
X71	JY	0.344
X103	Jurs_RASA	0.341
X2	ALogP98	0.332
X111	Jurs_TPSA	0.329
X31	QED	0.325
X15	ES_Sum_sCH3	0.325
X112	Jurs_WNSA_1	0.322
X85	Dipole_X	0.320
X56	CHI_3_P	0.317
X65	E_DIST_equ	0.317
X130	Shadow_XYfrac	0.315
X75	Kappa_2_AM	0.314
X59	CHI_V_2	0.314
X137	Molecular_3D_PolarSASA	0.311

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X37	QED_MW	0.310
X55	CHI_3_C	0.307
X93	Jurs_FNSA_3	0.305
X50	Molecular_SurfaceArea	0.303
X92	Jurs_FNSA_2	0.299
X117	Jurs_WPSA_3	0.299
X217	Zagreb	0.298
X78	PHI	0.298
X99	Jurs_PNSA_3	0.295
X97	Jurs_PNSA_1	0.291
X30	Molecular_Solubility	0.290
X61	CHI_V_3_P	0.289
X133	Shadow_Ylength	0.286
X121	RadOfGyration	0.284
X102	Jurs_PPSA_3	0.284
X135	Shadow_YZfrac	0.283
X83	V_DIST_mag	0.282
X38	QED_PSA	0.281
X77	Kappa_3_AM	0.278
X138	Molecular_3D_SASA	0.276
X124	PMI_X	0.273
X20	ES_Sum_sOH	0.270

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X132	Shadow_XZfrac	0.270
X131	Shadow_XZ	0.269
X84	Dipole_mag	0.266
X27	Apol	0.266
X89	Jurs_DPSA_2	0.262
X86	Dipole_Y	0.261
X106	Jurs_RPCG	0.260
X115	Jurs_WPSA_1	0.254
X118	AverageBondLength	0.251
X79	SIC	0.250
X80	V_ADJ_equ	0.249
X116	Jurs_WPSA_2	0.248
X62	CIC	0.247
X187	Num_ExplicitAtoms	0.242
X96	Jurs_FPSA_3	0.241
X63	E_ADJ_equ	0.238
X95	Jurs_FPSA_2	0.236
X66	E_DIST_mag	0.236
X46	Molecular_PolarSASA	0.236
X122	Strain_Energy	0.227
X109	Jurs_SASA	0.227
X183	Num_Bonds	0.226

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X29	Molecular_Mass	0.222
X98	Jurs_PNSA_2	0.217
X33	QED_ALOGP	0.216
X210	Organic_Count	0.214
X47	Molecular_PolarSurfaceArea	0.211
X168	ES_Count_ssCH2	0.210
X44	Molecular_FractionalPolarSASA	0.209
X39	QED_ROTB	0.205
X129	Shadow_XY	0.204
X194	Num_Hydrogens	0.203
X185	Num_Chains	0.202
X36	QED_HBD	0.201
X90	Jurs_DPSA_3	0.200
X82	V_DIST_equ	0.199
X88	Jurs_DPSA_1	0.198
X193	Num_H_Donors_Lipinski	0.198
X215	SC_3_P	0.197
X21	ES_Sum_ssCH2	0.195
X3	ALogP_MR	0.190
X54	CHI_2	0.189
X72	Kappa_1	0.187
X76	Kappa_3	0.184

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X94	Jurs_FPSA_1	0.183
X204	Num_RotatableBonds	0.182
X64	E_ADJ_mag	0.179
X212	SC_1	0.175
X126	PMI_Z	0.174
X1	pKa(max20)	0.174
X28	LogD	0.169
X178	Num_AliphaticSingleBonds	0.167
X51	BIC	0.166
X113	Jurs_WNSA_2	0.165
X68	IAC_Total	0.165
X101	Jurs_PPSA_2	0.165
X74	Kappa_2	0.163
X140	Molecular_Volume	0.163
X145	H_Count	0.163
X53	CHI_1	0.162
X162	ES_Count_sCH3	0.162
X192	Num_H_Donors	0.159
X25	ES_Sum_sssN	0.151
X11	ES_Sum_dO	0.149
X49	Molecular_SAVol	0.148
X58	CHI_V_1	0.139

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X188	Num_ExplicitBonds	0.137
X213	SC_2	0.135
X182	Num_Atoms	0.134
X22	ES_Sum_ssNH	0.130
X23	ES_Sum_ssO	0.129
X214	SC_3_C	0.127
X189	Num_ExplicitHydrogens	0.124
X169	ES_Count_ssNH	0.122
X4	ES_Sum_aaaC	0.120
X147	N_Count	0.119
X123	PMI_mag	0.118
X176	NPlusO_Count	0.114
X177	Num_AliphaticDoubleBonds	0.113
X52	CHI_0	0.110
X186	Num_DoubleBonds	0.105
X125	PMI_Y	0.104
X205	Num_SingleBonds	0.103
X158	ES_Count_dO	0.099
X142	C_Count	0.098
X175	HBD_Count	0.096
X105	Jurs_RNCS	0.094
X172	ES_Count_sssN	0.087

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

Variabel	Nama Variabel	MDG
X100	Jurs_PPSA_1	0.086
X184	Num_ChainAssemblies	0.082
X211	SC_0	0.080
X181	Num_AtomClasses	0.079
X35	QED_HBA	0.073
X198	Num_RingBonds	0.071
X191	Num_H_Acceptors_Lipinski	0.070
X10	ES_Sum_ddssS	0.069
X190	Num_H_Acceptors	0.063
X32	QED_ALERTS	0.058
X148	O_Count	0.058
X167	ES_Count_sOH	0.056
X151	ES_Count_aaCH	0.055
X152	ES_Count_aaN	0.055
X14	ES_Sum_sBr	0.050
X208	Num_TerminalRotomers	0.049
X16	ES_Sum_sCl	0.048
X202	Num_Rings6	0.043
X206	Num_StereoAtoms	0.043
X81	V_ADJ_mag	0.037
X34	QED_AROM	0.034
X17	ES_Sum_sF	0.033

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

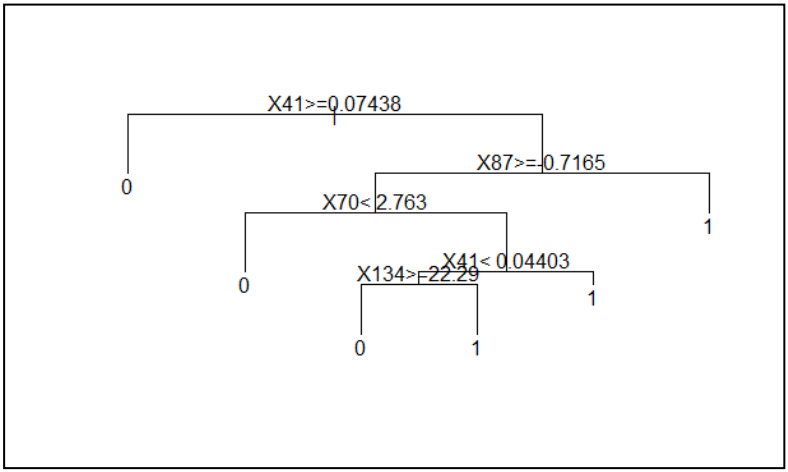
Variabel	Nama Variabel	MDG
X155	ES_Count_aasC	0.032
X157	ES_Count_ddssS	0.030
X170	ES_Count_ssO	0.028
X197	Num_RingAssemblies	0.028
X195	Num_NegativeAtoms	0.027
X149	S_Count	0.024
X203	Num_Rings7	0.023
X174	HBA_Count	0.023
X24	ES_Sum_sssCH	0.020
X166	ES_Count_sNH2	0.017
X173	ES_Count_ssssC	0.016
X171	ES_Count_sssCH	0.016
X160	ES_Count_dssC	0.013
X200	Num_Rings	0.013
X19	ES_Sum_sNH2	0.013
X179	Num_AromaticBonds	0.012
X161	ES_Count_sBr	0.010
X13	ES_Sum_dssC	0.008
X26	ES_Sum_ssssC	0.007
X141	Br_Count	0.006
X143	Cl_Count	0.006
X199	Num_RingFusionBonds	0.006

Lampiran 2. Tingkat Kepentingan Variabel Berdasarkan *Mean Decrease Gini* (MDG) (Lanjutan)

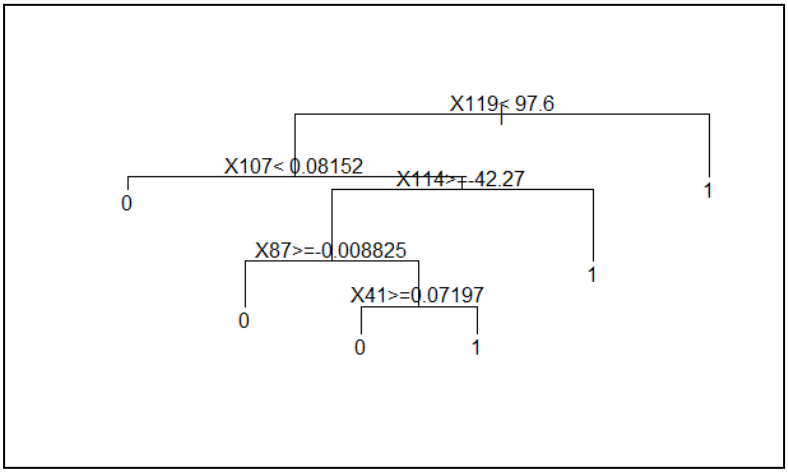
Variabel	Nama Variabel	MDG
X18	ES_Sum_sI	0.004
X196	Num_PositiveAtoms	0.003
X180	Num_AromaticRings	0.002
X144	F_Count	0.000
X7	ES_Sum_aaO	0.000
X9	ES_Sum_ddsN	0.000
X12	ES_Sum_dsN	0.000
X146	I_Count	0.000
X150	ES_Count_aaaC	0.000
X153	ES_Count_aaNH	0.000
X154	ES_Count_aaO	0.000
X156	ES_Count_ddsN	0.000
X159	ES_Count_dsN	0.000
X163	ES_Count_sCl	0.000
X164	ES_Count_sF	0.000
X165	ES_Count_sI	0.000
X201	Num_Rings5	0.000
X207	Num_StereoBonds	0.000
X209	Num_TrueStereoAtoms	0.000

Lampiran 3. Pohon Klasifikasi dengan Dua Kelas Proteksi Radiasi Menggunakan Metode CART

(a) Pohon Klasifikasi CART dengan Prediktor 5%

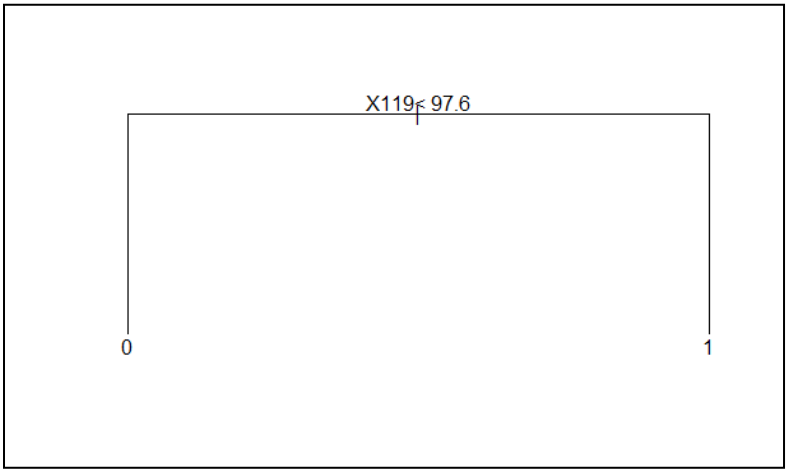


(b) Pohon Klasifikasi CART dengan Prediktor 10%

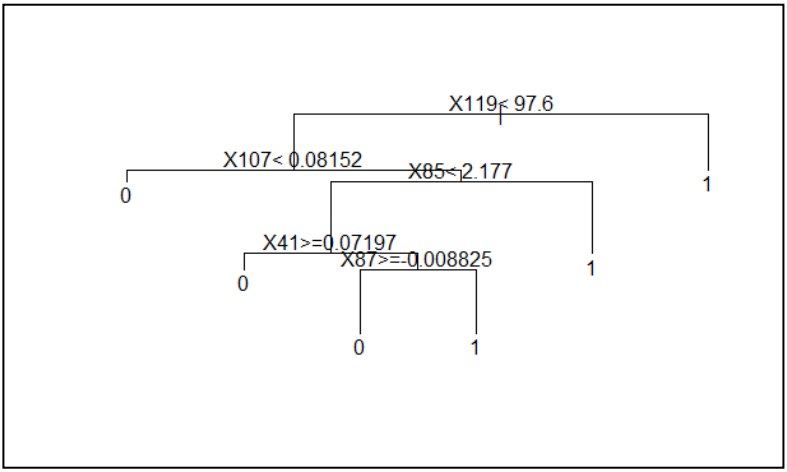


Lampiran 3. Pohon Klasifikasi dengan Dua Kelas Proteksi Radiasi Menggunakan Metode CART (Lanjutan)

(c) Pohon Klasifikasi CART dengan Prediktor 15%

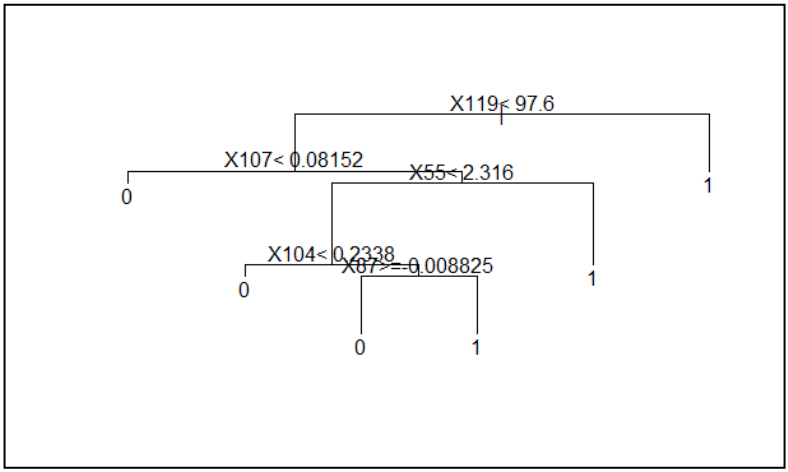


(d) Pohon Klasifikasi CART dengan Prediktor 20%

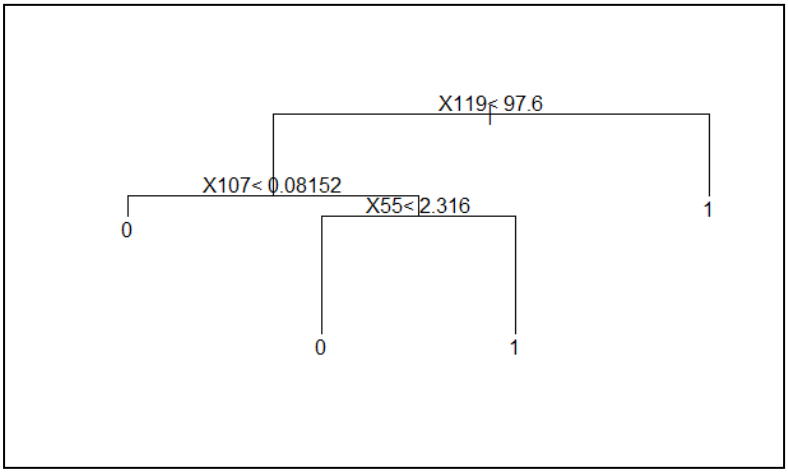


Lampiran 3. Pohon Klasifikasi dengan Dua Kelas Proteksi Radiasi Menggunakan Metode CART (Lanjutan)

(e) Pohon Klasifikasi CART dengan Prediktor 25%

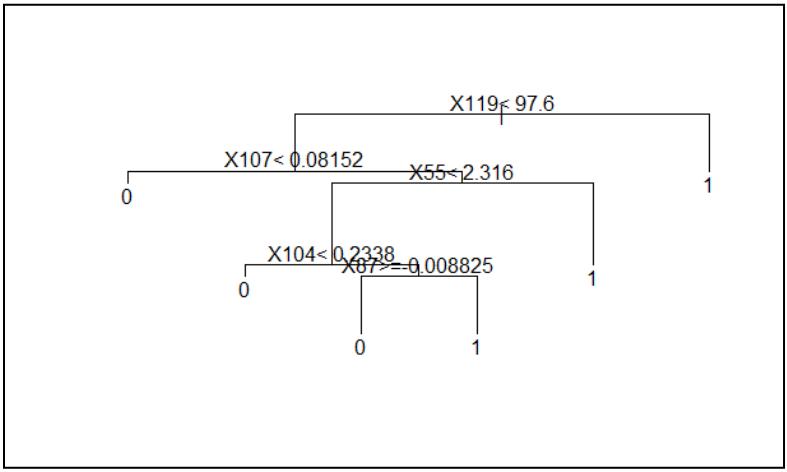


(f) Pohon Klasifikasi CART dengan Prediktor 30%

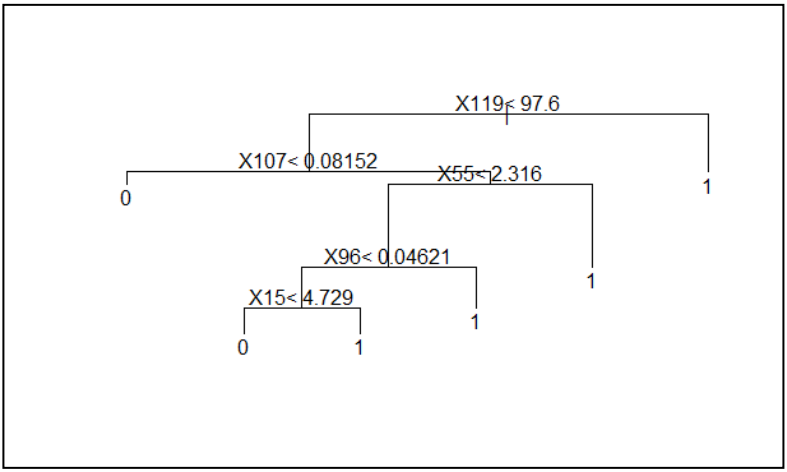


Lampiran 3. Pohon Klasifikasi dengan Dua Kelas Proteksi Radiasi Menggunakan Metode CART (Lanjutan)

(g) Pohon Klasifikasi CART dengan Prediktor 35%



(h) Pohon Klasifikasi CART dengan Prediktor 100%



Lampiran 4. Model CART yang Dihasilkan dari *Software* R

(a) Model Klasifikasi CART dengan Prediktor 5%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.52631579 0.47368421)
   2) X41>=0.07437925 16  3 0 (0.81250000 0.18750000) *
   3) X41< 0.07437925 60 27 1 (0.45000000 0.55000000)
     6) X87>=-0.716525 48 22 0 (0.54166667 0.45833333)
      12) X70< 2.762935 16  3 0 (0.81250000 0.18750000) *
      13) X70>=2.762935 32 13 1 (0.40625000 0.59375000)
        26) X41< 0.0440333 23 11 0 (0.52173913 0.47826087)
          52) X134>=22.29165 16  5 0 (0.68750000 0.31250000) *
          53) X134< 22.29165 7  1 1 (0.14285714 0.85714286) *
        27) X41>=0.0440333 9  1 1 (0.11111111 0.88888889) *
     7) X87< -0.716525 12  1 1 (0.08333333 0.91666667) *
```

(b) Model Klasifikasi CART dengan Prediktor 10%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.5263158 0.4736842)
   2) X119< 97.6 69 29 0 (0.5797101 0.4202899)
     4) X107< 0.081515 10  0 0 (1.0000000 0.0000000) *
     5) X107>=0.081515 59 29 0 (0.5084746 0.4915254)
      10) X114>=-42.2658 49 20 0 (0.5918367 0.4081633)
        20) X87>=-0.008825 22  4 0 (0.8181818 0.1818182) *
        21) X87< -0.008825 27 11 1 (0.4074074 0.5925926)
          42) X41>=0.071971 11  4 0 (0.6363636 0.3636364) *
          43) X41< 0.071971 16  4 1 (0.2500000 0.7500000) *
      11) X114< -42.2658 10  1 1 (0.1000000 0.9000000) *
   3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

Lampiran 4. Model CART yang Dihasilkan dari *Software* R (Lanjutan)

(c) Model Klasifikasi CART dengan Prediktor 15%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 76 36 0 (0.5263158 0.4736842)
  2) X119< 97.6 69 29 0 (0.5797101 0.4202899) *
  3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

(d) Model Klasifikasi CART dengan Prediktor 20%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.52631579 0.47368421)
   2) X119< 97.6 69 29 0 (0.57971014 0.42028986)
     4) X107< 0.081515 10  0 0 (1.00000000 0.00000000) *
     5) X107>=0.081515 59 29 0 (0.50847458 0.49152542)
      10) X85< 2.17656 48 19 0 (0.60416667 0.39583333)
        20) X41>=0.071971 14  1 0 (0.92857143 0.07142857) *
        21) X41< 0.071971 34 16 1 (0.47058824 0.52941176)
          42) X87>=-0.008825 18  5 0 (0.72222222 0.27777778) *
          43) X87< -0.008825 16  3 1 (0.18750000 0.81250000) *
      11) X85>=2.17656 11  1 1 (0.09090909 0.90909091) *
   3) X119>=97.6 7  0 1 (0.00000000 1.00000000) *
```

Lampiran 4. Model CART yang Dihasilkan dari *Software* R (Lanjutan)

(e) Model Klasifikasi CART dengan Prediktor 25%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.5263158 0.4736842)
   2) X119< 97.6 69 29 0 (0.5797101 0.4202899)
     4) X107< 0.081515 10  0 0 (1.0000000 0.0000000) *
     5) X107>=0.081515 59 29 0 (0.5084746 0.4915254)
      10) X55< 2.315635 45 17 0 (0.6222222 0.3777778)
        20) X104< 0.233805 16  2 0 (0.8750000 0.1250000) *
        21) X104>=0.233805 29 14 1 (0.4827586 0.5172414)
          42) X87>=-0.008825 15  4 0 (0.7333333 0.2666667) *
          43) X87< -0.008825 14  3 1 (0.2142857 0.7857143) *
      11) X55>=2.315635 14  2 1 (0.1428571 0.8571429) *
   3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

(f) Model Klasifikasi CART dengan Prediktor 30%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.5263158 0.4736842)
   2) X119< 97.6 69 29 0 (0.5797101 0.4202899)
     4) X107< 0.081515 10  0 0 (1.0000000 0.0000000) *
     5) X107>=0.081515 59 29 0 (0.5084746 0.4915254)
      10) X55< 2.315635 45 17 0 (0.6222222 0.3777778) *
      11) X55>=2.315635 14  2 1 (0.1428571 0.8571429) *
   3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

Lampiran 4. Model CART yang Dihasilkan dari *Software* R (Lanjutan)

(g) Model Klasifikasi CART dengan Prediktor 35%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.5263158 0.4736842)
   2) X119< 97.6 69 29 0 (0.5797101 0.4202899)
     4) X107< 0.081515 10  0 0 (1.0000000 0.0000000) *
     5) X107>=0.081515 59 29 0 (0.5084746 0.4915254)
      10) X55< 2.315635 45 17 0 (0.6222222 0.3777778)
        20) X104< 0.233805 16  2 0 (0.8750000 0.1250000) *
        21) X104>=0.233805 29 14 1 (0.4827586 0.5172414)
          42) X87>=-0.008825 15  4 0 (0.7333333 0.2666667) *
          43) X87< -0.008825 14  3 1 (0.2142857 0.7857143) *
      11) X55>=2.315635 14  2 1 (0.1428571 0.8571429) *
   3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

(h) Model Klasifikasi CART dengan Prediktor 100%

```
n= 76
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 76 36 0 (0.5263158 0.4736842)
   2) X119< 97.6 69 29 0 (0.5797101 0.4202899)
     4) X107< 0.081515 10  0 0 (1.0000000 0.0000000) *
     5) X107>=0.081515 59 29 0 (0.5084746 0.4915254)
      10) X55< 2.315635 45 17 0 (0.6222222 0.3777778)
        20) X96< 0.04621 34  9 0 (0.7352941 0.2647059)
          40) X15< 4.729 27  4 0 (0.8518519 0.1481481) *
          41) X15>=4.729 7  2 1 (0.2857143 0.7142857) *
        21) X96>=0.04621 11  3 1 (0.2727273 0.7272727) *
      11) X55>=2.315635 14  2 1 (0.1428571 0.8571429) *
   3) X119>=97.6 7  0 1 (0.0000000 1.0000000) *
```

Lampiran 5. Nilai *Posterior Probability* pada $k = 2$

Senyawa	Komponen 1	Komponen 2	Kategori
AS-1	0.232	0.768	2
AS-10	0.558	0.442	1
AS-11	1.000	0.000	1
AS-12	0.558	0.442	1
AS-13	1.000	0.000	1
AS-15	0.558	0.442	1
AS-16	0.558	0.442	1
AS-17	0.037	0.963	2
AS-2	0.050	0.950	2
AS-3	1.000	0.000	1
AS-4	1.000	0.000	1
AS-5	0.884	0.116	1
AS-6	0.884	0.116	1
AS-7	0.985	0.015	1
AS-8	0.884	0.116	1
AS-9	1.000	0.000	1
KH-1	0.985	0.015	1
KH-10	0.999	0.001	1
KH-12	1.000	0.000	1
KH-13	0.884	0.116	1
KH-16	1.000	0.000	1
KH-18	0.232	0.768	2

Lampiran 5. Nilai *Posterior Probability* pada $k = 2$ (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Kategori
KH-19	0.999	0.001	1
KH-2	0.884	0.116	1
KH-20	0.095	0.905	2
KH-21	1.000	0.000	1
KH-22	0.050	0.950	2
KH-23	1.000	0.000	1
KH-24	0.884	0.116	1
KH-25	0.884	0.116	1
KH-3	0.232	0.768	2
KH-4	0.985	0.015	1
KH-5	1.000	0.000	1
KH-6	1.000	0.000	1
KT-1	1.000	0.000	1
KT-2	0.095	0.905	2
MH-1	0.095	0.905	2
MH-10	1.000	0.000	1
MH-11	1.000	0.000	1
MH-12	0.999	0.001	1
MH-13	0.985	0.015	1
MH-14	1.000	0.000	1
MH-15	0.232	0.768	2
MH-16	0.985	0.015	1

Lampiran 5. Nilai *Posterior Probability* pada *k* = 2 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Kategori
MH-2	0.999	0.001	1
MH-3	1.000	0.000	1
MH-4	1.000	0.000	1
MH-5	1.000	0.000	1
MH-6	0.999	0.001	1
MH-7	0.999	0.001	1
MH-8	1.000	0.000	1
MH-9	1.000	0.000	1
MY-1	0.558	0.442	1
MY-2	0.232	0.768	2
naphthalene	0.999	0.001	1
naphthol	0.884	0.116	1
phenol	1.000	0.000	1
quinoform	1.000	0.000	1
quinoline	1.000	0.000	1
SAr-1	0.999	0.001	1
SAr-2	0.985	0.015	1
SAr-3	1.000	0.000	1
SAr-4	0.999	0.001	1
SAr-5	1.000	0.000	1
SAr-7	0.232	0.768	2
ST-1	1.000	0.000	1

Lampiran 5. Nilai *Posterior Probability* pada *k* = 2 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Kategori
TPEN	1.000	0.000	1
UM-10	0.050	0.950	2
UM-7	0.558	0.442	1
UM-8	1.000	0.000	1
UM-9	0.884	0.116	1
Vitamine_C	1.000	0.000	1
Vitamine_E	0.095	0.905	2
YM-13	0.095	0.905	2
YM-14	1.000	0.000	1
YM-16	1.000	0.000	1
YN-1	1.000	0.000	1
YN-4	1.000	0.000	1
YN-5	1.000	0.000	1
YN-6	1.000	0.000	1
YN-7	0.037	0.963	2
YN-8	0.999	0.001	1
YN-9	0.095	0.905	2
YT-1	0.095	0.905	2

Lampiran 6. Nilai *Posterior Probability* pada $k = 3$

Senyawa	Komponen 1	Komponen 2	Komponen 3	Kategori
AS-1	2.09E-97	9.98E-01	1.77E-03	2
AS-10	5.08E-80	9.82E-01	1.85E-02	2
AS-11	2.89E-19	1.04E-02	9.90E-01	3
AS-12	5.08E-80	9.82E-01	1.85E-02	2
AS-13	2.89E-19	1.04E-02	9.90E-01	3
AS-15	5.08E-80	9.82E-01	1.85E-02	2
AS-16	5.08E-80	9.82E-01	1.85E-02	2
AS-17	2.86E-159	1.00E+00	4.01E-07	2
AS-2	4.93E-137	1.00E+00	8.20E-06	2
AS-3	9.54E-01	2.64E-05	4.57E-02	1
AS-4	5.10E-12	3.64E-03	9.96E-01	3
AS-5	2.66E-64	8.62E-01	1.38E-01	2
AS-6	2.66E-64	8.62E-01	1.38E-01	2
AS-7	2.14E-50	4.76E-01	5.24E-01	3
AS-8	2.66E-64	8.62E-01	1.38E-01	2
AS-9	4.84E-28	3.59E-02	9.64E-01	3
KH-1	2.14E-50	4.76E-01	5.24E-01	3
KH-10	2.23E-38	1.42E-01	8.58E-01	3
KH-12	4.84E-28	3.59E-02	9.64E-01	3
KH-13	2.66E-64	8.62E-01	1.38E-01	2
KH-16	4.84E-28	3.59E-02	9.64E-01	3
KH-18	2.09E-97	9.98E-01	1.77E-03	2

Lampiran 6. Nilai *Posterior Probability* pada $k = 3$ (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Kategori
KH-19	2.23E-38	1.42E-01	8.58E-01	3
KH-2	2.66E-64	8.62E-01	1.38E-01	2
KH-20	2.06E-116	1.00E+00	1.34E-04	2
KH-21	2.89E-19	1.04E-02	9.90E-01	3
KH-22	4.93E-137	1.00E+00	8.20E-06	2
KH-23	2.71E-06	1.59E-03	9.98E-01	3
KH-24	2.66E-64	8.62E-01	1.38E-01	2
KH-25	2.66E-64	8.62E-01	1.38E-01	2
KH-3	2.09E-97	9.98E-01	1.77E-03	2
KH-4	2.14E-50	4.76E-01	5.24E-01	3
KH-5	2.89E-19	1.04E-02	9.90E-01	3
KH-6	4.84E-28	3.59E-02	9.64E-01	3
KT-1	5.10E-12	3.64E-03	9.96E-01	3
KT-2	2.06E-116	1.00E+00	1.34E-04	2
MH-1	2.06E-116	1.00E+00	1.34E-04	2
MH-10	4.84E-28	3.59E-02	9.64E-01	3
MH-11	4.84E-28	3.59E-02	9.64E-01	3
MH-12	2.23E-38	1.42E-01	8.58E-01	3
MH-13	2.14E-50	4.76E-01	5.24E-01	3
MH-14	5.10E-12	3.64E-03	9.96E-01	3
MH-15	2.09E-97	9.98E-01	1.77E-03	2
MH-16	2.14E-50	4.76E-01	5.24E-01	3

Lampiran 6. Nilai *Posterior Probability* pada *k* = 3 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Kategori
MH-2	2.23E-38	1.42E-01	8.58E-01	3
MH-3	4.84E-28	3.59E-02	9.64E-01	3
MH-4	2.89E-19	1.04E-02	9.90E-01	3
MH-5	2.89E-19	1.04E-02	9.90E-01	3
MH-6	2.23E-38	1.42E-01	8.58E-01	3
MH-7	2.23E-38	1.42E-01	8.58E-01	3
MH-8	4.84E-28	3.59E-02	9.64E-01	3
MH-9	2.71E-06	1.59E-03	9.98E-01	3
MY-1	5.08E-80	9.82E-01	1.85E-02	2
MY-2	2.09E-97	9.98E-01	1.77E-03	2
naphthalene	2.23E-38	1.42E-01	8.58E-01	3
naphthol	2.66E-64	8.62E-01	1.38E-01	2
phenol	2.89E-19	1.04E-02	9.90E-01	3
quinoform	4.84E-28	3.59E-02	9.64E-01	3
quinoline	4.84E-28	3.59E-02	9.64E-01	3
SAr-1	2.23E-38	1.42E-01	8.58E-01	3
SAr-2	2.14E-50	4.76E-01	5.24E-01	3
SAr-3	4.84E-28	3.59E-02	9.64E-01	3
SAr-4	2.23E-38	1.42E-01	8.58E-01	3
SAr-5	5.10E-12	3.64E-03	9.96E-01	3
SAr-7	2.09E-97	9.98E-01	1.77E-03	2
ST-1	9.97E-01	1.58E-06	3.28E-03	1

Lampiran 6. Nilai *Posterior Probability* pada *k* = 3 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Kategori
TPEN	2.89E-19	1.04E-02	9.90E-01	3
UM-10	4.93E-137	1.00E+00	8.20E-06	2
UM-7	5.08E-80	9.82E-01	1.85E-02	2
UM-8	5.10E-12	3.64E-03	9.96E-01	3
UM-9	2.66E-64	8.62E-01	1.38E-01	2
Vitamine_C	4.84E-28	3.59E-02	9.64E-01	3
Vitamine_E	2.06E-116	1.00E+00	1.34E-04	2
YM-13	2.06E-116	1.00E+00	1.34E-04	2
YM-14	5.10E-12	3.64E-03	9.96E-01	3
YM-16	4.84E-28	3.59E-02	9.64E-01	3
YN-1	4.84E-28	3.59E-02	9.64E-01	3
YN-4	5.10E-12	3.64E-03	9.96E-01	3
YN-5	4.84E-28	3.59E-02	9.64E-01	3
YN-6	5.10E-12	3.64E-03	9.96E-01	3
YN-7	2.86E-159	1.00E+00	4.01E-07	2
YN-8	2.23E-38	1.42E-01	8.58E-01	3
YN-9	2.06E-116	1.00E+00	1.34E-04	2
YT-1	2.06E-116	1.00E+00	1.34E-04	2

Lampiran 7. Nilai *Posterior Probability* pada $k = 4$

Senyawa	Komponen 1	Komponen 2	Komponen 3	Komponen 4	Kategori
AS-1	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3
AS-10	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
AS-11	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
AS-12	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
AS-13	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
AS-15	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
AS-16	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
AS-17	2.65E-105	4.55E-18	1.0000e+00 6	1.85E-160	2
AS-2	5.84E-86	4.11E-14	1.0000e+00 1	5.63E-138	2
AS-3	6.15E-08	9.32E-07	6.44E-04	9.99E-01	4
AS-4	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
AS-5	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
AS-6	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
AS-7	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3
AS-8	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
AS-9	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
KH-1	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3
KH-10	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
KH-12	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
KH-13	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
KH-16	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
KH-18	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3

Lampiran 7. Nilai *Posterior Probability* pada *k* = 4 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Komponen 4	Kategori
KH-19	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
KH-2	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
KH-20	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
KH-21	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
KH-22	5.84E-86	4.11E-14	1.0000e+00 1	5.63E-138	2
KH-23	9.63E-01	2.13E-02	1.56E-02	3.55E-06	1
KH-24	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
KH-25	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
KH-3	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3
KH-4	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3
KH-5	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
KH-6	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
KT-1	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
KT-2	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
MH-1	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
MH-10	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
MH-11	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
MH-12	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
MH-13	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3
MH-14	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
MH-15	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3
MH-16	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3

Lampiran 7. Nilai *Posterior Probability* pada *k* = 4 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Komponen 4	Kategori
MH-2	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
MH-3	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
MH-4	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
MH-5	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
MH-6	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
MH-7	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
MH-8	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
MH-9	9.63E-01	2.13E-02	1.56E-02	3.55E-06	1
MY-1	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
MY-2	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3
naphthalene	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
naphthol	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
phenol	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
quinoform	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
quinoline	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
SAr-1	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
SAr-2	2.42E-18	1.64E-01	8.36E-01	3.68E-50	3
SAr-3	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
SAr-4	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
SAr-5	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
SAr-7	4.45E-53	1.28E-07	1.00E+00	3.80E-97	3
ST-1	1.04E-14	6.45E-10	7.04E-05	1.00E+00	4

Lampiran 7. Nilai *Posterior Probability* pada *k* = 4 (Lanjutan)

Senyawa	Komponen 1	Komponen 2	Komponen 3	Komponen 4	Kategori
TPEN	3.99E-02	9.05E-01	5.47E-02	5.23E-19	2
UM-10	5.84E-86	4.11E-14	1.0000e+00 1	5.63E-138	2
UM-7	1.54E-39	4.38E-05	1.00E+00	8.02E-80	3
UM-8	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
UM-9	6.17E-28	5.03E-03	9.95E-01	3.89E-64	3
Vitamine_C	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
Vitamine_E	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
YM-13	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
YM-14	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
YM-16	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
YN-1	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
YN-4	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
YN-5	8.14E-06	9.18E-01	8.18E-02	5.63E-28	2
YN-6	8.71E-01	1.15E-01	1.40E-02	4.29E-12	1
YN-7	2.65E-105	4.55E-18	1.0000e+00 6	1.85E-160	2
YN-8	4.43E-11	7.19E-01	2.81E-01	3.22E-38	2
YN-9	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2
YT-1	1.50E-68	1.25E-10	1.0000e+00 4	1.69E-117	2

Lampiran 8. *Syntax Feature Importance* dengan *Mean Decrease Gini*

```
#import library
library(caret)
library(randomForest)

#import data
data=read.csv("D:/DATA/Kuliah/! Tugas Akhir/Hasil Olahan
Data/Data Protection Radiation.csv")
data$Y <- as.factor(data$Y)

#parameter tuning random forest
customRF <- list(type = "Classification", library = "randomForest",
loop = NULL)
customRF$parameters <- data.frame(parameter = c("mtry", "ntree"),
class = rep("numeric", 2), label = c("mtry", "ntree"))
customRF$grid <- function(x, y, len = NULL, search = "grid") {}
customRF$fit <- function(x, y, wts, param, lev, last, weights,
classProbs, ...) {
  randomForest(x, y, mtry = param$mtry, ntree=param$ntree, ...)
}
customRF$predict <- function(modelFit, newdata, preProc = NULL,
submodels = NULL)
  predict(modelFit, newdata)
customRF$prob <- function(modelFit, newdata, preProc = NULL,
submodels = NULL)
  predict(modelFit, newdata, type = "prob")
customRF$sort <- function(x) x[order(x[,1]),]
customRF$levels <- function(x) x$classes

control <- trainControl(method="repeatedcv", number=10, repeats=10)
tunegrid <- expand.grid(.mtry=c(1:20),
.ntree=c(100,200,300,400,500,600,700,800,900,1000))
```

Lampiran 8. Syntax Feature Importance dengan Mean Decrease Gini (Lanjutan)

```
custom <- train(Y~., data=data, method=customRF, tuneGrid=tunegrid,
trControl=control)
custom
summary(custom)
plot(custom)

#function var.share
var.share <- function(rf.obj, members) {
  count <- table(rf.obj$forest$xbestsplit)[-1]
  names(count) <- names(rf.obj$forest$ncat)
  share <- count[members] / sum(count[members])
  return(share)
}

#function group.importance
group.importance <- function(rf.obj, groups) {
  var.imp <- as.matrix(sapply(groups, function(g) {
    sum(importance(rf.obj, 2)[g, ]*var.share(rf.obj, g))
  }))
  colnames(var.imp) <- "MeanDecreaseGini"
  return(var.imp)
}

#mean decrease gini
rf.obj <- randomForest(Y ~ ., data=data, ntree=300, mtry=3)
groups=list(X1=c("X1"),
            X2=c("X2"),
            …
            X217=c("X217"))
group.importance(rf.obj, groups)
```

Lampiran 9. *Syntax Naïve Bayes Classifier*

```
#import library
library(MXM)
library(e1071)
library(caret)
library(rminer)

#import data
data=read.csv("D:/DATA/Kuliah/! Tugas Akhir/Hasil Olahan
Data/Data Protection Radiation 100%.csv")
data$Y <- as.factor(data$Y)

#NBC
a=generatefolds(data$Y, nfolds=10, stratified=TRUE, seed=12345)
TotalAccuracyTrain=rep(0,10)
SensTrain=rep(0,10)
SpesTrain=rep(0,10)
TotalAccuracyTest=rep(0,10)
SensTest=rep(0,10)
SpesTest=rep(0,10)
for(i in 1:10)
{
  train=data[-a[[i]],]
  test=data[a[[i]],]
  model = naiveBayes(Y~., data=train)
  predtrain=predict(model, train)
  predtest=predict(model, test)
  tabel1=table(train$Y, predtrain)
  tabel2=table(test$Y, predtest)
```

Lampiran 9. Syntax Naïve Bayes Classifier (Lanjutan)

```
  TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
  SensTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SpesTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))

  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  SensTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
  SpesTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))
}

mean(TotalAccuracyTrain)
mean(SensTrain)
mean(SpesTrain)
mean(TotalAccuracyTest)
mean(SensTest)
mean(SpesTest)
hasil=data.frame(TotalAccuracyTrain, SensTrain, SpesTrain,
TotalAccuracyTest, SensTest, SpesTest)
```

Lampiran 10. *Syntax* CART

```
#import library
library(MXM)
library(e1071)
library(caret)
library(rpart)
library(rminer)

#import data
data=read.csv("D:/DATA/Kuliah/! Tugas Akhir/Hasil Olahan
Data/Data Protection Radiation 100%.csv")
data$Y <- as.factor(data$Y)

# Fit the model on the training set
set.seed(12345)
model1 <- train(
  Y ~., data = data, method = "rpart",
  trControl = trainControl("cv", number = 10),
  tuneLength = 10
)
model1
plot(model1)

#cart
a=generatefolds(data$Y, nfolds=10, stratified=TRUE, seed=12345)
TotalAccuracyTrain=rep(0,10)
SensTrain=rep(0,10)
SpesTrain=rep(0,10)
TotalAccuracyTest=rep(0,10)
SensTest=rep(0,10)
SpesTest=rep(0,10)
for(i in 1:10)
```

Lampiran 10. *Syntax* CART (Lanjutan)

```
{
  train=data[-a[[i]],]
  test=data[a[[i]],]
  model = rpart(Y~., data=train, method = "class", cp=0.01994302)
  predtrain=predict(model, train, type = "class")
  predtest=predict(model, test, type = "class")
  tabel1=table(train$Y, predtrain)
  tabel2=table(test$Y, predtest)

  TotalAccuracyTrain[i]=((tabel1[1,1]+tabel1[2,2])/sum(tabel1))
  SensTrain[i]=((tabel1[1,1])/(tabel1[1,1]+tabel1[1,2]))
  SpesTrain[i]=((tabel1[2,2])/(tabel1[2,1]+tabel1[2,2]))

  TotalAccuracyTest[i]=((tabel2[1,1]+tabel2[2,2])/sum(tabel2))
  SensTest[i]=((tabel2[1,1])/(tabel2[1,1]+tabel2[1,2]))
  SpesTest[i]=((tabel2[2,2])/(tabel2[2,1]+tabel2[2,2]))
}
mean(TotalAccuracyTrain)
mean(SensTrain)
mean(SpesTrain)
mean(TotalAccuracyTest)
mean(SensTest)
mean(SpesTest)

# Decision rules in the model
model

# Plot the tree model
par(xpd = NA)
plot(model)
text(model,  digits = 3)
```

Lampiran 11. *Syntax Mixture Distribution*

```
#import library
library(mixtools)

#import data
data=read.csv("D:/DATA/Kuliah/! Tugas Akhir/Hasil Olahan
Data/Tingkat Kematian Sel.csv")

#model
set.seed(12345)
wait = data$R
mixmdl = normalmixEM(wait, k=4)
summary(mixmdl)
mixmdl
mixmdl$lambda
mixmdl$mu
mixmdl$sigma
mixmdl$posterior

#plot
plot(mixmdl,which=2)
lines(density(wait), lty=2, lwd=2)
```

Lampiran 12. Surat Keterangan Pengambilan Data

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, mahasiswa Departemen Statistika FMKSD ITS,

Nama : Rizky Mubarok
NRP : 062114 4000 0074

menyatakan bahwa data yang digunakan dalam Tugas Akhir ini merupakan data sekunder yang diambil dari penelitian yaitu :

Sumber : Data Penelitian Ariyasu, *et al.,* (2014) dengan Judul "*Design and Synthesis of 8-Hydroxyquinoline-Based Radioprotective Agents*"

Keterangan : Data proteksi radiasi dengan 84 observasi (senyawa), dua variabel respon yaitu tingkat kematian sel kanker dan kelas proteksi radiasi, serta 217 variabel prediktor yang merupakan penyusun senyawa.

Surat pernyataan ini dibuat dengan sebenarnya. Apabila terdapat pemalsuan data maka saya siap menerima sanksi sesuai aturan yang berlaku.

Mengetahui,	Surabaya, Juli 2018
Pembimbing Tugas Akhir	Mahasiswa
Dr. rer. pol. Heri Kuswanto, M.Si.	Rizky Mubarok
NIP. 19820326 200312 1 004	NRP.062114 4000 0074

(*Halaman ini sengaja dikosongkan*)

BIODATA PENULIS

Penulis dengan nama lengkap Rizky Mubarok dilahirkan di Kabupaten Tuban pada 23 Juni 1996. Penulis menempuh pendidikan formal di SDN Mojoagung, SMPN 2 Bojonegoro, dan SMAN 10 Malang (Sampoerna Academy). Kemudian penulis diterima sebagai Mahasiswa Departemen Statistika ITS melalui jalur SBMPTN pada tahun 2014. Selama masa perkuliahan, penulis aktif di organisasi Himpunan Mahasiswa Statistika ITS (HIMASTA-ITS) sebagai staff Departemen Pengembangan Sumber Daya Mahasiswa (PSDM) periode 2015/2016 dan Ketua HIMASTA-ITS periode 2016/2017. Penulis juga aktif sebagai staff Kementrian Komunikasi dan Informasi BEM ITS periode 2015/2016. Selain itu, penulis juga berkesempatan dalam mengikuti pelatihan manajerial yaitu LKMM tingkat dasar, tingkat menengah, dan tingkat lanjut. Di bidang akademik, penulis diberi kesempatan menjadi semifinalis pada Pekan Analisis Statistik tahun 2017 yang diselenggarakan oleh Universitas Mulawarman dan Kompetisi Statistika Nasioal tahun 2017 yang diselenggarakan oleh Institut Pertanian Bogor. Penulis juga pernah mengikuti kegiatan ITS *Goes Global* Singapura tahun 2017 yang diselenggarakan oleh ITS *International Office*. Apabila pembaca ingin memberi kritik dan saran serta diskusi lebih lanjut mengenai Tugas Akhir ini, dapat menghubungi penulis melalui email mubarokrizky06@gmail.com atau melalui nomor telepon 082140796868.