PREDIKSI DETEKSI PENYAKIT KANKER PAYUDARA DENGAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

Oleh:
AYU DIAN FITRI MELLINA
NIM. 17650087

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

PREDIKSI DETEKSI PENYAKIT KANKER PAYUDARA DENGAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

Oleh:
AYU DIAN FITRI MELLINA
NIM. 17650087

Diajukan kepada:
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

HALAMAN PERSETUJUAN

PREDIKSI DETEKSI PENYAKIT KANKER PAYUDARA DENGAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

Oleh:
AYU DIAN FITRI MELLINA
NIM. 17650087

Telah Diperiksa dan Disetujui untuk Diuji
Tanggal: 17 Mei 2023

Pembimbing I	Pembimbing II
Prof. Dr. Suhartono, M.Kom	Dr. M. Ainul Yaqin, M.Kom
NIP. 19680519 200312 1 001	NIP. 19761013 200604 1 004

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

PREDIKSI DETEKSI PENYAKIT KANKER PAYUDARA DENGAN MENGGUNAKAN ALGORITMA DECISION TREE

SKRIPSI

Oleh:
AYU DIAN FITRI MELLINA
NIM. 17650087

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S. Kom)
Pada Tanggal: 5 Juni 2023

Susunan Dewan Penguji

Ketua Penguji	: Dr. Muhammad Faisal, M.T NIP. 19740510 200501 1 007	( )
Anggota Penguji I	: Syahiduz Zaman, M.Kom NIP. 19700502 200501 1 005	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. Ainul Yaqin, M.Kom NIP. 19761013 200604 1 004	( )

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Ayu Dian Fitri Mellina

NIM : 17650087

Program Studi : Teknik Informatika

Fakultas :Sains dan Teknologi

Judul Skripsi : Prediksi Keputusan Deteksi Penyakit Kanker Payudara dengan Menggunakan Algoritma Decision Tree.

Menyatakan dengan sebenarnya bahwa skripsi yang saya tulis ini benar – benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang,17 Mei 2023

Yang membuat pernyataan,

Ayu Dian Fitri Mellina

NIM. 17650087

MOTTO

“Dan tidak ada kesuksesan bagiku melainkan atas pertolongan Allah”

(Q.S. Huud:88)

HALAMAN PERSEMBAHAN

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Penulis persembahkan skripsi ini kepada keluarga penulis, terutama untuk ayah Sahirah dan ibu Hartini yang telah banyak memberikan pelajaran kehidupan kepada penulis lewat kerja keras, kesabaran, dan kesederhanaan mereka. Kepada keluarga besar penulis dari pihak Ayah maupun pihak Ibu yang senantiasa ikut mendoakan agar cepat menyelesaikan masa studi. Teruntuk om saya Mulyadi, yang turut membantu dan mendukung penulis secara finansial. Terakhir untuk Abdullah Muhammad Kuddah, Sabrina Gabiella, Nuristiana Izzatul, Halimatus, yang selalu menjadi pendengar yang baik, menjadi tempat keluh kesah, bertukar pikiran, dan menjadi salah satu *support system* penulis. Semoga kasih sayang Allah *subhanahu wa ta'ala* selalu menyertai mereka.

KATA PENGANTAR

Assalamu'alaikum warahmatullahi wabarakatuh

Segala puji bagi Allah *subhanahu wa ta'ala* yang telah melimpahkan rahmat dan karunia-Nya serta shalawat beriring salam tak lupa dihanturkan kepada baginda Rasulullah *shalallahu 'alaihi wa sallam* sehingga penulis mampu merampungkan penulisan skripsi yang berjudul **"Prediksi Deteksi Penyakit Kanker Payudara dengan Menggunakan Algoritma Decision Tree"** sebagai salah satu syarat kelulusan untuk mendapatkan gelar sarjana pada Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Skripsi ini tidak dapat terwujud tanpa adanya doa, bantuan, bimbingan dan motivasi dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih sedalam-dalamnya kepada:

1. Prof. Dr. H. M. Zainuddin, MA, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Sri Harini, M. Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, ST., M.MT., IPM selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Juniardi Nur Fadila, M.T., selaku Dosen Wali yang telah bersedia meluangkan waktu dalam membimbing, dan memberikan motivasi sehingga skripsi ini dapat terselesaikan.
5. Prof. Dr. Suhartono, M,Kom selaku Dosen Pembimbing I yang telah banyak bersedia meluangkan waktunya dalam membimbing, memberi saran dan arahan kepada penulis sehingga skripsi ini dapat terselesaikan.

6. Dr. M. Ainul Yaqin, M,Kom selaku Dosen Pembimbing II yang juga bersedia meluangkan waktunya dalam membimbing dan memberi arahan kepada penulis sehingga skripsi ini dapat terselesaikan.
7. Dr. Muhammad Faisal, M.T., M.Kom dan Syahiduz Zaman, M.Kom selaku Dosen Penguji yang telah memberikan kritik dan masukan membangun kepada penulis selama proses penyelesaian skripsi ini.
8. Seluruh Dosen dan Jajaran Staf Program Studi Teknik Informatika yang telah mengajarkan ilmu yang bermanfaat kepada penulis.
9. Teman-teman Teknik Informatika Angkatan 2017 UNOCORE, khususnya Rafika Syahranita, Jayanti Galuh, Shinta Rizky, Aldy Destra, Hamdan, Fahim Fikri, Ramadhana fardian, Ardisca Evanandy, M.Syahiruddin yang telah menemani dan saling bertukar pikiran saat mengerjakan skripsi, serta memberikan banyak pengalaman dan dukungan berharga.

Penulis menyadari bahwa dalam pengerjaan skripsi ini masih terdapat banyak kekurangan sehingga penulis terbuka terhadap kritik dan saran yang membangun dari para pembaca. Penulis berharap semoga skripsi ini dapat memberikan manfaat tidak hanya bagi penulis namun juga bagi para pembaca.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Malang, 17 Mei 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Mellina, Ayu Dian Fitri. 2023. **Prediksi Deteksi Penyakit Kanker Payudara dengan Menggunakan Metode Decision Tree.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. M. Ainul Yaqin, M.Kom

---

Kata kunci: Kanker Payudara, Klasifikasi, Prediksi, Decision Tree, *Machine Learning*

Kanker merupakan salah satu penyakit pembunuh terbesar di dunia dan sulit untuk disembuhkan. Deteksi dini kanker dapat dilakukan melalui serangkaian uji laboratorium yang dapat mengidentifikasi kanker jinak atau ganas. Kanker payudara termasuk dalam jenis kanker ganas dan memiliki gejala awal berupa benjolan yang semakin membesar. Data mining, khususnya metode klasifikasi, dapat digunakan untuk menganalisis data uji laboratorium dan mengkategorikan kanker payudara menjadi jinak dan ganas. Decision tree adalah metode klasifikasi yang digunakan dalam penelitian ini, dengan algoritma *iterative dichotomiser-3* (ID3) dan C5.0 sebagai pilihan untuk deteksi kanker payudara. Data yang digunakan pada penelitian ini merupakan *Breast Cancer Coimbra Dataset* yang dapat diunduh secara gratis di website resmi UCI *Machine Learning* tahun 2018. *iterative dichotomiser-3* (ID3) memiliki keterbatasan dalam menangani data tidak terstruktur dan atribut kontinu, sementara C5.0 merupakan pengembangan dari *iterative dichotomiser-3* (ID3) yang lebih baik dalam menangani jenis data tersebut. Kedua algoritma ini menghasilkan model pohon yang berbeda dengan tingkat keakuratan yang bervariasi. Hasil dari penelitian ini menunjukan bahwa algoritma C5.0 memperoleh hasil klasifikasi terbaik dibandingkan dengan algoritma *iterative dichotomiser-3* (ID3) dengan nilai *accuracy* sebesar 80%, *precision* sebesar 84,2%, *recall* sebesar 80%, dan F1 *score* sebesar 80%. Nilai *accuracy* sebesar 80% menyatakan bahwa sistem dapat melakukan klasifikasi dengan baik sehingga model algoritma C5.0 dapat diterima dan dapat digunakan untuk memprediksi deteksi penyakit kanker payudara.

ABSTRACT

Mellina, Ayu Dian Fitri. 2023. **Prediction Detection of Breast Cancer using a decision tree algorithm.** Undergraduate Thesis. Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang. Supervisor: (I) Prof, Dr, Suhartono, M.Kom (II) Dr. M. Ainul Yaqin, M.Kom

---

Cancer is one of the biggest killers in the world and difficult to cure. Early detection of cancer can be done through a series of laboratory tests that can identify benign or malignant cancer. Breast cancer is a type of malignant cancer and has early symptoms in the form of a growing lump. Data mining, especially the classification method, can be used to analyze laboratory test data and categorize breast cancer into benign and malignant. Decision tree is the classification method used in this study, with iterative dichotomizer-3 (ID3) and C5.0 algorithms as the choice for breast cancer detection. The data used in this study is the Breast Cancer Coimbra Dataset which can be downloaded for free on the official UCI Machine Learning website in 2018. Iterative dichotomizer-3 (ID3) has limitations in handling unstructured data and continuous attributes, while C5.0 is a development of iterative dichotomiser-3 (ID3) which is better at handling this type of data. These two algorithms produce different tree models with varying degrees of accuracy. The results of this study indicate that the C5.0 algorithm obtains the best classification results compared to the iterative dichotomous-3 (ID3) algorithm with an accuracy value of 80%, a precision of 84.2%, a recall of 80%, and an F1 score of 80%. An accuracy value of 80% indicates that the system can perform the classification properly so that the C5.0 algorithm model is acceptable and can be used to predict breast cancer detection.

Keywords: *Breast Cancer, Classification, Prediction, Machine Learning*

مستخلص البحث

ميلينا، أيو ديان فيتري. 2023. التنبؤ بالكشف عن سرطان الثدي باستخدام طريقة شجرة القرار. أُطرُوحَة. برنامج دراسة هندسة المعلوماتية ، كلية العلوم والتكنولوجيا ، جامعة الولاية الإسلامية مولانا مالك إبراهيم مالانج. المشرفون: (1) الأستاذ الدكتور سوهارتون محمد كوم (2) د. عين اليقين ، م. كوم

الكلمات المفتاحية: سرطان الثدي ، التصنيف ، التنبؤ ، شجرة القرار ، التعلم الآلي

السرطان هو أحد أكبر الأمراض الفتاكة في العالم ويصعب علاجه. يمكن الكشف المبكر عن السرطان من خلال سلسلة من الاختبارات المعملية التي يمكن أن تحدد السرطان الحميد أو الخبيث. سرطان الثدي هو نوع من السرطانات الخبيثة وله أعراض مبكرة على شكل كتلة متنامية. يمكن استخدام التنقيب عن البيانات ، وخاصة طريقة التصنيف ، لتحليل بيانات الاختبارات المعملية وتصنيف سرطان الثدي إلى حميدة وخبيثة. شجرة القرار هي طريقة التصنيف المستخدمة في هذه الدراسة ، مع خوارزميات ثنائية التكرار 3 (ID3)و C5.0 كخيار للكشف عن سرطان الثدي. البيانات المستخدمة في هذه الدراسة هي مجموعة بيانات Coimbraلسرطان الثدي والتي يمكن تنزيلها مجانًا على موقع UCI Machine Learning الرسمي في عام 2018. يوجد قيود في التعامل مع البيانات غير المنظمة والسمات المستمرة ، بينما C5.0 هو تطوير dichotomiserالتكراري 3 (ID3) وهو أفضل في التعامل مع هذا النوع من البيانات. تنتج هاتان الخوارزميتان نماذج شجرية مختلفة بدرجات متفاوتة من الدقة. تشير نتائج هذه الدراسة إلى أن خوارزمية C5.0تحصل على أفضل نتائج التصنيف مقارنة بالخوارزمية التكرارية ثنائية التفرع -3 (ID3) بقيمة دقة 80٪ ودقة 84.2٪ واسترجاع 80٪ و درجة F1 بنسبة 80٪. تشير قيمة الدقة البالغة 80٪ إلى أن النظام يمكنه إجراء التصنيف بشكل صحيح بحيث يكون نموذج خوارزمية C5.0 مقبولًا ويمكن استخدامه للتنبؤ باكتشاف سرطان الثدي

BAB I

PENDAHULUAN

1.1 Latar Belakang

Salah satu penyakit dengan kategori 10 besar penyakit pembunuh di dunia adalah kanker (Imaduddin, Hermansyah, & Salsabilla B, 2021). Hal tersebut merupakan pernyataan dari *World Health Organization* (WHO) karena kanker menyerang jaringan sel lainnya yang berada di dalam tubuh, dimana sel-sel tersebut bersifat abnormal yang kemudian membelah diri tanpa kontrol. Kanker menjadi penyakit yang mudah menyebar dalam tubuh dan menjadi penyakit yang sulit untuk disembuhkan. Salah satu cara yang dapat digunakan untuk mendeteksi penyakit kanker adalah dengan melakukan serangkaian uji laboratorium. Hasil dari uji laboratorium tersebut dapat dijadikan sebagai acuan apakah seseorang tersebut mengidap penyakit kanker dengan kategori jinak atau ganas (Kemenkes, 2019).

Penyakit kanker memiliki berbagai macam jenis, salah satunya adalah kanker payudara. Dalam Islam, penjelasan mengenai penyakit yang diturunkan pada manusia ini dijelaskan pada Al-Quran surah Yunus ayat 57.

يَٰٓأَيُّهَا ٱلنَّاسُ قَدْ جَآءَتْكُم مَّوْعِظَةٌ مِّن رَّبِّكُمْ وَشِفَآءٌ لِّمَا فِى ٱلصُّدُورِ وَهُدًى وَرَحْمَةٌ لِّلْمُؤْمِنِ

"Hai manusia, sesungguhnya telah datang kepadamu pelajaran dari Tuhanmu dan penyembuh bagi penyakit-penyakit (yang berada) dalam dada dan petunjuk serta rahmat bagi orang-orang yang beriman."(Q.S Yunus .57).

Pada ayat tersebut dijelaskan bahwasanya Allah SWT memberikan pelajaran yang berguna kepada manusia dan menyembuhkan penyakit-penyakit dalam hati,

serta menjadi petunjuk dan rahmat bagi orang yang beriman. Dalam konteks kanker payudara, ayat ini dapat memberikan penghiburan serta harapan bagi penderita penyakit kanker payudara. Ayat ini juga mengingatkan bahwasaynya Allah SWT adalah sumber penyembuhan dan rahmat yang tak terbatas. Sebagau umat yang beriman kita diperintahkan untuk percaya bahwasanya dalam pelajaran yang Allah berikan, terdapat pula jalan menuju kesembuhan. Ayat ini juga mengingatkan mengenai pentingnya beriman dan memiliki hubungan yang kuat dengan Tuhan. Iman serta keayakinan yang kuat akan membawa penghiburan dan ketenangan di tengah cobaan, termasuk dalam menghadapi penyakit seperti kanker payudara. Dalam kekuatan iman dan pengharapan kepada Allah SWT, kita dapat menemukan penyembuhan dan bimbingan-Nya. Dalam konteks medis, tentunya pengobatan dan perawatan yang tepat juga penting. Namun, ayat ini mengajarkan bahwa penyembuhan sejati berasal dari Allah SWT. Oleh karena itu, kita dianjurkan untuk mencari pengobatan medis yang tersedia dan juga memperkuat iman kita kepada Allah dalam proses kesembuhan.

Kanker payudara sendiri termasuk pada golongan penyakit kanker ganas yang mana kasusnya banyak dijumpai di kalangan wanita. Penyakit ini dapat menyerang wanita pada usia berapapun, tetapi risiko meningkat dengan bertambahnya usia. Penyakit ini juga dapat menyerang pria, meskipun hal ini sangat jarang terjadi. Jumlah kasus penyakit kanker payudara di seluruh dunia semakin bertambah setiap tahunnya. Salah satu gejala awal pada kanker payudara adalah munculnya benjolan kecil yang semakin lama akan bertambah semakin besar. Perubahan atau mutasi pada DNA sel payudara merupakan penyebab awal

timbulnya kanker payudara. Mutasi gen biasanya terjadi karena diwariskan dari generasi sebelumnya, akan tetapi mutasi ini dapat juga terjadi tanpa penyebab yang pasti. Perempuan dengan resiko terkena kanker lebih besar adalah perempuan yang mengalami siklus menstruasi lebih banyak daripada perempuan normal lainnya, terlambat *menopause*, serta menstruasi dini. Peningkatan jumlah kasus kanker payudara di seluruh dunia juga disebabkan oleh perubahan pola gaya hidup yang tidak sehat serta kurangnya aktivitas fisik (Musa & Aliyu, 2018).

Terdapat beberapa atribut yang menjadi acuan dalam mendeteksi jenis kanker pada penyakit kanker payudara. Atribut tersebut membentuk sebuah pola yang kemudian dikategorikan sesuai dengan kelas yang sudah ada. Dalam menentukan jenis kanker payudara yang dialami oleh pasien di rumah sakit, pihak laboratorium rumah sakit tentunya membutuhkan waktu yang cukup lama untuk menganalisis hasil diagnosa mengenai jenis kanker tersebut. Hasil data laboratorium tidak sepenuhnya memberikan hasil yang konkrit, tentunya diperlukan hipotesis yang dapat memperkuat hasil laboratorium tersebut. Hipotesis disini bertujuan agar dapat membantu dokter menentukan jenis kanker pasien sehingga dapat ditangani dengan segera. Terdapat banyak cara atau metode dalam menentukan pola pada data mengenai penyakit kanker, salah satu cara tersebut adalah menggunakan metode *data mining. Data mining* merupakan sebuah proses yang bertujuan untuk menemukan pola tersembunyi agar dapat menghasilkan sebuah informasi yang mudah dicerna oleh manusia. Pada proses *data mining* terdapat beberapa metode di dalamnya, salah satunya adalah metode prediktif yang memiliki teknik yang dapat digunakan, yakni regresi dan klasifikasi

(Sunjana, 2010). Dalam penelitian ini, proses klasifikasi dapat diterapkan untuk mengolah hasil data uji laboratorium dengan mengkategorikannya menjadi dua kategori, yakni kategori jinak (*benign*) dan ganas (*malignant*).

Klasifikasi merupakan sebuah proses yang bertujuan untuk mengelompokkan sebuah data menjadi beberapa kelompok sesuai dengan kategori yang sudah ditetapkan. Beberapa contoh dari teknik klasifikasi pada *data mining* adalah *K-Nearest Neighbor, Naive Bayes, Rules Base* dan *Decision Tree.* Metode yang diterapkan pada penelitian ini adalah *decision tree. Decision tree* merupakan sebuah metode sistem prediksi yang strukturnya menyerupai pohon bercabang atau biasa juga disebut dengan struktur hierarki sehingga metode ini cocok untuk diterapkan pada permasalah penelitian ini yang di dalamnya menggambarkan sebuah persoalan dan mencari atau membutuhkan sebuah solusi dari persoalan tersebut (Wahyudin, 2009). Kedua algoritma *iterative dichotomiser-3* (ID3) dan C5.0 merupakan salah satu algoritma decision tree yang dapat digunakan untuk tujuan tersebut.

Kedua algoritma *iterative dichotomiser-3* (ID3) dan C5.0 merupakan algoritma decision tree yang dapat digunakan untuk deteksi penyakit kanker payudara. Algoritma *iterative dichotomiser-3* (ID3) merupakan algoritma yang pertama kali dikembangkan oleh *decision tree* yang memiliki kemampuan yang cukup baik dalam menangani data yang terstruktur. Namun, algoritma ini tidak dapat menangani data yang tidak terstruktur dan tidak dapat menangani atribut yang bernilai kontinu. Sedangkan, C5.0 merupakan pengembangan dari algoritma *iterative dichotomiser-3* (ID3) yang memiliki kemampuan yang lebih baik dalam

menangani data yang tidak terstruktur dan dapat menangani atribut yang bernilai kontinu. Algoritma ini juga memiliki kemampuan untuk membangun model yang lebih akurat dan memiliki fitur pruning yang dapat membantu mengurangi overfitting pada model. Kedua algoritma tersebut dapat menghasilkan model pohon (*tree*) yang berbeda dengan *dataset* yang sama. Model yang dihasilkan dari kedua algoritma tersebut tentunya memilki tingkat keakuratan yang berbeda.

Latar belakang penelitian ini adalah untuk mengembangkan suatu sistem prediksi menggunakan metode *decision tree* serta membandingkan model yang dihasilkan oleh algoritma yang ada pada *decision tree* yakni *iterative dichotomiser-3* (ID3) dan C5.0 untuk deteksi penyakit kanker payudara. Sistem ini diharapkan dapat membantu dokter dalam mengambil keputusan yang lebih akurat dan tepat waktu dalam menegakkan diagnosis kanker payudara. Dengan demikian, diharapkan dapat mengurangi risiko kematian akibat kanker payudara dan meningkatkan tingkat kesembuhan pasien. Berdasarkan latar belakang yang telah dijabarkan, maka peneliti mengajukan penelitian dengan judul "Prediksi Deteksi Penyakit Kanker Payudara dengan Menggunakan Algoritma *Decision Tree".*

1.2 Pernyataan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka pernyataan masalah yang dirumuskan dalam penelitian ini adalah bagaimana cara memantau penyakit kanker payudara dengan menggunakan algoritma *decision tree*?

1.3 Batasan Masalah

Agar penelitian ini terhindar dari kegiatan diluar sasaran dan untuk memudahkan pekerjaan, maka ditetapkan batasan-batasan masalah yakni sebagai berikut:

1. Data pada penelitian ini merupakan *dataset breast cancer coimbra* yang didapat dari *UCI machine learning* tahun 2018. Data berjumlah 116, yang mana pada data tersebut terdapat 1 atribut ID serta 10 atribut utama yaitu age, BMI, glukosa, insulin, HOMA, leptin, adiponectin, resistin, MCP.1, classification.
2. Hasil dari proses penelitian ini nantinya akan mengklasifikasikan jenis penyakit kanker menjadi dua kategori, yakni kanker jinak (*benign*) atau kanker ganas (*malignant*).
3. Sistem aplikasi pada penelitian ini dibangun dengan menggunakan Rstudio.

1.4 Tujuan Penelitian

Berdasarkan pernyataan masalah yang sudah dijabarkan, tujuan yang ingin dicapai dalam penelitian ini adalah mengklasifikasikan penyakit kanker payudara menggunakan algoritma *decision tree* untuk pemantauan dalam proses pengobatan.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Sebagai sarana bagi masyarakat untuk mencegah penyakit kanker sejak dini.

2. Sebagai sarana untuk membuktikan hipotesis klasifikasi awal penyakit kanker payudara.
3. Mengkategorikan penyakit kanker payudara sesuai dengan kelas yang sudah ada.

1.6 Sistematika Penulisan

Dalam menyusun laporan penelitian ini sistematika penulisan yang digunakan yakni sebagai berikut:

Bab I Pendahuluan

Bab ini berisi beberapa sub bab yang menguraikan latar belakang masalah, pernyataan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, serta sistematika penulisan.

Bab II Tinjauan Pustaka

Bab ini berisi tentang penjabaran penelitian-penelitian terdahulu sebagai pembeda dari penelitian yang diangkat oleh penulis. Serta, berisi tentang konsep serta teori dari berbagai sumber yang berkaitan dengan pembahasan dalam penelitian ini.

Bab III Desain dan Implementasi Sistem

Bab ini berisi tentang pemaparan dari perancangan desain sistem penelitian, serta uraian mengenai langkah-langkah penelitian, yang didalamnya memuat sumber data, jenis data, pengolahan serta analisis data.

Bab IV Hasil dan Pembahasan

Bab ini berisi hasil dan implementasi sistem yang telah dibuat. Serta pengujian yang telah dilakukan sehingga dapat ditarik kesimpulan.

Bab V Penutup

Bab ini berisi tentang kesimpulan yang mana kesimpulan tersebut didapat dari hasil implementasi sistem yang dibuat, serta beberapa saran yang bertujuan untuk pengembangan penelitian di masa mendatang.

BAB II

TINJAUAN PUSTAKA

Bab ini berisi tentang penjabaran hal-hal yang berkaitan dengan tinjauan pustaka, penelitian terdahulu, serta dasar teori yang digunakan sebagai pendukung dalam penelitian ini. Hal tersebut didapat dari berbagai sumber literatur yang berkaitan dengan pokok pembahasan dalam penelitian ini.

2.1 Penelitian Terdahulu

Penelitian mengenai "*Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Networks Algorithms*" yang dilakukan oleh (Higa, 2018) bertujuan untuk mendiagnosis dan membuat prognosis mengenai kanker payudara yang mana diagnosis tersebut dikategorikan menjadi dua bagian yakni jinak dan ganas. Sementara prognosis kanker payudara pada penelitian tersebut memprediksi kapan kanker payudara kemungkinan akan kambuh pada pasien yang pernah menderita kanker. Kedua algoritma berhasil mengklasifikasikan dengan benar lebih dari 92% kasus dalam 10 percobaan. Namun, algoritma Neural Network memiliki rata-rata tingkat akurasi prediksi yang lebih baik (tingkat klasifikasi yang benar adalah 95,9%). Hasil penelitian tersebut menunjukkan bahwa kedua algoritma tersebut memiliki hasil akurasi dengan prediksi keseluruhan masing-masing 94% dan 95,4%.

Penelitian lainnya juga pernah dilakukan oleh (Musa & Aliyu, 2018) yakni mengenai penerapan algoritma *machine learning* dengan menggunakan *decision tree* serta statistik deskriptif untuk mengevaluasi kinerja model dalam

memprediksi kemungkinan kanker payudara metastasis. Dataset yang digunakan memiliki 259 instances dan 10 atribut. Klasifikasi pada penelitian tersebut menggunakan *decision tree* pada perangkat lunak IMB SPSS (versi 23), dengan mengkategorikan penelitian tersebut menjadi 2 kelas yakni kelas 0 = tidak bermetastasis, kelas 1 = bermetastasis. Hasil dari penelitian tersebut menunjukan bahwa 259 kasus kanker payudara, 218 (84,2%) kasus tidak bermetastasis, sedangkan 41 (15,8%) bermetastasis ke bagian tubuh lainnya. Tingkat akurasi yang didapat pada keseluruhan model adalah 87% dengan sensitivitas 88%, spesifitas 75% dan presisi 98%. Berdasarkan hasil tersebut penerapan algoritma *decision tree* memprediksikan bahwa 87% tumor muncul pada stadium IV, yang berarti bahwa tumor tersebut dapat menyebar ke bagian tubuh lainnya.

Penelitian selanjutnya mengenai "*Breast Cancer Classification using Decision Tree Algorithms*" oleh (Tarawneh et al., 2022). Peneliti pada penelitian tersebut menguji tingkat akurasi pada penerapan metode *decision tree* dalam mengklasifikasikan jenis kanker sesuai dengan kategori jinak dan ganas berdasarkan karakteristik yang berkaitan dengan rekam medis pasien. Pengujian tersebut dilakukan dengan menggunakan beberapa metrik akurasi seperti F-measure, ROC Area, Precision, Recall, TP rate, and FP rate. Data yang digunakan pada penelitian tersebut diambil dari arsip Kaggle, studi utama pada penelitian tersebut menggunakan 10 sampel kanker payudara, sedangkan pada studi lanjutan data yang digunakan sebanyak 286 sampel kanker payudara dari kelompok data yang sama. Hasil dari penelitian tersebut menunjukan bahwa penerapan metode *decision tree* dapat mewakili diagnosis kanker payudara, serta dapat menjadi

sarana atau strategi tambahan yang dapat digunakan dalam pengobatan medis. Nilai akurasi yang didapat pada metode *decision tree* pada percobaan pertama adalah 100%, sedangkan pada penyelidikan lanjutan adalah 97,9%.

Penelitian yang dilakukan oleh (Imaduddin et al., 2021) mengenai perbandingan metode *decision tree* dengan *support vector machine* untuk mengklasifikasikan kanker payudara, kedua metode tersebut digunakan untuk mengetahui metode *machine learning* mana yang memiliki performa terbaik. Pada penelitian tersebut terdapat proses seleksi fitur dimana hal tersebut bertujuan untuk memilih dan memfiltrasi fitur apa saja yang memiliki pengaruh serta dampak yang besar terhadap proses klasifikasi, proses ini berpengaruh terhadap perolehan hasil klasifikasi yang lebih baik. Hasil perbandingan dari penelitian tersebut menunjukkan bahwa penerapan algoritma *support vector machine* memiliki performa terbaik dengan hasil akurasi yang didapat sebesar 87,5%, sensitivitas 90%, dan spesifitas 85%. Sedangkan hasil terburuk diperoleh oleh skema *support vector machine* tanpa dilengkapi dengan fitur seleksi dengan hasil akurasi yang didapat sebesar 70%, sensitivitas 64%, serta spesifitas sebesar 80%.

(Keerthika, Sruthi, Swathi, Swetha, & Vinupriya, 2021) dalam penelitiannya yang berjudul "*Diagnosis of Breast Cancer using Decision Tree Data Mining Technique*" dimana pada penelitian tersebut menerapkan teknik *data mining* untuk mendiagnosa penyakit kanker payudara, serta menentukan faktor yang mempengaruhi resiko terjadinya penyakit kanker payudara yang kemudian mengidentifikasi ke dalam dua kategori yakni kanker yang berisiko rendah atau tinggi. Strategi berlapis campuran khusus dari pohon keputusan (*decision tree*)

serta klaster diterapkan untuk membuat skema penilaian risiko kanker yang paling banyak. Penelitian ini menggunakan pengujian statistik seperti grouping, clustering dan estimasi untuk mengetahui aptitude pasien kanker. Sistem yang dibuat pada penelitian ini dilengkapi dilengkapi dengan pengetahuan klinis pasien sebelumnya. Tujuan dari model ini adalah untuk melindungi pasien serta dapat dijangkau oleh penggunanya dan model prediktif pada sistem ini akan membantu dalam pencegahan dan diagnosis awal pada kanker payudara. Hasil dari penelitian tersebut dalam hal membedakan antara massa jinak dan ganas, SVM mendapat peringkat tertinggi sebesar 97% untuk akurasi sistem, serta untuk tingkat klasifikasi sebesar 97%.

Selanjutnya, dilakukan perbandingan penelitian dengan beberapa penelitian terdahulu. Perbandingan penelitian dari segi metode yang digunakan dapat dijelaskan melalui tiga tahapan yakni prefix, metode utama, dan post. Pada penelitian ini tiga tahapan tesebut sebagai berikut:

1. Prefix

   Tahap prefix pada penelitian ini menggunakan pendekatan yang sama yaitu *decision tree.*

2. Metode utama

   Penelitian ini membangun sebuah model menggunakan algoritma *decision tree* berdasarkan dataset yang sudah ada. Dengan menerapkakan dua metode pada algoritma *decision tree* yakni *iterative dichotomiser-3* (ID3) untuk membangun pohon keputusan. Sedangkan metode C5.0 digunakan untuk menghasilkan pohon keputusan yang lebih optimal.

3. Post

   Menggunakan metode testing terpisah (*holdout testing*) yakni dengan cara membagi dataset menjadi dua bagian yakni data pelatihan (*training* data) dan data pengujian (*testing* data) untuk menguji performa dari dua metode *decision tree* tersebut.

Tabel 2.1 Matriks Jurnal Penelitian Terdahulu

No	Riset	Metode		
		Prefix	Utama	Post
1.	*Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models* (Naveen, Sharma, & Ramachandran Nair, 2019).	Penskalaan fitur (*feature scaling*), *cross validation*, dan teknik *bagging*.	Decision Tree, SVM, KNN, Multilayer Perceptron, Logistics Regression, Random Forest.	Dataset dibagi menjadi *training* dan *testing* dengan rasio 90:10. Evaluasi prediksi dengan menggunakan *confusion matrix* dan *classification report.*
2.	*Comparison of Machine Learning Algorithms in Breast Cancer Prediction using the Coimbra Dataset* (Austria et al., 2019)	Menerapkan *unique hyper-parameter* pada setiap klasifikasi untuk melakukan prediksikinerja.	KNN, Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, Naive-Bayes.	Pembagian data yang digunakan dibagi menjadi dua, dengan ratio 70:30, yakni 70% data *training* dan 30% data *testing.*
3.	*Comparison of Classification Models for Early Prediction of Breast Cancer* (Ghani, Alam, & Jaskani, 2019).	Menerapkan *recursive feature elimination.*	Naïve Bayes, Decision Tree, KNN, Artificial Neural Networks.	Pembagian data yang digunakan tidak disebutkan secara rinci, hanya menyebutkan total data sebanyak 116 untuk mencari nilai akurasi.
4.	*Identification of Breast Cancer Using The Decision Tree Algorithm* (Sathiyanarayanan, Pavithra, Sai Saranya, & Makeswari, 2019).	Data *distribution* atau *Exploratory data analysis* (EDA)	Support Vector Machine (SVM), KNN, Decision Tree.	Pembagian data training dan testing dengan rasio yang digunakan 80:20. Untuk mengukur akurasi dan kinerja sistem
5.	*Machine Learning for Breast Cancer Classification With ANN and Decision Tree* (Hazra, Banerjee, & Badia, 2020).	Seleksi fitur dan ekstraksi fitur statistik	Neural Networks (ANN), Decision Tree (DT)	Nilai akurasi, presisi, recall yang dihasilkan baik. Sehingga, sudahcukup untuk dapat diimplementasikan dalam sebah sistem.

2.2 Kanker Payudara

Kanker payudara merupakan salah satu jenis penyakit kanker yang mematikan di dunia, penyakit ini umumnya terjadi di kalangan wanita akan tetapi tidak menutup kemungkinan penyakit ini juga menyerang kalangan pria, tetapi perbandingannya sangat kecil yakni hanya sebesar 1:100 (Imaduddin et al., 2021) Kanker payudara merupakan sebuah tumor ganas yang berkembang dan menyerang sel pada payudara, gejala awal penyakit ini hanya berupa benjolan kecil yang semakin lama semakin membesar. Menurut Kementerian Kesehatan kasus kanker di Indonesia mencapai angka 42,1 orang per 100 ribu penduduk, angka tersebut bukanlah angka yang terbilang kecil.Selain itu, *World Health Organization* (WHO) pada tahun 2018 menetapkan Indonesia sebagai negara dengan kasus kanker payudara terbanyak, yakni sebesar 58.256 kasus atau setara dengan 6.7% dari total 348.809 kasus kanker di seluruh dunia. Menurut data Globocan jumlah kasus baru kanker payudara di Indonesia pada tahun 2020 mencapai 68.858 kasus (16,6%) dari total 396.914 kasus, dengan jumlah kematian mencapai lebih dari 22 ribu jiwa (Kemenkes, 2022).

Kanker payudara dikategorikan menjadi 3 subtipe utama berdasarkan ada atau tidak adanya penanda molekuler untuk reseptor estrogen atau progesteron dan faktor pertumbuhan epidermal manusia 2 (ERBB2; sebelumnya HER2): reseptor hormon positif/ERBB2 negatif (70% pasien), ERBB2 positif (15% - 20%), dan triple-negatif (tumor tidak memiliki semua 3 penanda molekuler standar; 15%). Lebih dari 90% kanker payudara tidak bermetastasis pada saat diagnosis (Waks & Winer, 2019).

2.3 Data Mining

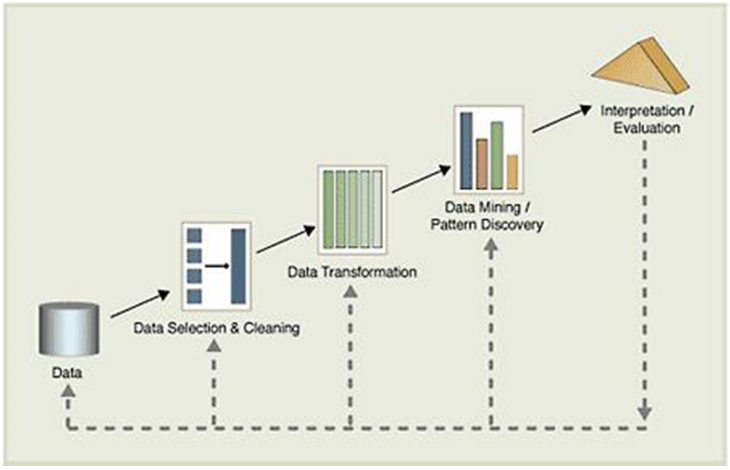
Awal mula istilah *data mining* digunakan oleh komunitas basis data pada tahun 1990-an. Namun, teori dasar serta metode dari *data mining* sendiri telah dikembangkan jauh sebelum tahun 90-an. *Data mining* sendiri berasal dari berbagai disiplin ilmu, dua yang paling mendasar adalah statistika dan pembelajaran mesin (*machine learning*). Teori statistika berakar pada teori matematika, dan berfokus pada pembentukan model. Model adalah pendekatan hipotesis atau struktural yang memperkirakan data secara aktual. Sementara pembelajaran mesin (*machine learning*) pada saat ini lebih tertarik untuk mengembangkan algoritma (Han, J., Kamber, M., & Pei, J., 2012).

Data Mining merupakan sebuah proses eksplorasi yang mana didalamnya terdapat sebuah proses pengumpulan informasi penting dari sekumpulan data. Informasi yang dikumpulkan berupa pola yang tersembunyi pada data, hubungan antar elemen pada data, serta pembuatan sebuah model dengan tujuan penelaahan pada data (Adinugroho & Sari, 2018). Bisa disimpulkan bahwa *data mining* merupakan proses penggalian, penambangan, atau pengumpulan informasi mengenai pengetahuan penting dari sebuah data.

Secara umum, operasi dalam *data mining* dapat diklasifikasikan menjadi dua kategori, yakni metode deskriptif dan prediktif. Metode deskriptif bertujuan untuk menemukan hubungan, pola, ataupun anomali pada sebuah data agar mudah dipahami oleh manusia. Salah satu contoh dari metode deskriptif adalah *association rules* dan *clustering*. Sedangkan metode deskriptif bertujuan untuk

memperkirakan nilai suatu variabel berdasarkan nilai variabel lainnya. Klasifikasi dan regresi merupakan contoh dari metode prediktif (Adinugroho & Sari, 2018).

Knowledge discovery (mining) in database (KDD) merupakan sebutan lain yang dimiliki oleh *data mining,* beberapa sebutan lainnya yang dimiliki oleh *data mining* adalah *knowledge extraction* (ekstraksi pengetahuan), *business intelligence,* dan lain sebagainya (Han, J., Kamber, M., & Pei, J., 2012).



Gambar 2.1 Proses Data Mining

Proses pada *data mining* ditunjukan oleh gambar 2.1. Adapun tahapan-tahapannya yakni sebagai berikut:

1. Data

   Tahapan pertama pada proses ini adalah hasil seleksi data yang digunakan pada proses ini akan disimpan ke dalam bentuk berkas yang kemudian akan disimpan terpisah dari operasional basis data.

2. *Cleaning* dan *Selection* Data

Tahapan kedua adalah *cleaning* (pembersihan) dan *selection* (seleksi) data, dimana data baru dari kumpulan data operasional harus diseleksi terlebih dahulu sebelum memasuki tahap penggalian informasi. Kemudian data yang sudah diseleksi dibersihkan (*cleaning*) sebelum memasuki proses *data mining*. Pembersihan data disini berupa perbaikan terhadap kesalahan cetak, data tidak konsisten, serta membuang data yang terduplikat.

3. Transformasi

   Tahapan ketiga hasil seleksi data kemudian diubah menyesuaikan bentuk dan format pada proses *data mining.* pada *knowledge discovery (mining) in database* (KDD) terdapat format tertentu yang dapat diaplikasikan. Seperti pada teknik *clustering* format yang diterima berupa input data kategorikal. Begitu pula pada data yang digunakan pada teknik *data mining*, data yang digunakan perlu dilakukan pemilihan terlebih dahulu.

4. *Data Mining*

   Tahapan selanjutnya adalah melakukan pencarian informasi atau pola dari sumber data yang sudah ada dan kemudian menerapkan metode tertentu pada data tersebut. Metode atau biasa juga disebut algoritma pada *data mining* memiliki jenis yang beragam, sehingga pemilihan metode pada tahap ini disesuaikan dengan tujuan dan kebutuhan proses secara keseluruhan.

5. Interpretasi dan Evaluasi

   Tahap terakhir dari proses *data mining* adalah interpretasi atau menarik kesimpulan dari hasil pola informasi yang sudah dilakukan pada proses sebelumnya. Pola informasi tersebut kemudian diubah kedalam bahasa yang

mudah dimengerti dan dipahami oleh manusia. Pada tahap ini juga dilakukan pemeriksaan terhadap hasil informasi yang sudah didapatkan apakah sesuai dengan hipotesis yang ada atau tidak (Suhartono, Kurniawan, & Imran, 2018).

2.4 Klasifikasi

Klasifikasi merupakan suatu tahapan penting dari *data mining.* Teknik klasifikasi sudah banyak digunakan dalam berbagai masalah pada suatu penelitian. Klasifikasi sendiri merupakan sebuah metode pengelompokan data yang mana didalamnya terdapat sebuah proses untuk menemukan sebuah model yang dapat membedakan, serta menggambarkan kelas pada konsep suatu data. Model ini diturunkan berdasarkan analisis yang didapat dari set data pelatihan. Penggunaan model ini biasa digunakan untuk memprediksi sebuah label kelas pada objek yang sebelumnya belum diketahui (Nikmatun & Waspada, 2019).

Konsep klasifikasi pada *data mining* secara umum memiliki kemiripan dengan klasifikasi yang terdapat pada bidang biologi. Klasifikasi pada kedua bidang ini memiliki tujuan yang sama yakni melakukan proses pengelompokan pada sebuah data. Perbedaannya terletak pada data yang digunakan pada masing-masing bidang tersebut, pada bidang biologi data yang dikelompokan berupa data makhluk hidup, sedangkan pada *data mining* klasifikasi disini mengelompokkan data yang kemudian dijadikan beberapa kelompok (Nikmatun & Waspada, 2019). Secara umum, proses klasifikasi terbagi menjadi dua tahap, yakni sebagai berikut:

1. *Learning*

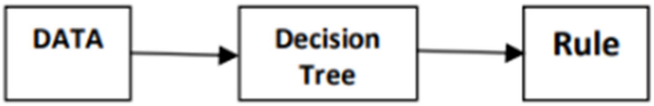
Tahap pertama dalam proses klasifikasi merupakan sebuah proses dimana kumpulan data yang sudah diketahui kelasnya diumpamakan, dengan tujuan untuk membentuk sebuah model perkiraan

2. *Test*

   Proses pada tahap ini adalah menguji model yang sudah terbentuk dengan data lainnya, dengan tujuan untuk mengetahui nilai akurasi dari model tersebut. Apabila nilai akurasi mencukupi, maka model tersebut dapat digunakan sebagai prediksi kelas data yang sebelumnya belum diketahui.

2.5 Pohon Keputusan (*Decision Tree*)

Decision tree merupakan sebuah metode prediksi dan klasifikasi yang sangat populer yang berbentuk seperti pohon. Metode ini dapat mengubah data yang berukuran besar menjadi data berukuran kecil. Konsep dari metode *decision tree* ditunjukan pada gambar 2.2.

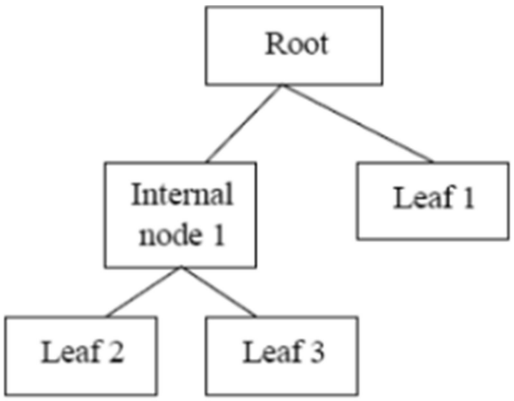


Gambar 2.2 Konsep Desicion Tree

Konsep dasar pada metode *decision tree* adalah merubah sebuah data menjadi suatu keputusan yang berbentuk pohon dengan beberapa aturan pengambilan keputusan. Manfaat dari penerapan metode ini adalah kemampuan dalam menyederhanakan pengambilan sebuah keputusan yang bersifat kompleks, sehingga solusi yang didapat dari pengambilan keputusan tersebut nantinya akan lebih menginterpretasikan permasalahan dari studi kasus yang diambil. *Decision*

tree juga dapat disebut sebagai struktur pada analisis pemecahan suatu masalah, serta dapat dijadikan sebagai sarana pemetaan alternatif dalam memecahkan suatu masalah.

Proses awal untuk membangun *tree* dimulai dengan data yang berada pada simpul akar (*root node*) yang kemudian dilanjutkan dengan langkah selanjutnya yaitu pemilihan atribut, perumusan uji logika (*logical test*) pada atribut yang sudah dipilih, serta percabangan pada setiap hasil pengujian dari tes tersebut. Sederhananya,alur pada proses ini dimulai dari simpul akar menuju simpul daun (Defiyanti & Pardede, 2008).



Gambar 2.3 Proses Desicion Tree

Proses pada *decision tree* terdiri dari akar (*root*)*, internal node,* dan *leaf* sebagaimana yang ditunjukan oleh gambar 2.3.

1. *Root* merupakan *node* teratas pada *decision tree* yang tidak memiliki input, dapat menghasilkan *output* lebih dari satu atau tidak menghasilkan *output* sama sekali.

2. *Internal Node* merupakan *node* percabangan, dimana pada *node* ini hanya terdapat satu input, dengan *output* yang dihasilkan minimal dua.
3. *Leaf* merupakan node *node* terakhir atau *terminal node* pada proses *decision tree,* dimana pada *node* ini hanya terdapat satu input, tanpa menghasilkan *output.*

Dilihat dari segi fungsionalnya *decision tree* merupakan salah satu metode *data mining* yang merepresentasikan bentuk pohon (*tree*) dengan tujuan untuk menentukan aturan pada proses klasifikasi. Tipe pada *decision tree* terdapat dua jenis, yaikni *classification tree* dan *regression tree*. *Classification tree* merupakan klasifikasi dengan proses memberi label serta memasukan *record* ke dalam suatu kelas yang sebelumnya telah disediakan, sedangkan *regression tree* merupakan proses klasifikasi membuat estimasi nilai dari sebuah variabel target yang bernilai numerik (Defriani & Jaelani, 2020).

2.6 Algoritma *Iterative Dichotomiser-3* (ID3)

Algoritma *iterative dichotomiser-3* atau biasa dikenal dengan algoritma ID3 merupakan metode dalam *data mining* dan merupakan metode dasar pada *decision tree.* Algoritma *iterative dichotomiser-3* (ID3) pertama kali dikebangkan oleh J. Ross Quinlan pada tahun 1979. Algoritma ID3 dirancang dengan sederhana, dimana terdapat banyak atribut dan set pelatihan yang berisi banyak objek, klasifikasi pada pohon keputusan baik digunakan tanpa banyak perhitungan (Quinlan, 1986). Pencarian yang dilakukan oleh metode ID3 dilakukan secara menyeluruh dan tamak (*greedy*) terhadap setiap kemungkinan pada sebuah pohon keputusan. Algoritma *iterative dichotomiser-3* (ID3) digunakan karena

menghasilkan aturan klasifikasi yang mudah dipahami, memungkinkan pembuatan pohon keputusan secara cepat, serta memerlukan sedikit konfigurasi (Pribadi, Athiry, Saputra, Supiandi, & Prayudi, 2018).

Algoritma *iterative dichotomiser-3* (ID3) membangun *decision tree* secara menyeluruh dari atas ke bawah (*top-down*) untuk memeriksa atribut yang cocok untuk ditempatkan pada *root*. Semua atribut tersebut diperiksa dengan cara dievaluasi menggunakan ukuran statistik yang sudah ditentukan, umumnya ukuran statistik yang digunakan adalah *information gain*. Evaluasi yang dilakukan pada setiap atribut bertujuan untuk mengukur efektivitas dalam mengklasifikasikan sampel data (Elmande & Widodo, 2012). Sederhananya, tahapan dalam pemilihan atribut menggunakan *information gain* dapat digambarkan sebagai berikut:

1. Pilih atribut yang memiliki nilai *information gain* terbesar.
2. Membuat simpul yang berisikan atribut yang sudah dipilih.
3. Melakukan perhitungan pada *information gain* secara menyeluruh pada semua data, hingga menjadi satu dalam kelas yang sama. Atribut yang sudah dipilih tidak dapat disertakan lagi dalam perhitungan *information gain.*

Algoritma *iterative dichotomiser-3* (ID3) memiliki keunggulan dibandingkan dengan algoritma lainnya pada metode *decision tree,* kelebihan tersebut yakni sebagai berikut:

1. Data *training* yang diterapkan pada algoritma ini menghasilkan aturan prediksi yang mudah dipahami.

2. Dapat membangun pohon keputusan secara cepat dan makasimal dengan ukuran pohon yang pendek.
3. ID3 melakukan pencarian secara menyeluruh pada kumpulan data untuk membuat keseluruhan pohon keputusan.
4. Dapat menemukan *leaf node* atau terminal *node* dengan cepat sehingga memungkinkan pemangkasan serta mengurangi jumlah pengujian pada data.
5. Waktu perhitungan ID3 adalah fungsi linier dari perkalian bilangan karakteristik dan bilangan simpul.

Proses perhitungan pada algoritma *iterative dichotomiser-3* (ID3) dilakukan dengan cara pembentukan pohon klasifikasi menggunakan dua langkah. Langkah pertama yang dilakukan adalah dengan cara menentukan nilai *entropy*, kemudian dilanjut dengan langkah kedua yakni menghitung nilai *information gain* pada tiap variabel (Pribadi et al., 2018). *Entropy* pada proses ini berfungsi untuk mengukur *node* yang digunakan sebagai parameter pada sampel data. Apabila *entropy* semakin besar maka kumpulan data semakin bervariasi (heterogen), sedangkan *gain* berfungsi untuk mengukur seberapa baik kinerja atribut dalam memisahkan *training example* ke dalam target kelas. Perhitungan nilai *entropy* pada algoritma ID3 yakni sebagaimana persamaan berikut:

$$Entropy(S) = -P_{+}log_2P_{+} - P_{-}log_2P_{-} \quad \textbf{(4.1)}$$

dimana:

S = data sampel yang digunakan sebagai *training*

P_{+} = probablitas sampel S dengan kelas positif.

$P\_$ = probablitas sampel S dengan kelas positif.

Pengurangan *entropy* pada algoritma ID3 disebut dengan *information gain.* Pembagian sampel S terhadap atribut X dihitung dengan menggunakan rumus *information gain* yakni sebagaimana persamaan berikut:

$$Gain(S,X) = Entropy(S) - \sum_{V\ \epsilon\ value(X)} \frac{|S_V|}{|S|}\ Entropy(S_V) \qquad \textbf{(4.2)}$$

dimana:

X = atribut

V = nilai yang memungkin untuk atribut X

$Value(X)$ = himpunan yang mungkin untuk atribut (X)

Setelah *information gain* pada semua atribut dihitung, kemudian dipilih nilai *information gain* tertinggi untuk dijadikan *root* pada suatu pohon keputusan. Hal ini dilakukan seterusnya hingga parameter pada tiap-tiap atribut terklasifikasi dengan sempurna.

Algoritma *iterative dichotomiser-3* (ID3) memiliki kelemahan dibandingkan dengan algoritma lainnya pada metode *decision tree.* Kelemahan tersebut merupakan masalah *overfitting*. *Overfitting* merupakan suatu kondisi di mana sebuah model terlalu memaksakan diri untuk mempelajari pola-pola dari data training, sehingga model tersebut tidak dapat menangani data baru dengan baik. Hal ini dapat menyebabkan model tersebut tidak dapat memberikan hasil yang akurat ketika digunakan untuk memprediksi data baru.

Salah satu penyebab *overfitting* pada algoritma *iterative dichotomiser-3* (ID3) adalah bahwa algoritma ini tidak memiliki mekanisme untuk

mengendalikan kompleksitas pohon keputusan yang dibangun. Algoritma ini akan terus membuat cabang baru hingga tidak ada lagi data yang dapat dipisahkan dengan baik, sehingga dapat menyebabkan pohon keputusan yang terlalu kompleks dan tidak generalisasinya buruk.

Untuk mengatasi masalah *overfitting* pada algoritma *iterative dichotomiser-3* (ID3), salah satu cara yang dapat dilakukan adalah dengan menggunakan teknik pruning. Teknik pruning ini akan memotong cabang-cabang pohon keputusan yang tidak relevan atau tidak penting, sehingga pohon keputusan yang dihasilkan akan lebih sederhana dan lebih mudah dibaca. Dengan demikian, diharapkan dapat mengurangi masalah *overfitting* pada algoritma *iterative dichotomiser-3* (ID3).

2.7 Algoritma C5.0

Algoritma C5.0 adalah sebuah algoritma klasifikasi yang dikembangkan oleh John Quinlan. Algoritma ini merupakan perkembangan dari algoritma C4.5 yang sebelumnya telah dikembangkan oleh Quinlan. Algoritma C5.0 memiliki beberapa keunggulan dibandingkan dengan algoritma C4.5, diantaranya adalah kemampuan C5.0 untuk menangani data yang lebih besar dengan lebih cepat, serta kemampuan untuk menangani data yang memiliki lebih banyak atribut (fitur). Algoritma C5.0 menggunakan teknik pembelajaran klasifikasi dengan menggunakan decision tree (pohon keputusan). Algoritma ini menggunakan prinsip "divide and conquer" (membagi dan menaklukkan) dengan membagi data menjadi kelompok-kelompok yang lebih kecil dan lebih homogen (memiliki ciri

yang sama). Kemudian, algoritma tersebut akan mencoba untuk memprediksi kelas dari data tersebut dengan menggunakan pohon keputusan.

Algoritma C5.0 biasa digunakan dalam berbagai aplikasi, seperti pengenalan pola, analisis data, dan sistem rekomendasi. Contohnya, algoritma ini dapat digunakan untuk memprediksi apakah seseorang akan mengalami stroke atau tidak berdasarkan data seperti usia, riwayat penyakit, dan gaya hidup. Algoritma ini juga dapat digunakan dalam sistem rekomendasi untuk menyarankan produk atau layanan kepada pelanggan berdasarkan data seperti preferensi, riwayat pembelian, dan perilaku pembelian.

Tahapan pada proses perhitungan yang terdapat pada algoritma C5.0 dimulai dengan langkah pertama yakni penentuan *root node*, *root node* adalah *node* pertama yang memisahkan semua data. Langkah berikutnya adalah *splitting*, pada tahap ini setiap *node* diperiksa dan dibagi menjadi beberapa *child node* berdasarkan fitur yang paling berguna diterminasi dengan mengunakan *entropy*, *gain ratio*. Tahap terakhir adalah pembentukan pohon, yang mana pada tahap ini merupaka proses *splitting* yang dilakukan pada setiap *child node* hingga setiap node memiliki satu kelas (Han, Kamber, & Pei, 2012). Proses perhitungan tersebut memiliki kesamaan dengan algoritma ID-3 yakni mulai dari peritungan *entropy* dan *information gain*, akan tetapi pada algoritma C5.0 atribut dengan *gain ratio* tertinggi akan dipilih sebagai *root node.* Adapun rumus perhitungan *gain ratio* yakni sebagaimana persamaan berikut:

$$Gain\ Ratio = \frac{Gain(S, X)}{\sum_{i=1}^{m} Entropy(S_i)} \quad \textit{(4.3)}$$

dimana:

$Gain(S, X)$ = nilai *gain* dari setiap atribut

i = 1 *entropy*

$(S)_i$ = jumlah nilai *entropy* dalam satu atribut

Beberapa keunggulan yang dimiliki oleh metode C5.0 dibandingkan dengan *iterative dichotomiser-3* (ID3), diantaranya adalah:

Algoritma C5.0 dapat menangani data yang tidak terstruktur dengan lebih baik dibandingkan *iterative dichotomiser-3* (ID3). Algoritma C5.0 menggunakan teknik discretization untuk mengubah atribut-atribut kontinu menjadi atribut-atribut diskrit sebelum membuat pohon keputusan. Dengan demikian, algoritma ini dapat menangani data yang memiliki banyak data yang hilang atau tidak valid dengan lebih baik dibandingkan *iterative dichotomiser-3* (ID3). Algoritma C5.0 dapat membuat pohon keputusan yang lebih sederhana dan lebih mudah dibaca dibandingkan *iterative dichotomiser-3* (ID3) (Witten, I. H., & Frank, E.2002).

Algoritma C5.0 menggunakan teknik pruning untuk memotong cabang-cabang pohon keputusan yang tidak relevan atau tidak penting, sehingga pohon keputusan yang dihasilkan akan lebih sederhana dan lebih mudah dibaca. Dengan demikian, algoritma ini dapat membantu dokter dalam mengambil keputusan yang lebih akurat dan tepat waktu. Pohon keputusan yang dihasilkan oleh algoritma C5.0 dapat membuat pohon keputusan yang lebih akurat dibandingkan *iterative dichotomiser-3* (ID3). Algoritma C5.0 menggunakan teknik boosting untuk meningkatkan keakuratan pohon keputusan yang dihasilkan. Teknik ini akan membuat beberapa pohon keputusan yang kecil dan saling berhubungan, lalu

menggabungkan semua pohon keputusan tersebut menjadi satu pohon keputusan yang lebih akurat. Dengan demikian, algoritma ini dapat memprediksi data baru dengan lebih akurat dibandingkan *iterative dichotomiser-3* (ID3) (Witten, I. H., & Frank, E.2002).

2.8 Integrasi Keislaman

Dalam kajian keislaman terdapat dua sudut pandang yakni sudut pandang secara medis atau kesehatan serta sudut pandang secara statistik atau teknis. Sudut pandang secara medis Al-Quran telah mengajarkan kepada seluruh umat manusia agar mengetahui tanda-tanda kebesaran Allah dengan memahami pentingnya kesehatan serta upaya dalam mengatasi penyakit. Sebagaimana firman Allah SWT dalam QS. Al-Anbiya' ayat 51:

وَلَقَدْ اٰتَيْنَآ اِبْرٰهِيْمَ رُشْدَهٗ مِنْ قَبْلُ وَكُنَّا بِهٖ عٰلِمِيْنَ

"*Dan sungguh, sebelum dia (Musa dan Harun) telah kami berikan kepada Ibrahim petunjuk, dan kami telah mengetahui dia*".

Dari ayat tersebut, Allah telah memberikan petunjuk serta bimbingan kepada hamba-Nya. Dalam konteks kesehatan, ayat tersebut dapat diartikan bahwasanya Allah memberikan pengetahuan dan petunjuk bagi manusia dalam mencari pengetahuan medis serta dapat mengembangkan teknologi untuk mengatasi penyakit, termasuk kanker payudara.

Dalam perspektif islam Allah SWT berkehendak untuk menciptakan seluruh alam semesta, begitupula Allah menunjukan belas kasih-Nya terhadap hambanya-Nya dengan menciptakan penyakit disertai dengan obat penawarnya, hal ini menunjukan bahwa Allah tidak menciptakan sesuatu tanpa tujuan dan

makna yang mendalam. Setiap ciptaan-Nya di dunia memiliki hikmah-hikmah tertentu. Hal ini tercermin dalam sabda Nabi Muhammad SAW yang tercantum dalam hadits riwayat Ibnu Majah.

مَا أَنْزَلَ اللهُ دَاءً إِلَّا أَنْزَلَ لَهُ شَفَاءً

"Allah tidak menciptakan sesuatu suatu penyakit tanpa menciptakan pula obat untuknya". (HR. Ibnu Majah)

Hadits tersebut menyampaikan keyakinan bahwa Allah SWT menciptakan penyembuhan untuk setiap penyakit. Dalam konteks kanker payudara, meskipun tidak ada penjelasan langsung tentang decision tree, ayat ini dapat mengingatkan kita bahwa Allah SWT menciptakan berbagai cara penyembuhan, termasuk pengembangan teknologi dan metode klasifikasi seperti *decision tree* untuk membantu diagnosis dan pengobatan penyakit.

Sedangkan sudut pandang secara teknis atau statistik Ayat Al-Quran memberikan petunjuk moral, etika, dan prinsip-prinsip hidup yang dapat membimbing kita dalam menggunakan teknologi dan pengetahuan dengan bijak. Dalam konteks kesehatan dan penggunaan teknik statistik seperti *decision tree*, beberapa prinsip Islam yang relevan yang dapat diterapkan, sebagaimana firman Allah SWT dalam QS. An-Nisa' ayat 135:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا كُوْنُوْا قَوّٰمِيْنَ بِالْقِسْطِ شُهَدَاۤءَ لِلّٰهِ وَلَوْ عَلٰٓى اَنْفُسِكُمْ اَوِ الْوَالِدَيْنِ وَالْاَقْرَبِيْنَ ۚ اِنْ يَّكُنْ غَنِيًّا اَوْ فَقِيْرًا فَاللّٰهُ اَوْلٰى بِهِمَا ۗ فَلَا تَتَّبِعُوا الْهَوٰٓى اَنْ تَعْدِلُوْا ۚ وَاِنْ تَلْوٗٓا اَوْ تُعْرِضُوْا فَاِنَّ اللّٰهَ كَانَ بِمَا تَعْمَلُوْنَ خَبِيْرًا

"*Wahai orang-orang yang beriman! Jadilah kamu penegak keadilan, menjadi saksi karena Allah, walaupun terhadap dirimu sendiri atau terhadap ibu bapak dan kaum kerabatmu. Jika dia (yang terdakwa) kaya ataupun miskin, maka Allah lebih tahu kemaslahatan (kebaikannya)*".

Ayat tersebut mengajarkan pentingnya keadilan serta ketepatan dalam pengambilan sebuah keputusan, termasuk dalam penerapan teknik statistik seperti *decision tree*. Ketika menggunakan metode *decision tree* untuk klasifikasi penyakit, hal penting yang harus dipaastikan adalah data yang digunakan harus akurat, terpercaya, dan tidak memihak, agar menghasilkan keputusan yang adil dan tepat. Sedangkan dari sudut pandang pengetahuan dan pembelajaran Allah berfirman dalam QS. Ar-Ra'd 19.

اَفَمَنْ يَّعْلَمُ اَنَّمَآ اُنْزِلَ اِلَيْكَ مِنْ رَّبِّكَ الْحَقُّ كَمَنْ هُوَ اَعْمٰىۗ اِنَّمَا يَتَذَكَّرُ اُولُوا الْاَلْبَابِۙ

"*Maka apakah orang yang mengetahui bahwa apa yang diturunkan Tuhan kepadamu adalah kebenaran, sama dengan orang yang buta? Hanya orang berakal saja yang dapat mengambil pelajaran*".

Ayat tersebut mengingatkan mengenai pentingnya ilmu pengetahuan dan pembelajaran. Dalam penggunaan *decision tree*, pengetahuan yang mendalam menganai metode statistik, interpretasi hasil, serta validitas data sangat penting. Hal ini bertujuan agar kita sebagai mukmin terus belajar dan meningkatkan pemahaman tentang teknik statistik untuk memastikan penggunaannya yang benar dan efektif. Dalam konteks klasifikasi kanker payudara menggunakan *decision tree*, ayat ini juga mengingatkan untuk tidak mengabaikan pengetahuan dan pemahaman yang diberikan melalui upaya penelitian dan pengembangan ilmu dalam bidang kedokteran (Snow, 2012). Menurut sudut pandang secara teknis atau statistic dari segi kewaspadaan dan pertimbangan Allah berfirman dalam QS. Al-Isra' ayat 36.

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهٖ عِلْمٌۗ اِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ اُولٰۤىِٕكَ كَانَ عَنْهُ مَسْـُٔوْلًا

"*Dan janganlah kamu mengikuti yang tidak kamu ketahui. Karena pendengaran, penglihatan dan hati nurani, semua itu akan diminta pertanggungjawabannya*".

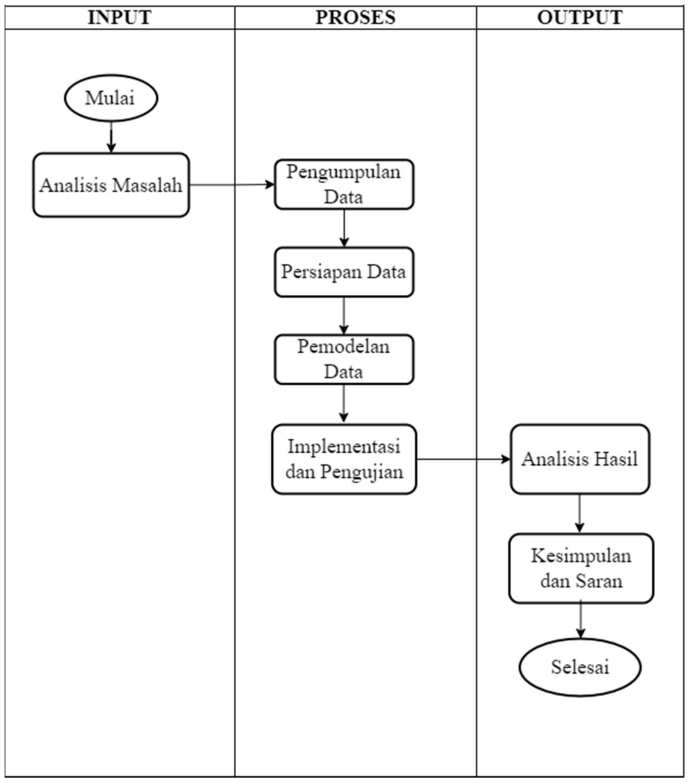
Ayat tersebut menekankan pentingnya kewaspadaan dan pertimbangan sebelum mengambil keputusan. Dalam konteks penggunaan *decision tree* untuk klasifikasi penyakit, penting untuk memahami batasan dan asumsi yang mendasari model tersebut, serta mempertimbangkan kualitas data yang digunakan. Keputusan yang diambil harus didasarkan pada pengetahuan serta pemahaman yang akurat (Freedman, D. A. 2010).

BAB III

METODELOGI PENELITIAN

3.1 Desain Penelitian

Pendekatan yang diterapkan pada penelitian ini merupakan pendekatan kuantitatif, yang mana pendekatan ini merupakan suatu metode yang menggunakan data konkrit dalam bentuk numerik yang nantinya akan dianalisis dengan menggunakan metode statistik sebagai alat uji perhitungan. Perencanaan tahapan kegiatan tentunya diperlukan agar tujuan penelitian dapat tercapai dan terlaksana dengan baik. Adapun tahapan-tahapan secara general mengenai prosedur yang akan dilakukan pada penelitian ini dijabarkan dalam bentuk flowchart pada gambar 3.1.



Gambar 3.1Prosedur Penelitian

3.2 Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder, dimana data tersebut merupakan data yang diperoleh secara tidak langsung melalui sumber lain, atau berupa dokumentasi (Sugiyono, 2011). Dokumentasi tersebut dapat berupa dokumentasi tertulis seperti buku, jurnal, dan informasi lainnya yang memiliki hubungan dengan penelitian ini.

Data sekunder pada penelitian ini diperoleh dari *website* resmi UCI *Machine Learning Repository*. Data tersebut merupakan *dataset breast cancer coimbra* tahun 2018 (UCI, 2018). Pada *dataset* tersebut terdapat 9 atribut dan 1 kelas klasifikasi. Atribut pada *dataset* yang digunakan pada penelitian ini ditunjukkan oleh tabel 3.1

Tabel 3.1 Atribut Kanker Payudara

No	Atribut	Deskripsi
1	*Age*	Umur Responden
2	BMI	*Body Mass Index* (BMI) atau Indeks Massa Tubuh (IMT)
3	*Glucose*	Level Glukosa
4	Insulin	Level Insulin
5	HOMA	*Homeostasis Model Assessment*
6	*Leptin*	Level Leptin
7	*Adiponectin*	Level Adiponectin
8	*Resistin*	Level *Resistin*
9	MCP-1	*Monocytes Chemoattractant Protein*-1
10	*Classification*	Klasifikasi Jenis Kanker

Dari 10 atribut yang pada tabel tersebut, atribut yang digunakan dalam proses klasifikasi hanya 9 atribut. Setiap atribut pada data tersebut memiliki parameter yang sama yakni 1-4, dimana nilai 1 merupakan nilai yang paling dekat dengan kanker jinak, sedangkan nilai 4 merupakan nilai atau parameter yang mendekati kanker ganas

Tabel 3.2 Parameter setiap Atribut

No	Kode	Atribut	Parameter
1	A	*Age*	1-4
2	B	BMI	1-4
3	C	*Glucose*	1-4
4	D	Insulin	1-4
5	E	HOMA	1-4
6	F	*Leptin*	1-4
7	G	*Adiponectin*	1-4
8	I	*Resistin*	1-4
9	J	MCP-1	1-4

Jumlah keseluruhan *dataset* yang digunakan pada penelitian ini sebanyak 116 data, dengan perincian 52 data merupakan kelas *benign* (jinak) serta 64 data merupakan kelas *malign* (ganas).*missing value* pada keseluruhan data yang digunakan berjumlah 0 atau tidak ada. Rentang nilai yang dimiliki oleh masing-masing atribut memiliki nilai yang sama, seperti yang ditunjukan oleh tabel 3.2.

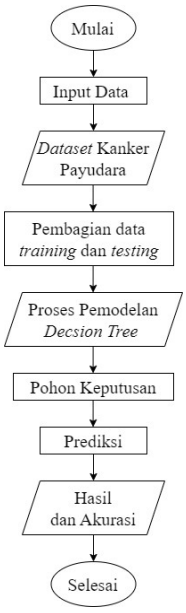
3.3 Objek Penelitian

Fokus utama penelitian ini adalah klasifikasi jenis kanker payudara. Terdapat beberapa atribut data atau variabel data yang menjadi faktor yang menyebabkan kanker payudara. Teknik *data mining* dengan menggunakan metode *decision tree iterative dichotomiser-3* (ID3) digunakan dengan tujuan untuk mengukur nilai akurasi serta mengklasifikasikan jenis kanker payudara.

3.4 Analisis dan Desain Sistem

3.4.1 Desain Sistem

Gambaran umum desain sistem proses menggunakan metode *decision tree iterative dichotomiser-3* (ID3), dijabarkan pada *flowchart* gambar 3.2



Gambar 3.2 Desain Sistem

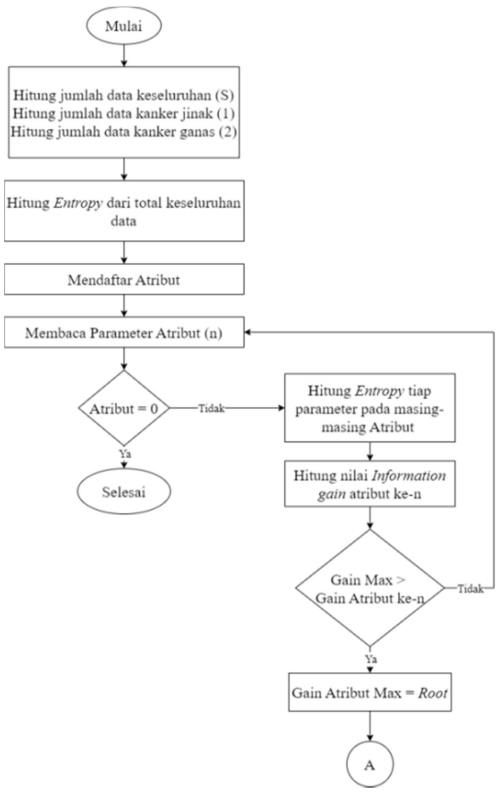
3.4.2 *Dataset* Kanker Payudara

Tahap pertama yang dilakukan pada proses perancangan sistem adalah memasukan data. yang mana data tersebut merupakan *data training* yang

didalamnya memuat seluruh atribut beserta informasinya. yang terdapat pada atribut tersebut. Informasi yang diambil oleh sistem pada setiap atribut antara lain atribut, nilai parameter atribut, serta kelas pada setiap data. *data training* yang digunakan pada penelitian ini tidak memiliki *missing value.*

3.4.3 Pemodelan Data Menggunakan *Decision Tree* ID-3 dan C5.0

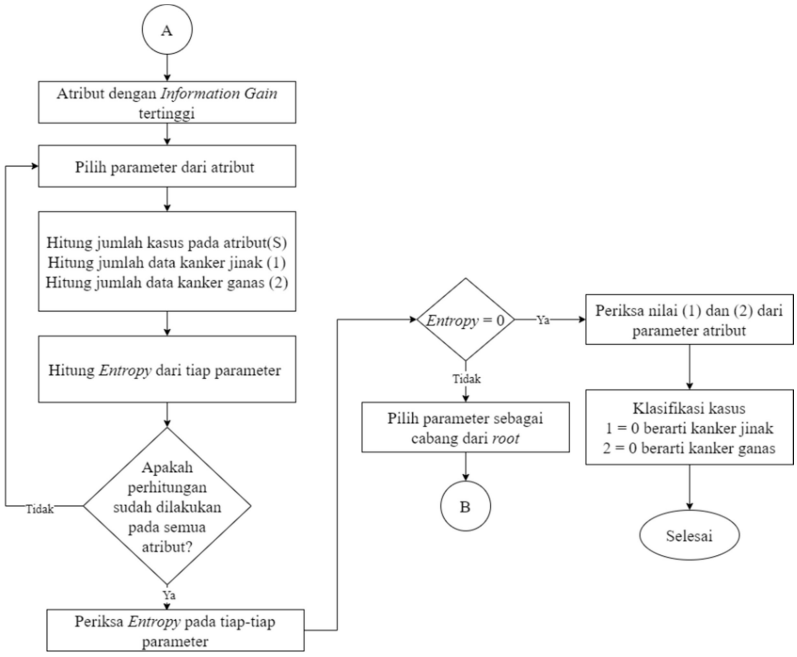
Setelah sistem menerima data untuk diproses, maka proses selanjutnya adalah membuat model klasifikasi. Sistem membangun model klasifikasi dalam bentuk pohon keputusan (*decision tree*). *Flowchart* pada gambar 3.3 merupakan tahapan pemodelan sistem menggunakan *decision tree* ID-3:



Gambar 3.3 Penentuan Akar (Root)

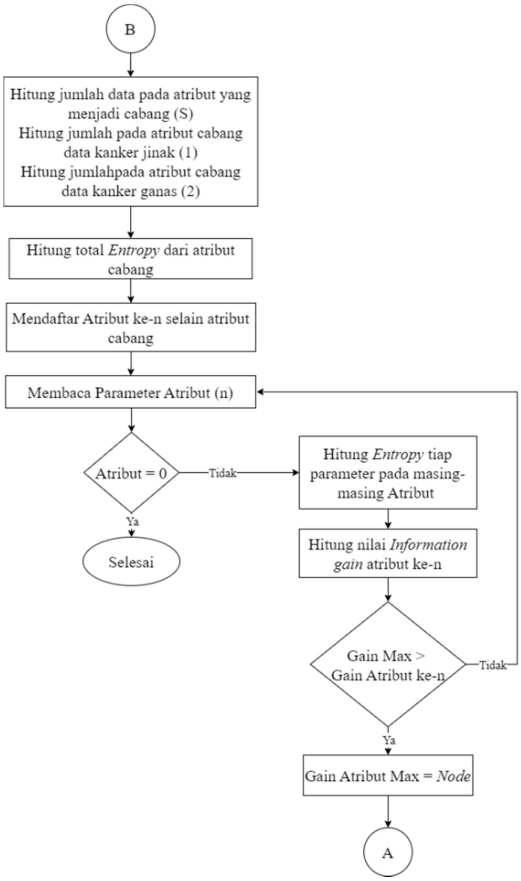
1. Sistem akan menghitung keseluruhan total nilai dari data *training.*
2. Setelah keseluruhan informasi pada data *training* dihitung, langkah selanjutnya adalah menghitung nilai *entropy* sistem dari seluruh data.
3. Setelah nilai *entropy* sistem diperoleh, maka langkah selanjutnya adalah mendaftar atribut.
4. Masing-masing atribut akan dihitung nilai *entropy* beserta *information gain.*
5. Sistem akan membandingkan seluruh nilai *information gain* dari masing-masing atribut untuk mencari nilai *information gain* tertinggi.
6. Atribut yang memiliki nilai *information gain* tertinggi akan dijadikan sebagai akar (*root*).

Setelah atribut yang dijadikan akar (*root*) ditemukan, maka tahap selanjutnya adalah menentukan cabang selanjutnya pada pohon keputusan (*decision tree*), penentuan cabang tersebut yakni sebagaimana *flowchart* berikut:



Gambar 3.4 Penentuan Cabang

1. Atribut dengan *information gain* tertinggi akan terpilih menjadi *root*
2. Pilih parameter yang tidak memiliki nilai 0 pada atribut
3. Hitung total keseluruhan kasus pada parameter atribut terpilih, beserta total kelas pada parameter tersebut
4. Hitung nilai *entropy* pada tiap-tiap parameter atribut, hingga semua atribut habis
5. Nilai *entropy* pada tiap-tiap atribut akan diperiksa, apabila nilai *entropy* pada atribut tersebut adalah nol (0), maka atribut tersebut akan dikoreksi untuk menentukan klasifikasi kasus.
6. Sedangkan apabila nilai *entropy* pada atribut tersebut berjumlah lebih dari nol (0), maka atribut tersebut akan dijadikan cabang node selanjutnya.



Gambar 3.5 Penentuan Node

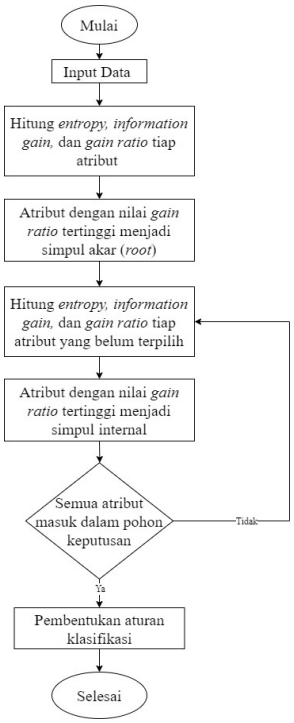
Proses selanjutnya adalah menentukan *node,* gambar 3.5 merupakan *flowchart* dari penentuan *node* pada pohon keputusan (*decision tree*):

1. Sistem akan menghitung keseluruhan total nilai dari atribut yang menjadi cabang.
2. Langkah berikutnya adalah menghitung nilai *entropy* sistem dari seluruh data pada atribut cabang.
3. Mendaftar atribut selain atribut yang memiliki *information gain* tertinggi
4. Seluruh e*ntropy* ke-n padai setiap atribut kemudian dihitung sampai seluruh atribut terhitung habis.

5. Masing-masing atribut kemudian dihitung *information gain*, kemudian atribut dengan nilai *information gain* tertinggi akan dijadikan sebagai *node* selanjutnya.

Setelah *node* terpilih, maka tahap selanjutnya adalah melakukan proses perhitungan kembali secara berulang sampai seluruh atribut dieksekusi habis hingga mencapai bagian terakhir dari pohon keputusan (*end of tree*).

Sedangkan proses pembentukan model klasifikasi pada metode *decision tree* algortima C5.0 memiliki kesamaan dengan proses perhitungan pada algoritma *iterative dichotomiser-3* (ID3). Pada algortitma ID3 penentu *root node* dan atribut pilihan lainnya terletak pada atribut dengan nilai *information gain* tertinggi, akan tetapi pada algoritma C5.0 atribut dengan nilai *gain ratio* tertinggi. ditunjukan oleh *flowchart* gambar 3.6.



Gambar 3.6 Flowchart Algoritma C5.0

1. Data diinputkan kemudian sistem akan menghitung keseluruhan data, yang kemudian dihitung nilai *entropy, information gain,* dan *gain ratio* dari masing-masing atribut.
2. Atribut dengan nilai *gain ratio* tertinggi akan menjadi simpul akar (*root*).
3. Atribut yang belum terpilih kemudian dihitung kembali nilai *entropy, information gain,* dan *gain ratio,* atribut dengan *gain ratio* tertinggi akan menjadi simpul internal pada pohon keputusan.
4. Proses perhitungan dilakukan kembali secara berulang sampai seluruh atribut dieksekusi habis hingga mencapai bagian terakhir dari pohon keputusan (*end of tree*).

3.5 Perhitungan Manual

Berikut ini merupakan perhitungan manual pemodelan data dengan menggunakan metode *decision tree* ID-3:

3.5.1 Memasukkan Data

Data sampel yang akan digunakan pada perhitungan manual sebanyak 10 data, yang mana sampel data tersebut diambil acak dari *dataset* yang sudah ada, adapun data sampel tersebut yakni sebagai berikut:

Tabel 3.3 Data Sampel

A	B	C	D	E	F	G	H	I	Kelas
3	1	2	3	1	4	3	2	1	1
4	3	3	1	3	3	2	4	4	2
1	1	1	4	1	2	1	3	2	2
2	3	3	2	4	2	4	1	3	1

A	B	C	D	E	F	G	H	I	Kelas
4	4	1	1	1	1	1	2	3	1
3	1	2	2	1	3	4	4	1	2
1	4	4	1	1	4	1	3	2	1
2	1	4	3	1	3	2	2	3	2
4	3	2	2	1	1	2	1	3	1
3	1	1	4	3	2	3	1	1	2

3.5.2 Menentukan *Root*

Data *testing* = 10 data

Kelas Jinak (1) = 5 data

Kelas Ganas (2) = 5 data

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

$= -\left(\frac{5}{10}\right) * log_2\left(\frac{5}{10}\right) - \left(\frac{5}{10}\right) * log_2\left(\frac{5}{10}\right)$

$= 1$

3.5.2.1 Perhitungan Atribut A

Parameter atribut A = 1,...4

Perhitungan entropy pada masing-masing atribut dilakukan dengan menggunakan persamaan (2.1) yang mana komponen bagian kanan merupakan jumlah kanker jinak, sedangkan bagian kiri merupakan jumlah kanker ganas.

A1 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

A2 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

A3 = [1, 2] = 3 $= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) = 1$

A4 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

$= 0.918296$

Average Entropy:

$(2/10)1 + (2/10)1 + (3/10)0.918296 + (3/10)0.918296 = 0.950978$

Information gain (S, A):

$$Gain(S, A) = Entropy(S) - \sum_{V\ \epsilon\ value(A)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.950978 = 0.049022$$

3.5.2.2 Perhitungan Atribut B

B1 = [1,4] = 5 $= -\left(\frac{1}{5}\right) * log_2\left(\frac{1}{5}\right) - \left(\frac{1}{5}\right) * log_2\left(\frac{1}{5}\right) = 0.721928$

B2 = [0, 0] = 0 = 0

B3 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

B4 = [2, 0] = 2 = 0

Average Entropy:

$(5/10)0.721928 + (0/10)0 + (3/10)0.918296 + (2/10)0 = 0.636453$

Information gain (S, B):

$$Gain(S, B) = Entropy(S) - \sum_{V\ \epsilon\ value(B)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.636453 \ = \ 0.363547$$

3.5.2.3 Perhitungan Atribut C

C1 = [1, 2] = 3 $= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) = 0.918296$

C2 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

C3 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

C4 = [1, 1] =2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

Average Entropy:

$$(3/10)0.918296 + (3/10)0.918296 + (2/10)1 + (2/10)1 \ = \ 0.950978$$

Information gain (S, C):

$$Gain(S, C) = Entropy(S) - \sum_{V\ \epsilon\ value(C)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.950978$$

$$= \ 0.049022$$

3.5.2.4 Perhitungan Atribut D

D1 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

D2 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

D3 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

D4 = [0, 2] = 2 = 0

Average Entropy:

$(3/10)0.918296 + (3/10)0.918296 + (2/10)1 + (2/10)0 \ = \ 0.750978$

Information gain (S, D):

$$Gain(S, D) = Entropy(S) - \sum_{V \ \epsilon \ value(D)} \frac{|S_V|}{|S|} \ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.750978$$

$$= \ 0.249022$$

3.5.2.5 Perhitungan Atribut E

E1 = [4, 3] = 7 $= -\left(\frac{4}{7}\right) * log_2\left(\frac{4}{7}\right) - \left(\frac{3}{7}\right) * log_2\left(\frac{3}{7}\right) = \ 0.68966$

E2 = [0, 0] = 0 = 0

E3 = [0, 2] = 2 = 0

E4 = [1, 0] = 1 = 0

Average Entropy:

$(7/10)0.985228 + (0/10)0 + (2/10)0 + (1/10)0 \ = \ 0.68966$

Information gain (S, E):

$$Gain(S, E) = Entropy(S) - \sum_{V \ \epsilon \ value(E)} \frac{|S_V|}{|S|} \ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - \ 0.68966$$

$$= \ 0.31034$$

3.5.2.6 Perhitungan Atribut F

F1 = [2, 0] = 2 = 0

F2 = [1, 2] = 3 $= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) = 0.918296$

F3 = [0, 3] = 3 = 0

F4 = [2, 0] = 2 = 0

Average Entropy:

$(2/10)0 + (3/10)0.918296 + (3/10)0 + (2/10)0 = 0.275489$

Information gain (S, F):

$$Gain(S,F) = Entropy(S) - \sum_{V\ \epsilon\ value(F)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.275489 = 0.724511$$

3.5.2.7 Perhitungan Atribut G

G1 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

G2 = [1, 2] = 3 $= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) = 0.918296$

G3 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

G4 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

Average Entropy:

$(3/10)0.918296 + (3/10)0.918296 + (2/10)1 + (2/10)1 = 0.950978$

Information gain (S, G):

$$Gain(S,G) = Entropy(S) - \sum_{V\ \epsilon\ value(G)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.950978$$

$$= \ 0.049022$$

3.5.2.8 Perhitungan Atribut H

H1 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

H2 = [2, 1] = 3 $= -\left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

H3 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

H4 = [0, 2] = 2 = 0

Average Entropy:

$(3/10)0.918296 + (3/10)0.918296 + (2/10)1 + (2/10)0 \ = \ 0.750978$

Information gain (S, H):

$$Gain(S,H) = Entropy(S) - \sum_{V\ \epsilon\ value(H)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.750978$$

$$= \ 0.249022$$

3.5.2.9 Perhitungan Atribut I

I1 = [1, 2] = 3 $= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) = 0.918296$

I2 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

I3 = [3, 1] = 4 $= -\left(\frac{3}{4}\right) * log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) * log_2\left(\frac{1}{4}\right) = 0.811278$

I4 = [0, 1] = 1 = 0

Average Entropy:

$$(3/10)0.918296 + (2/10)1 + (4/10)0.811278 + (1/10)0 = 0.8$$

Information gain (S, I):

$$Gain(S, I) = Entropy(S) - \sum_{V\ \epsilon\ value(I)} \frac{|S_V|}{|S|}\ Entropy(S_V)$$

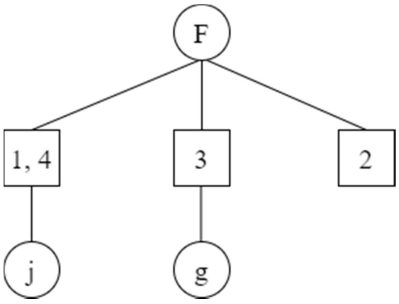
$$= Entropy(S) - Average\ Entropy$$

$$= 1 - 0.8$$

$$= 0.2$$

Berdasarkan hasil perhitungan manual yang sudah dilakukan diatas, nilai *information gain* tertinggi diperoleh atribut F sebesar 0.820112, sehingga atribut F dijadikan sebagai akar (*root*) pada pohon (*tree*).

Setelah akar (*root*) sudah diperoleh, maka langkah selanjutnya adalah melakukan pengecekan terhadap seluruh parameter pada atribut F (*root*), hal tersebut bertujuan untuk mengklasifikasikan apakah seluruh parameter sudah terklasifikasi dengan sempurna atau belum. Hasil dari klasifikasi pada atribut F memperoleh informasi bahwa parameter 1 dan 4 termasuk ke dalam kategori jinak, parameter 3 termasuk ke dalam kategori ganas, sedangkan untuk parameter 2 masih belum terklasifikasi atau terdefinisi dengan sempurna. Atribut F beserta turunannya dapat dilihat pada gambar 3.7.



Gambar 3.7 Root Atribut F dan Parameter Turunannya

3.5.3 Pencarian *Child*

3.5.3.1 *Child* Parameter 2 Atribut F

Tabel 3.4 Parameter 2 Atribut F

Parameter	A	B	C	D	E	G	H	I	Kls
2	1	1	1	4	1	1	3	2	2
2	2	3	3	2	4	4	1	3	1
2	3	1	1	4	3	3	1	1	2

Data *testing* = 3 data

Kelas Jinak (1) = 1 data

Kelas Ganas (2) = 2 data

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

$= -\left(\frac{1}{3}\right) * log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right) * log_2\left(\frac{2}{3}\right)$

$= 0.9183$

Langkah selanjutnya yakni menghitung kembali nilai *entropy* serta *information gain* dari masing-masing atribut:

Entropy Atribut A

A1 = [0, 1] = 0

A2 = [1, 0] = 0

A3 =[0, 1] = 0

Information gain (S, A):

$= 0.9183 - (1/3)0 - (1/3)0 - (1/3)0 = 0.91829$

Entropy Atribut B

B1 = [0, 2] = 0

B3 = [1, 0] = 0

Information gain (S, B):

$= 0.9183 - (2/3)0 - (1/3)0 = 0.91829$

Entropy Atribut C

C1 = [0, 2] = 0

C3 = [1, 0] = 0

Information gain (S, C):

$= 0.9183 - (2/3)0 - (1/3)0 = 0.91829$

Entropy Atribut D

D2 = [1, 0] = 0

D4 = [0, 2] = 0

Information gain (S, D):

$= 0.9183 - (1/3)0 - (2/3)0 = 0.91829$

Entropy Atribut E

E1 = [0, 1] = 0

E3 = [0, 1] = 0

E4 = [1, 0] = 0

Information gain (S, E):

$$= 0.9183 - (1/3)0 - (1/3)0 - (1/3)0 = 0.91829$$

Entropy Atribut G

G1 = [0, 1] = 0

G3 = [0, 1] = 0

G4 = [1, 0] = 0

Information gain (S, G):

$$= 0.9183 - (1/3)0 - (1/3)0 - (1/3)0 = 0.91829$$

Entropy Atribut H

H1 = [1, 1] = 2 $= -\left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * log_2\left(\frac{1}{2}\right) = 1$

H3 = [0, 1] = 0

Information gain (S, H):

$$= 0.9183 - (2/3)1 - (1/3)0 = 0.25162$$

Entropy Atribut I

I1 = [0, 1] = 0

I2 = [0, 1] = 0

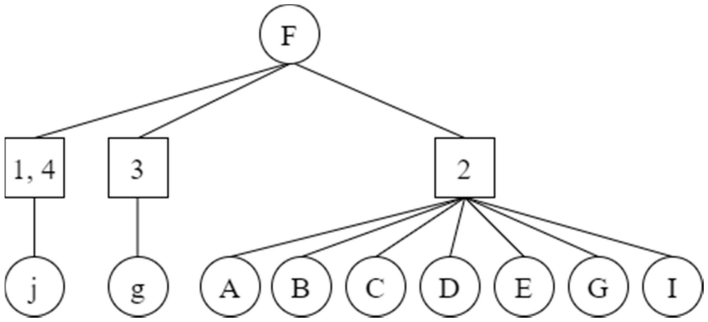
I3 = [1, 0] = 0

Information gain (S, I):

$$= 0.9183 - (1/3)0 - (1/3)0 - (1/3)0 = 0.91829$$

Dari perhitungan diatas, nilai *information gain* tertinggi didapatkan oleh atribut A, B, C, D, E, G, dan I. Atribut-atribut tersebut menunjukan

child dari parameter 2 yaitu atribut F. *Child* dari parameter 2 atribut F dapat dilihat pada gambar 3.8.



Gambar 3.8 Child Parameter 2 Atribut F

3.5.4 Pencarian Parameter Level 3

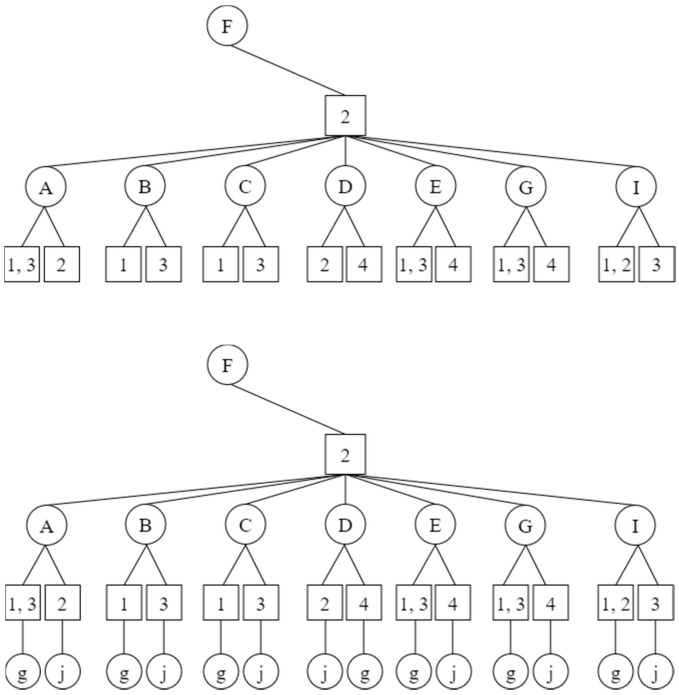
Berdasarkan parameter 2 terdapat 7 turunan atribut yakni, A, B, C, D, E, G, dan I seperti yang ditunjukan pada Tabel 3.5.

Tabel 3.5 Tabel Atribut F Parameter 2

Parameter	A	B	C	D	E	G	H	I	Kelas
2	1	1	1	4	1	1	3	2	2
2	2	3	3	2	4	4	1	3	1
2	3	1	1	4	3	3	1	1	2
Gain	0.91829	0.91829	0.91829	0.91829	0.91829	0.91829	0.25162	0.91829	

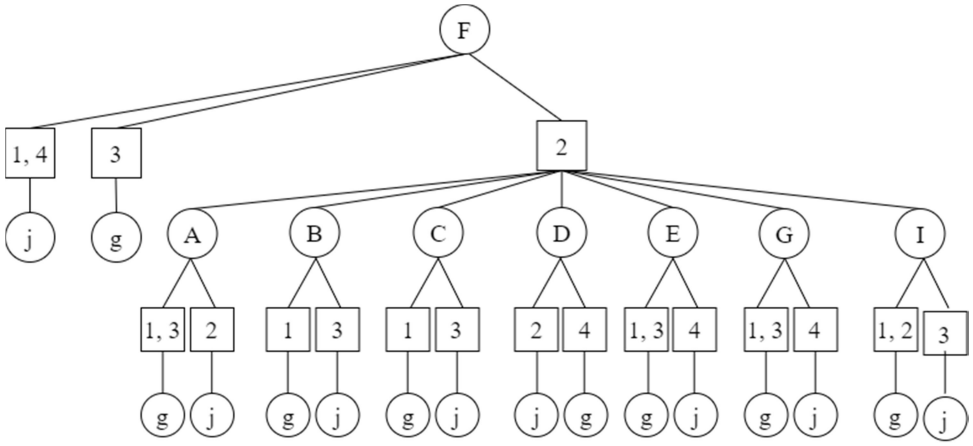
Berdasarkan nilai *information gain* yang diperoleh pada tabel atribut A parameter 1 dan 3 termasuk ke dalam kelas ganas, sedangkan parameter 2 termasuk ke dalam kelas jinak. Pada atribut B parameter 1 termasuk ke dalam kelas ganas, sedangkan parameter 3 kelas jinak. Pada atribut C, parameter 1 termasuk ke dalam kelas ganas, dan parameter 3 termasuk kelas ganas. Pada atribut D parameter 2 termasuk kelas jinak, sedangkan parameter 4 termasuk kelas ganas. Pada atribut E parameter 1 dan 3 termasuk kelas ganas, sedangkan

parameter 4 termasuk kelas jinak. Pada atribut G parameter 1 dan 3 termasuk ke dalam kelas ganas, sedangkan parameter 4 termasuk kelas jinak. Pada atribut I parameter 1 dan 2 termasuk ke dalam kelas ganas, sedangkan parameter 3 termasuk kelas jinak. Model *decision tree* yang dihasilkan ditunjukkan oleh gambar 3.9.



Gambar 3.9 Child Parameter 2 Level 3

Sehingga model *decision tree* yang diperoleh adalah sebagaimana gambar 3.10



Gambar 3.10 Model Decision Tree

BAB IV

HASIL DAN PEMBAHASAN

4.1 Skenario Pengujian

Pengujian yang dilakukan pada penelitian ini bertujuan untuk mengevaluasi tingkat ketepatan atau akurasi dari hasil keputusan diagnosa penyakit kanker payudara dengan menggunakan beberapa skenario uji coba. Beberapa skenario uji coba pada penelitian ini dilakukan untuk mengetahui variasi nilai akurasi yang yang ada pada kedua metode *decision tree.* Langkah yang dilakukan pertamakali adalah memisahkan data menjadi dua bagian, yakni data latih dan data uji dengan jumlah persentase beragam. Tahap pengujian dalam penelitian ini dilakukan pada setiap iterasi menggunakan skenario yang dijelaskan oleh tabel 4.1.

Tabel 4.1 Skenario Pengujian

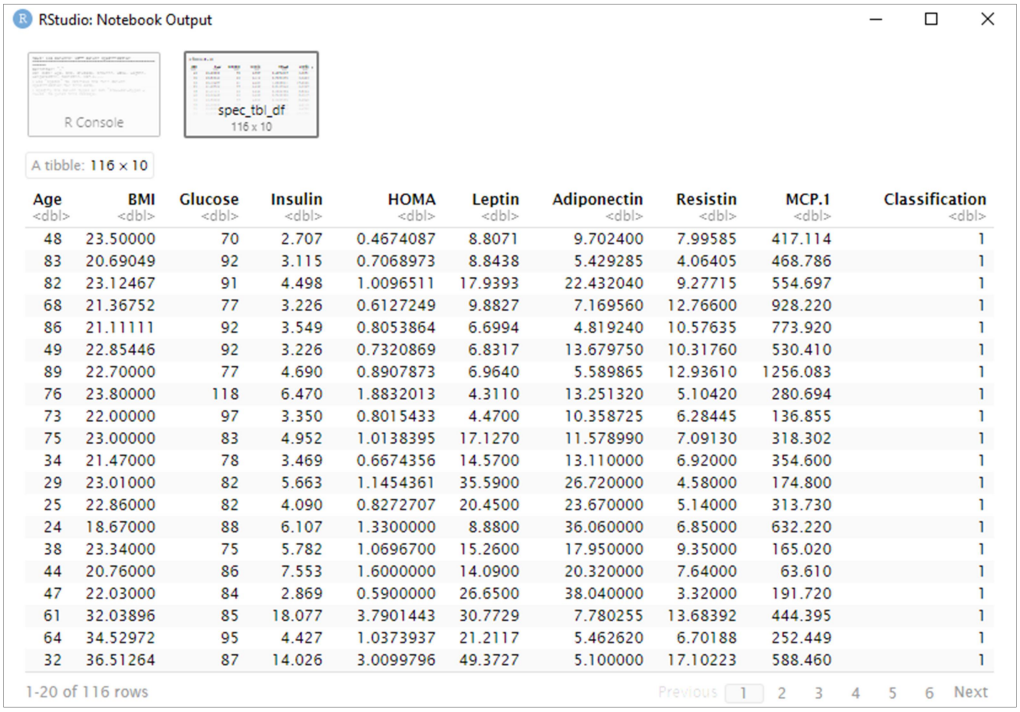
Iterasi	Jumlah Data Latih	Jumlah Data Uji	Keterangan
1	80%	20%	*Dataset* sebanyak 80% akan menjadi data latih (*data training*), sedangkan 20% sisanya akan menjadi data uji (*data testing*).
2	75%	25%	*Dataset* sebanyak 75% akan menjadi data latih (*data training*), sedangkan 25% sisanya akan menjadi data uji (*data testing*).
3	70%	30%	*Dataset* sebanyak 70% akan menjadi data latih (*data training*), sedangkan 30% sisanya akan menjadi data uji (*data testing*).
4	50%	50%	*Dataset* sebanyak 50% akan menjadi data latih (*data training*), sedangkan 50% sisanya akan menjadi data uji (*data testing*).
5	25%	75%	*Dataset* sebanyak 25% akan menjadi data latih (*data training*), sedangkan 75% sisanya akan menjadi data uji (*data testing*).

4.1.1 Input Data

Langkah yang dilakukan pertama kali adalah mengambil data mengenai kanker payudara yang diperoleh dari *website* resmi UCI *machine learning* dengan delapan atribut terpilih yakni berupa age, BMI, glukosa, insulin, HOMA, leptin, adiponectin, resistin, MCP.1, dengan total jumlah data sebanyak 116 data (UCI, 2018).

4.1.2 *Preprocessing*

Tahap selanjutnya yakni *preprocessing*, yang mana di dalamnya terdapat serangkaian proses yang mencakup seleksi data, pembersihan data, pembakuan format data, serta pelabelan pada data.


RStudio: Notebook Output

R Console

spec_tbl_df
116 x 10

A tibble: 116 × 10

Age <dbl>	BMI <dbl>	Glucose <dbl>	Insulin <dbl>	HOMA <dbl>	Leptin <dbl>	Adiponectin <dbl>	Resistin <dbl>	MCP.1 <dbl>	Classification <dbl>
48	23.50000	70	2.707	0.4674087	8.8071	9.702400	7.99585	417.114	1
83	20.69049	92	3.115	0.7068973	8.8438	5.429285	4.06405	468.786	1
82	23.12467	91	4.498	1.0096511	17.9393	22.432040	9.27715	554.697	1
68	21.36752	77	3.226	0.6127249	9.8827	7.169560	12.76600	928.220	1
86	21.11111	92	3.549	0.8053864	6.6994	4.819240	10.57635	773.920	1
49	22.85446	92	3.226	0.7320869	6.8317	13.679750	10.31760	530.410	1
89	22.70000	77	4.690	0.8907873	6.9640	5.589865	12.93610	1256.083	1
76	23.80000	118	6.470	1.8832013	4.3110	13.251320	5.10420	280.694	1
73	22.00000	97	3.350	0.8015433	4.4700	10.358725	6.28445	136.855	1
75	23.00000	83	4.952	1.0138395	17.1270	11.578990	7.09130	318.302	1
34	21.47000	78	3.469	0.6674356	14.5700	13.110000	6.92000	354.600	1
29	23.01000	82	5.663	1.1454361	35.5900	26.720000	4.58000	174.800	1
25	22.86000	82	4.090	0.8272707	20.4500	23.670000	5.14000	313.730	1
24	18.67000	88	6.107	1.3300000	8.8800	36.060000	6.85000	632.220	1
38	23.34000	75	5.782	1.0696700	15.2600	17.950000	9.35000	165.020	1
44	20.76000	86	7.553	1.6000000	14.0900	20.320000	7.64000	63.610	1
47	22.03000	84	2.869	0.5900000	26.6500	38.040000	3.32000	191.720	1
61	32.03896	85	18.077	3.7901443	30.7729	7.780255	13.68392	444.395	1
64	34.52972	95	4.427	1.0373937	21.2117	5.462620	6.70188	252.449	1
32	36.51264	87	14.026	3.0099796	49.3727	5.100000	17.10223	588.460	1

1-20 of 116 rows Previous 1 2 3 4 5 6 Next

Gambar 4.1 Input Data

Gambar 4.1 menunjukan data yang sudah diinputkan, yang mana di dalamnya terdapat 10 atribut. Kolom pada atribut *classification* merupakan kolom yang akan digunakan untuk memprediksi nilai pada penelitian ini, dengan artian kolom tersebut akan menjadi kelas *variable.* Sedangkan sembilan atribut atau kolom lainnya akan dijadikan sebagai input *variable*. Kelas *classification* akan digunakan untuk membentuk model *decision tree*. Pada algoritma *decision tree* tipe data pada kelas klasifikasi (*variable*) harus berupa *factor.* Apabila kolom pada kelas tersebut dibaca sebagai tipe data lain, maka harus dikonversi terlebih dahulu menjadi tipe data *factor*.

```
```{r}
df$Classification <- as.factor(df$Classification)
str(df)
```
```

Gambar 4.2 Source Code Konversi Tipe Data

Setelah kolom pada kelas *classification* berhasil dikonversi menjadi tipe data *factor,* maka kelas *variable* sudah terbentuk, dan untuk melihat tipe dan stuktur data diperlukan satu fungsi yakni *glimpse.* Sebagaimana ditunjukan oleh gambar 4.3.

```
'data.frame':   116 obs. of  10 variables:
 $ Age           : int  48 83 82 68 86 49 89 76 73 75 ...
 $ BMI           : num  23.5 20.7 23.1 21.4 21.1 ...
 $ Glucose       : int  70 92 91 77 92 92 77 118 97 83 ...
 $ Insulin       : num  2.71 3.12 4.5 3.23 3.55 ...
 $ HOMA          : num  0.467 0.707 1.01 0.613 0.805 ...
 $ Leptin        : num  8.81 8.84 17.94 9.88 6.7 ...
 $ Adiponectin   : num  9.7 5.43 22.43 7.17 4.82 ...
 $ Resistin      : num  8 4.06 9.28 12.77 10.58 ...
 $ MCP.1         : num  417 469 555 928 774 ...
 $ Classification: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Gambar 4.3 Tipe dan Struktur Data

Langkah selanjutnya adalah membagi menjadi dua bagian, yakni data latih (*training*) dan data uji (*testing*). *Source code* pembagian data latih serta data uji ditujukkan oleh gambar 4.4.

```r
#partition dan membagi data training dan testing
n <- round(nrow(df)*0.70);n
set.seed(116)
samp = sample(1: nrow(df),n)
trainData=df[samp,]
testData=df[-samp,]
```

Gambar 4.4 Source Code Pembangian Data

Pada *source code* tersebut fungsi set.seed adalah untuk menghasilkan dataset secara acak dengan menggunakan seluruh data yang digunakan, agar pembagian data latih dan data uji tidak berurutan.

Menghitung nilai entropy dari masing-masing atribut dilakukan dengan *source code* pada gambar 4.5 sebagaimana rumus pada persamaan (2.1).

```r
entropy <- function(target){
  freq <- table(target)/length(target)
  #vektorisasi kolom data frame
  vec <- as.data.frame(freq)[,2]
  #drop 0 to avoid NaN resulting from log2
  vec <- vec[vec>0]
  #menghitung nilai entropy
  -sum(vec * log2(vec))
}
#menghitung nilai entropy kolom classification
print(entropy(df$classification))
```

Gambar 4.5 Source Code Nilai Entropy Atribut

Kemudian langkah selanjutnya adalah mengetahui nilai *information gain* dari masing-masing atribut sebagaimana persamaan (2.2), *source code*

dalam menghitung nilai *information* gain pada penelitian ini ditunjukan oleh gambar 4.6.

```r
IG_numeric<-function(data, feature, target, bins=9) {
  #Hapus baris di mana fiturnya adalah NA
  data<-data[!is.na(data[,feature]),]
  #Menghitung entropi untuk induk(label data)
  e0<-entropy(data[,target])

  data$cat<-cut(data[,feature], breaks=bins, labels=c(1:bins))

  #gunakan dplyr untuk menghitung e dan p untuk setiap nilai fitur
  dd_data <- data %>% group_by(cat) %>% summarise(e=entropy(get(target)),
                n=length(get(target)),
                min=min(get(feature)),
                max=max(get(feature))
                )

  #hitung p untuk setiap nilai fitur
  dd_data$p<-dd_data$n/nrow(data)
  #menghitung IG
  IG<-e0-sum(dd_data$p*dd_data$e)

  return(IG)
}

IG_numeric(df, "Age", "Classification", bins=10)
```

Gambar 4.6 Source Code Nilai Informasi Gain Atribut

Setelah semua nilai *entropy* dan *information gain* dari masing-masing atribut telah ditemukan, maka hasil nilai *entropy* dan *information gain* ditunjukan oleh gambar 4.7 dibawah ini.

Description: df [10 × 3]

	Atribut <chr>	Entropy <dbl>	Information_Gain <dbl>
3	Glucose	5.3507992	0.17067961
1	Age	5.4332017	0.16933142
8	Resistin	6.8579810	0.12686298
4	Insulin	6.8062569	0.08097678
5	HOMA	6.8579810	0.08032743
9	MCP.1	6.7890155	0.07425622
7	Adiponectin	6.8407396	0.06069414
2	BMI	6.7545327	0.05883200
6	Leptin	6.8579810	0.02569859
10	Classification	0.9922666	0.00000000

1-10 of 10 rows

Gambar 4.7 Nilai Entropy dan Infromasi Gain Atribut

4.1.3 Membangun Model *Decision Tree*

Dalam membangun model *decision tree* digunakan beberapa *package* atau *library* pada penelitian ini, salah satu yang digunakan adalah *library* party yang mana fungsi dari *library* tersebut adalah untuk memperoleh model pohon klasifikasi (*classification tree*). Langkah selanjutnya adalah menginisialisasi terlebih dahulu model dari *decision tree*, pada *source code* yang ditunjukan oleh gambar 4.8 variabel target/respon adalah "*classification"* dan sembilan atribut yang menjadi variabel prediktornya dengan menggunkan data *training*.

```r
```{r}
#membuat model menggunakan library (rpart)
library(rpart)
pohon <- rpart(Classification~Age+BMI+Glucose+Insulin+HOMA+Leptin+Adiponectin+Resistin+MCP.1, data =
trainData, method = "class", control = rpart.control(minsplit = 0, cp=0))
pohon
```
```

Gambar 4.8 Source Code Inisialisasi Model

Library yang digunakan pada gambar 4.5 adalah *library* rpart yang mana fungsi dari *library* tersebut adalah untuk memperoleh model pohon klasifikasi (*classification tree*). Pada fungsi tersebut diperlukan opsi method = 'class' sehingga fungsi rpart akan mengenali variabel respon sebagai variabel kategorik dan fungsi tersebut akan menghasilkan pohon klasifikasi. Sehingga model pohon klasifikasi akan disimpan ke dalam objek dengan nama pohon, yang mana hasil yang didapatkan ditunjukan oleh gambar 4.9.

```
n= 81

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 81 32 2 (0.39506173 0.60493827)
    2) HOMA< 2.169985 49 22 1 (0.55102041 0.44897959)
      4) BMI>=30.0273 12  1 1 (0.91666667 0.08333333)
        8) MCP.1>=236.3445 10  0 1 (1.00000000 0.00000000) *
        9) MCP.1< 236.3445 2  1 1 (0.50000000 0.50000000)
         18) Age>=58.5 1  0 1 (1.00000000 0.00000000) *
         19) Age< 58.5 1  0 2 (0.00000000 1.00000000) *
      5) BMI< 30.0273 37 16 2 (0.43243243 0.56756757)
       10) Age>=74.5 8  0 1 (1.00000000 0.00000000) *
       11) Age< 74.5 29  8 2 (0.27586207 0.72413793)
         22) Adiponectin< 3.539117 3  0 1 (1.00000000 0.00000000) *
         23) Adiponectin>=3.539117 26  5 2 (0.19230769 0.80769231)
           46) Age< 37 2  0 1 (1.00000000 0.00000000) *
           47) Age>=37 24  3 2 (0.12500000 0.87500000)
             94) MCP.1< 192.795 3  1 1 (0.66666667 0.33333333)
              188) BMI>=20.56633 2  0 1 (1.00000000 0.00000000) *
              189) BMI< 20.56633 1  0 2 (0.00000000 1.00000000) *
             95) MCP.1>=192.795 21  1 2 (0.04761905 0.95238095)
              190) HOMA< 0.4876723 1  0 1 (1.00000000 0.00000000) *
              191) HOMA>=0.4876723 20  0 2 (0.00000000 1.00000000) *
    3) HOMA>=2.169985 32  5 2 (0.15625000 0.84375000)
      6) BMI>=36.13114 2  0 1 (1.00000000 0.00000000) *
      7) BMI< 36.13114 30  3 2 (0.10000000 0.90000000)
       14) Glucose< 90.5 6  3 1 (0.50000000 0.50000000)
         28) Leptin>=30.0234 3  0 1 (1.00000000 0.00000000) *
         29) Leptin< 30.0234 3  0 2 (0.00000000 1.00000000) *
       15) Glucose>=90.5 24  0 2 (0.00000000 1.00000000) *
```

Gambar 4.9 Hasil Klasifikasi

Pada hasil tersebut diketahui bahwa *rootnode* bersisi sebanyak 81 amatan yang terbagi menjadi dua bagian, yakni dengan aturan apakah HOMA < 2.169985 atau HOMA >= 2.169985. Node dengan HOMA < 2.169985 berisikan 49 amatan, sedangkan node HOMA >= 2.169985 berisikan 32 amatan. Selanjutnya node HOMA < 2.169985 terbagi lagi menjadi dua amatan yakni apakah BMI < 30.0273 atau BMI >= 30.0273, BMI < 30.0273 berisikan 37 amatan, sedangkan BMI >= 30.0273 berisikan kurang dari 30 amatan yakni hanya berisikan 12 amatan.

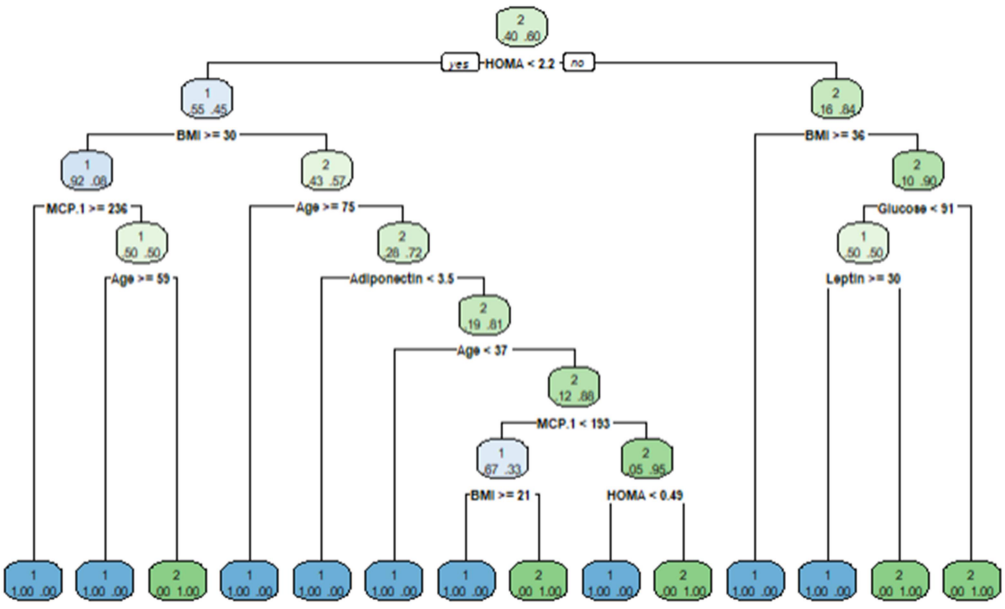
Jika diperhatikan pada baris awal *ouput*: node), split, n, loss, yval, (yprob) didefinisikan bahwa nilai 81 adalah n, nilai 32 adalah frekuensi amatan yang salah klasifikasi, 2 atau yval merupakan prediksi pada node tersebut yakni kelas yang lebih dominan, dan (0.39506173 0.60493827) adalah nilai dari masing-masing proporsi yval '2'. Bentuk pohon klasifikasi akan lebih mudah

dipahami apabila terbentuk dengan visualisasi yang baik, *package* yang dapat digunakan sebagai fungsi untuk membangun visualisasi pohon klasifikasi ditunjukan oleh gambar 4.10.

```r
```{r}
library(rpart.plot)
rpart.plot(pohon, extra = 4)
```
```

Gambar 4.10 Source Code Visualisasi Pohon Klasifikasi

Library yang digunakan untuk membangun visualisasi model pohon adalah rpart.plot yang mana fungsi pada *library* tersebut dapat digunakan untuk membuat visualisai pohon klasifikasi secara lebih baik dan mudah dipahami. Cara mengimplementasikan fungsi tersebut dengan menyebutkan nama kelas model yang sebelumnya sudah kita buat, dan memilih opsi nomor pilihan apa saja yang ingin ditampilkan pada setiap node. Hasil pohon klasifikasi ditunjukan oleh gambar 4.11.



Gambar 4.11 Hasil Pohon Klasifikasi

Sedangkan pada metode *decision tree* algoritma C5.0 langkah yang dilakukan pertama kali adalah menginisialisasi atribut yang akan digunakan, kemudian membuat model C5.0 dengan menggunakan *package/library* C50 dan tidak lupa untuk membagi data menjadi dua bagian, yakni data training dan data testing.

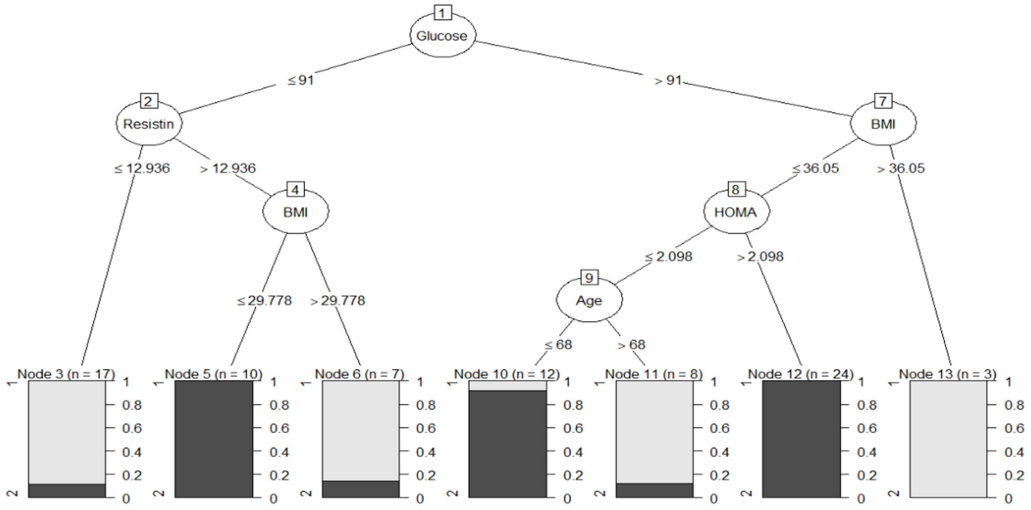
```r
set.seed(116)
testindexes <- sample(nrow(df), nrow(df)*0.7)
dataTest <- df[-testindexes, ]
dataTrain <- df[testindexes, ]
model <- C5.0(Classification~., data = dataTrain)
model
```

Gambar 4.12 Source Code Pembagian Data

Setelah data yang akan digunakan telah dibagi, maka langkah selanjutnya yang dilakukan maka langkah selanjutnya adalah menampilkan hasil model klasifikasi *decision tree* dengan menggunakan algoritma C5.0. *Source code* serta hasil model klasifikasi tersebut ditunjukan oleh gambar 4.13 dam 4.14.

```r
#Menampilkan hasil model decision tree c5.0
model
plot(model)
summary(model)
```

Gambar 4.13 Source Code Menampilkan Hasil Pohon Klasifikasi C5.0



Gambar 4.14 Hasil Klasifikasi Pohon dengan Algoritma C5.0

4.1.4 Evaluasi Sistem

Hasil prediksi dari model akan dibandingkan dengan data uji (*testing*). Tujuannya adalah untuk dapat mengukur tingkat akurasi dari sistem yang telah dikembangkan. Evaluasi sistem dilakukan dengan menggunakan *multiclass confusion matrix* karena penelitian ini termasuk dalam klasifikasi multikelas.

Pada tahap ini, terdapat aturan klasifikasi dimana nilai-nilai yang dibutuhkan dalam perhitungan *confusion matrix* adalah *True Positive, True Negative, False Positve,* dan *False Negative*, yang meliputi:

a. *True Positive* (TP) adalah data yang sebenarnya positif diprediksi dengan benar sebagai data positif.
b. *True Negative* (TN) adalah data yang sebenarnya negatif diprediksi dengan benar sebagai data negatif.

c. *False Positive* (FP) adalah data yang sebenarnya negatif salah diprediksi sebagai data positif.
d. *False Negative* (FN) adalah data yang sebenarnya positif salah diprediksi sebagai data negatif.

Apabila data aktual pada suatu kelas tidak konsisten dengan prediksinya, maka akan dimasukkan ke dalam nilai TN, yang merupakan jumlah dari kelas-kelas alternatif kelas yang tidak memiliki hubungan baik pada data aktual maupun prediksi. Dalam dua kelas tersebut, yang memiliki hubungan antara keduanya atau saling berkaitan akan masuk ke nilai FP dan FN. Apabila data aktual dan prediksi memiliki kelas yang sama, maka akan dimasukkan ke nilai TP, dan sisa dari kelas-kelas alternatif kelas yang tidak saling berkaitan dimasukkan ke nilai TN.

Pada tahap ini, dijelaskan tentang perhitungan pada *accuracy*, *precision*, *recall* dan *micro*-F1. Akurasi klasifikasi mengacu pada persentase data pengujian yang diklasifikasikan dengan benar oleh model. Jika akurasi klasifikasi dianggap memadai, maka model dapat digunakan untuk mengklasifikasikan set data di masa mendatang yang memiliki label kelas yang belum diketahui (Agarwal, 2014). Perhitungan *accuracy* ditunjukkan dalam persamaan (4.1).

$$accuracy = \frac{TP}{Total\ data\ testing} \tag{4.1}$$

Presisi atau yang dikenal juga sebagai *precision,* menggambarkan proporsi unit yang diprediksi sebagai positif oleh model yang juga benar-

benar positif dalam data yang sebenarnya. Presisi dapat diinterpretasikan sebagai tingkat kesesuaian antara permintaan informasi dan respons terhadap permintaan tersebut (Mayadewi & Rosely, 2015). *Precision* adalah hasil perhitungan yang menunjukan sejauh mana data uji diprediksi sebagai kelas positif yang benar-benar positif. Perhitungan pada *precision* ditunjukkan dalam persamaan (4.2).

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall adalah hasil perhitungan yang menunjukan sejauh mana semua data uji yang positif telah diprediksi dengan benar sebagai positif dalam klasifikasi. *Recall* juga dikenal sebagai *True Positive Rate* (TPR), sensitivitas, dan probabilitas deteksi (Grandini, Bagli, & Visani, 2020). Perhitungan pada *recall* ditunjukkan dalam persamaan (4.3).

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Dalam klasifikasi *multiclass* dimana setiap observasi hanya memiliki satu label, skor F1 yang dihitung dengan metode mikro (*micro-averaged* F1) sama dengan akurasi klasifikasi secara keseluruhan. Rumus *micro*-F1 ditunjukkan dalam persamaan (4.4). Seluruh kelas observasi digabungkan untuk menghitung presisi secara mikro (*micro-averaged precision*) dan *micro-averaged recall,* sehingga diperoleh nilai rata-rata harmonik *micro*-F1 (Zhang, Wang, & Zhao, 2015).

$$Micro\ F1 = \left(\frac{2 \times Recall \times Precision}{Recall + Precision}\right) \quad (4.4)$$

4.2 Hasil Uji Coba

Hasil dari proses uji coba klasifikasi yang dilakukan sesuai dengan pemaparan sub bab 4.1. Hasil uji coba tingkat akurasi pada skenario pengujian yang dilakukan mulai dari iterasi pertama hingga iterasi ke-lima dikelompokan dalam tabel 4.2.

Tabel 4.2 Hasil Akurasi pada Skenario Pengujian

No	Skenario Pengujian	Akurasi	
		***Iterative Dichotomiser-3* (ID3)**	**C5.0**
1	Skenario 1 (80%-20%)	69.57%	75.00%
2	Skenario 2 (75%-25%)	72.41%	68.97%
3	Skenario 3 (70%-30%)	77.14%	80.00%
4	Skenario 4 (50%-50%)	72.41%	63.79%
5	Skenario 5 (25%-75%)	68.57%	70.11%

Dari hasil pemaparan pada Tabel 4.2 tersebut nilai akurasi yang paling optimal didapatkan oleh skenario pengujian dengan pembagian data dengan perbandingan rasio yakni sebesar 70:30 dengan artian 70% dari total keseluruhhan data atau sebanyak 81 data digunakan sebagai data latih (*training*), sementara 30% % dari total keseluruhhan data atau sebanyak 35 data digunakan sebagai data uji (*testing*). Pembagian data pada penilitian ini menggunakan perbandingan Tabel 4.3 menunjukan data *training.*

Tabel 4.3 Data Training.

No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
1	29	32,270	84	5,81	1,203	45,619	6,209	24,603	904,981	1
2	66	36,212	101	15,533	3,869	74,706	7,539	22,320	864,968	1
3	86	21,111	92	3,549	0,805	6,699	4,819	10,576	773,92	1
4	69	28,444	108	8,808	2,346	14,748	5,288	16,485	353,568	2
5	51	22,892	103	2,74	0,696	8,016	9,349	11,554	359,232	2
6	62	26,84	100	4,53	1,117	12,45	21,42	7,32	330,16	2
7	58	29,154	139	16,582	5,685	22,888	10,262	13,973	923,886	2
8	44	27,887	99	9,208	2,248	12,675	5,478	23,033	407,206	2
9	54	36,05	119	11,91	3,495	89,27	8,01	5,06	218,28	2
10	49	20,956	94	12,305	2,853	11,240	8,412	23,117	573,63	2
11	60	26,349	103	5,138	1,305	24,299	2,194	20,253	378,996	1
12	24	18,67	88	6,107	1,33	8,88	36,06	6,85	632,22	1
13	71	27,915	104	18,2	4,668	53,499	1,656	49,241	256,001	2
14	66	31,238	82	4,181	0,845	16,224	4,267	3,291	634,602	1
15	86	27,18	138	19,91	6,777	90,28	14,11	4,35	90,09	2
16	28	35,855	87	8,576	1,840	68,510	4,7942	21,443	358,624	1
17	62	22,656	92	3,482	0,790	9,864	11,236	10,695	703,973	2
18	69	29,4	89	10,704	2,349	45,272	8,286	4,53	215,769	1
19	59	28,672	77	3,188	0,605	17,022	16,440	31,690	910,489	2
20	51	19,132	93	4,364	1,001	11,081	5,807	5,570	90,6	2
21	68	35,56	131	8,15	2,633	17,87	11,9	4,19	198,4	2
22	57	34,838	95	12,548	2,940	33,161	2,364	9,954	655,834	2
23	48	31,25	199	12,162	5,969	18,131	4,104	53,630	1698,44	2
24	35	35,250	90	6,817	1,513	50,609	6,966	22,037	667,928	1
25	47	22,03	84	2,869	0,59	26,65	38,04	3,32	191,72	1
26	75	23	83	4,952	1,0138	17,127	11,578	7,091	318,302	1
27	76	23,8	118	6,47	1,883	4,311	13,251	5,1042	280,694	1
28	81	31,640	100	9,669	2,385	38,806	10,636	29,558	426,175	2
29	89	22,7	77	4,69	0,890	6,964	5,589	12,936	1256,08	1
30	74	28,650	88	3,012	0,653	31,123	7,652	18,355	572,401	2
31	38	22,499	95	5,261	1,232	8,438	4,771	15,736	199,055	2
32	72	23,62	105	4,42	1,144	21,78	17,86	4,82	195,94	2
33	73	22	97	3,35	0,801	4,47	10,358	6,284	136,855	1
34	42	29,296	98	4,172	1,008	12,261	6,695	53,671	1041,84	2
35	86	26,667	201	41,611	20,630	47,647	5,357	24,370	1698,44	2
36	75	25,7	94	8,079	1,873	65,926	3,741	4,496	206,802	1
37	36	28,576	86	4,345	0,921	15,124	8,6	9,153	534,224	1
38	44	19,56	114	15,89	4,468	13,08	20,37	4,62	220,66	2
39	53	36,790	101	10,175	2,534	27,184	20,03	10,263	695,754	1
40	45	23,140	116	4,902	1,402	17,997	4,294	5,2633	518,586	2
41	40	30,836	128	41,894	13,227	31,038	6,160	17,556	638,261	2
42	49	21,367	78	2,64	0,507	6,334	3,886	22,942	737,672	2
43	72	25,59	82	2,82	0,5703	24,96	33,75	3,27	392,46	2
44	43	34,422	89	23,194	5,091	31,212	8,301	6,710	960,246	1
45	72	29,136	83	10,949	2,241	26,808	2,784	14,769	232,018	2
46	54	30,483	90	5,537	1,229	12,331	9,731	10,192	1227,91	1
47	60	31,231	131	30,13	9,736	37,843	8,404	11,501	396,021	2
48	82	23,124	91	4,498	1,009	17,939	22,432	9,277	554,697	1
49	66	26,562	89	6,524	1,432	14,908	8,429	14,919	269,487	2
50	45	21,303	102	13,852	3,485	7,647	21,056	23,034	552,444	2
51	68	21,082	102	6,2	1,559	9,699	8,574	13,742	448,799	2
52	48	28,125	90	2,54	0,563	15,532	10,223	16,110	1698,44	2
53	67	29,606	79	5,819	1,133	21,903	2,194	4,207	585,307	1
54	71	30,3	102	8,34	2,098	56,502	8,13	4,298	200,976	1
55	49	29,778	70	8,396	1,449	51,338	10,731	20,768	602,486	2
56	45	20,26	92	3,44	0,780	7,65	16,67	7,84	193,87	2
57	76	29,218	83	5,376	1,101	28,562	7,369	8,043	698,789	1
58	34	31,975	87	4,53	0,972	28,750	7,642	5,625	572,783	1
59	46	20,83	88	3,42	0,742	12,87	18,55	13,56	301,21	2
60	71	25,510	112	10,395	2,871	19,065	5,486	42,744	799,898	2
61	34	24,242	92	21,699	4,924	16,735	21,823	12,065	481,949	2
62	46	22,21	86	36,94	7,836	10,16	9,76	5,68	312	2
63	44	24,74	106	58,46	15,285	18,16	16,1	5,31	244,75	2
64	65	30,915	97	10,491	2,510	44,021	3,710	20,468	396,648	2

No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
65	52	30,801	87	30,212	6,483	29,273	6,268	24,245	764,667	2
66	85	27,688	196	51,814	25,050	70,882	7,901	55,215	1078,35	2
67	66	31,446	90	9,245	2,052	45,962	10,355	23,381	1102,11	1
68	45	20,83	74	4,56	0,832	7,752	8,237	28,032	382,955	2
69	48	23,5	70	2,707	0,467	8,807	9,702	7,995	417,114	1
70	46	33,18	92	5,75	1,304	18,69	9,16	8,89	209,19	2
71	45	29,384	90	4,713	1,046	23,847	6,644	15,556	621,273	2
72	61	32,038	85	18,077	3,790	30,772	7,780	13,683	444,395	1
73	83	20,690	92	3,115	0,706	8,843	5,429	4,064	468,786	1
74	51	27,688	77	3,855	0,732	20,092	3,192	10,375	473,859	1
75	50	38,578	106	6,703	1,752	46,640	4,667	11,783	887,16	1
76	45	26,85	92	3,33	0,755	54,68	12,1	10,96	268,23	2
77	35	30,276	84	4,376	0,906	39,213	9,048	16,437	733,797	1
78	69	35,092	101	5,646	1,406	83,482	6,796	82,1	263,499	1
79	48	32,461	99	28,677	7,002	46,076	21,57	10,157	738,034	2
80	75	30,48	152	7,01	2,628	50,53	10,06	11,73	99,45	2
81	43	26,562	101	10,555	2,629	9,8	6,420	16,1	806,724	2

Kemudian, proses selanjutnya yakni proses *learning* menggunakan *decision tree iterative dichotomiser-3* (ID3) yang diterapkan pada data *testing* yakni sebanyak 35 data atau 30% dari jumlah keseluruhan data. Adapaun daftar data *testing* ditunjukan dalam Tabel 4.4.

Tabel 4.4 Data Testing

No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
1	68	21,367	77	3,226	0,612	9,882	7,169	12,766	928,22	1
2	49	22,854	92	3,226	0,732	6,831	13,679	10,317	530,41	1
3	34	21,47	78	3,469	0,667	14,57	13,11	6,92	354,6	1
4	29	23,01	82	5,663	1,145	35,59	26,72	4,58	174,8	1
5	25	22,86	82	4,09	0,827	20,45	23,67	5,14	313,73	1
6	38	23,34	75	5,782	1,069	15,26	17,95	9,35	165,02	1
7	44	20,76	86	7,553	1,6	14,09	20,32	7,64	63,61	1
8	64	34,529	95	4,427	1,037	21,211	5,462	6,701	252,449	1
9	32	36,512	87	14,026	3,009	49,372	5,1	17,102	588,46	1
10	45	37,035	83	6,76	1,383	39,980	4,617	8,704	586,173	1
11	36	34,174	80	6,59	1,300	10,280	5,065	15,721	581,313	1
12	77	35,587	76	3,881	0,727	21,786	8,125	17,261	618,272	1
13	76	27,2	94	14,07	3,262	35,891	9,346	8,415	377,227	1
14	75	27,3	85	5,197	1,089	10,39	9,008	7,576	335,393	1
15	69	32,5	93	5,43	1,245	15,145	11,787	11,787	270,142	1
16	66	27,7	90	6,042	1,341	24,846	7,652	6,705	225,88	1
17	78	25,3	60	3,508	0,519	6,633	10,567	4,663	209,749	1
18	85	26,6	96	4,462	1,056	7,85	7,9317	9,613	232,006	1
19	76	27,1	110	26,211	7,112	21,778	4,935	8,493	45,843	1
20	77	25,9	85	4,58	0,960	13,74	9,753	11,774	488,829	1
21	42	21,359	93	2,999	0,687	19,082	8,462	17,376	321,919	2
22	69	21,513	112	6,683	1,846	32,58	4,138	15,698	713,239	2
23	59	22,832	98	6,862	1,658	14,903	4,230	8,204	355,31	2
24	54	24,218	86	3,73	0,791	8,687	3,705	10,344	635,049	2
25	64	22,222	98	5,7	1,377	12,190	4,783	13,912	395,976	2
26	51	18,37	105	6,03	1,561	9,62	12,76	3,21	513,66	2
27	55	31,975	92	16,635	3,775	37,223	11,018	7,165	483,377	2
28	43	31,25	103	4,328	1,099	25,781	12,718	38,653	775,322	2

No	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
29	41	26,672	97	22,033	5,271	44,705	13,494	27,832	783,796	2
30	65	29,665	85	14,649	3,071	26,516	7,282	19,463	1698,44	2
31	82	31,217	100	18,077	4,458	31,645	9,923	19,946	994,316	2
32	49	32,461	134	24,887	8,225	42,391	10,793	5,768	656,393	2
33	40	27,636	103	2,432	0,617	14,322	6,783	26,013	293,123	2
34	73	37,109	134	5,636	1,862	41,406	3,335	6,892	788,902	2
35	65	32,05	97	5,73	1,370	61,48	22,54	10,33	314,05	2

Dari data tersebut kemudian dilakukan proses klasifikasi dengan menggunakan model *decision tree iterative dichotomiser-3* (ID3) sehingga hasil prediksi klasifikasi ditunjukkan dalam tabel 4.5.

Tabel 4.5 Hasil Prediksi Klasifikasi ID3

No	Classification	prediksiId3
1	1	2
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	2
10	1	1
11	1	1
12	1	1
13	1	2
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	2
20	1	1
21	2	2
22	2	2
23	2	1

No	Classification	prediksiId3
24	2	1
25	2	2
26	2	1
27	2	2
28	2	1
29	2	2
30	2	2
31	2	2
32	2	2
33	2	2
34	2	1
35	2	1

Hasil prediksi dari klasifikasi diatas kemudian ditunjukan oleh *confussion matrix* pada gambar 4.15.

```
          aktual
testpred  1  2
       1 14  2
       2  6 13
```

Gambar 4.15 Hasil Prediksi Aloritma ID3

Penentuan hasil uji coba dari konfusi matrik diperlukan nilai *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) yang mana nilai-nilai tersebut didapatkan dari hasil klasifikasi pada sistem yang dibuat yang ditunjukan oleh gambar 4.12. Seluruh nilai tersebut diperlukan untuk proses perhitungan nilai *accuracy, precision, recall,* dan *micro* F1 pada *confusion matrix.* Pada Tabel 4.6 berisi konfusi matrix beserta deskripsi dari nilai-nilai tersebut.

Tabel 4.6 Cofussion Matrix Iterative dichotomiser -3 (ID3)

Confusion Matrix		**Aktual**	
		1	**2**
Prediksi	**1**	14 (TP)	2 (FP)
	2	6 (FN)	13 (TN)

Perhitungan nilai akurasi pada hasil klasifikasi tersebut kemudian dilakukan dengan persamaan (4.1) yakni sebagai berikut:

$$Accuracy = \frac{14 + 13}{35} = \frac{27}{35} = 0{,}7714 = 77{,}14\%$$

Selanjutnya adalah menghitung nilai presisi dari pengujian klasifikasi penyakit kanker payudara, perhitungan nilai presisi dilakukan dengan menggunakan persamaa (4.2) yakni sebagai berikut:

$$Precision = \frac{14}{14 + 2} \times 100\% = \frac{14}{16} \times 100\% = 87{,}5\%$$

Menghitung nilai *recall* dari pengujian klasifikasi penyakit kanker payudara, perhitungan nilai presisi dilakukan dengan menggunakan persamaan (4.3) yakni sebagai berikut:

$$Recall = \frac{14}{14 + 6} \times 100\% = \frac{14}{20} \times 100\% = 70\%$$

Menghitung nilai F1 *score* dari pengujian klasifikasi penyakit kanker payudara. Perhitungan nilai presisi dilakukan dengan menggunakan persamaan (4.4) yakni sebagai berikut:

$$\text{F1 } score = \frac{2 \times 70\% \times 87{,}5\%}{70\% + 87{,}5\%} = \frac{12{,}250}{157} = 78\%$$

Sedangkan hasil klasifikasi dengan menggunakan *decision tree* C5.0 ditunjukan oleh Tabel 4.7 dan *confussion matrix* ditunjukan oleh gambar 4.16.

Tabel 4.7 Hasil Prediksi Klasifikasi C5.0

No	Classification	prediksiC5
1	1	1
2	1	2
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	2
9	1	1
10	1	1
11	1	1
12	1	1
13	1	2
14	1	1
15	1	1
16	1	1
17	1	1
18	1	1
19	1	2
20	1	1
21	2	2
22	2	1
23	2	2
24	2	1
25	2	2
26	2	2
27	2	2
28	2	2

No	Classification	prediksiC5
29	2	2
30	2	2
31	2	2
32	2	2
33	2	2
34	2	1
35	2	2

```
                aktual
treeC5predict  1  2
            1 16  3
            2  4 12
```

Gambar 4.16 Hasil Prediksi dengan Algoritma C5.0

Tabel 4.8 Cofussion Matrix Algoritm C5.0

Confusion Matrix		**Aktual**	
		1	**2**
Prediksi	**1**	16 (TP)	3 (FP)
	2	4 (FN)	12 (TN)

Pada tabel 4.8 berisi konfusi matrix pada algoritma C5.0 beserta deskripsi dari nilai-nilai pada tiap prediksi tersebut. Perhitungan nilai akurasi pada hasil klasifikasi tersebut kemudian dilakukan dengan persamaan (4.1) yakni sebagai berikut:

$$Accuracy = \frac{16 + 12}{35} = \frac{28}{35} = 0{,}8 = 80\%$$

Selanjutnya adalah menghitung nilai presisi dari pengujian klasifikasi penyakit kanker payudara. Perhitungan nilai presisi dilakukan dengan menggunakan persamaa (4.2) yakni sebagai berikut:

$$Precision = \frac{16}{16+3} \times 100\% = \frac{16}{19} \times 100\% = 84{,}2\%$$

Menghitung nilai *recall* dari pengujian klasifikasi penyakit kanker payudara, perhitungan nilai presisi dilakukan dengan menggunakan persamaa (4.3) yakni sebagai berikut:

$$Recall = \frac{16}{16+4} \times 100\% = \frac{16}{20} \times 100\% = 80\%$$

Menghitung nilai F1 *score* dari pengujian klasifikasi penyakit kanker payudara. Perhitungan nilai presisi dilakukan dengan menggunakan persamaan (4.4) yakni sebagai berikut:

$$\text{F1 } score = \frac{2 \times 84{,}2\% \times 80\%}{84{,}2\% + 80\%} = \frac{13{,}472}{164{,}2} = 82\%$$

4.3 Pembahasan

Pembahasan pada sub bab ini merupakan hasil yang didapat dari hasil uji coba yang yang telah dilakukan pada dua algoritma pada metode *decision tree* yakni algoritma *iterative dichotomiser-3* (ID3) dan C5.0, dapat ditarik kesimpulan bahwasanya dari kedua algoritma dari metode *decision tree* tersebut tidak ada yang mengungguli jauh diatas algoritma yang lain. Hal ini disebabkan karena kedua algoritma tersebut algoritma tergolong ke dalam jenis metode *supervised*

learning, yang mana metode *supervised learning* sendiri merupakan sebuah pendekatan *machine learning* dimana data yang digunakan sudah diberi label dan diketahui oleh perancang *dataset* tersebut (Osisanwo et al., 2017). Sehingga model prediktif pada hasil prediksi yang dihasilkan oleh algoritma *iterative dichotomiser-3* (ID3) dan C5.0 secara lebih spesifik dan akurat.

Hasil evaluasi yang telah dilakukan pada uji coba skenario pertama hingga kelima dengan menggunakan ratio yang berbeda-beda pada data latih (*training*) dan data uji (*testing*), serta tidak saling beririsan. Dapat dilihat bahwasanya algoritma *iterative dichotomiser-3* (ID3) memiliki tingkat nilai akurasi yang konsisten pada ketiga skenario uji coba, yang berkisar 72.41%-77.14%. Hal ini disebabkan karna algoritma *iterative dichotomiser-3* (ID3) sulit memprediksi model pohon klasifikasi dengan data uji (*testing*) yang berubah-ubah secara signifikan dan bervariasi seperti pada skenario uji coba kedua sampai keempat, hal inilah yang menyebabkan hasil kinerja pada algoritma tersebut lebih konsisten, karena banyaknya data latih yang digunakan tidak mempengaruhi nilai akurasi. Serta pohon keputusan yang dihasilkan cenderung sederhana dan lebih umum. Sedangkan pada algoritma C5.0, nilai akurasi yang didapatkan lebih bervariasi pada tiap uji coba skenario. Hal ini dikarenakan algoritma C5.0 lebih memiliki kecenderungan untuk mempelajari pola yang sangat spesifik pada data latih (*training*) yang digunakan. Hal inilah yang menyebabkan hasil pohon keputusan pada algoritma C5.0 lebih sederhana jika dibandingkan dengan ID3, karena algoritma C5.0 akan melakukan pemangkasan (pruning) terhadap cabang-cabang yang tidak signifikan dari pohon keputusan untuk menghindari mempelajari pola

yang terlalu spesifik pada data pelatihan. Algoritma C5.0 memiliki lebih banyak hyperparameter dan aturan (*rule*) yang dapat mempengaruhi hasil akurasi. Jika ratio pembagian data yang digunakan berbeda dalam setiap skenario pengujian, maka hasil akurasi C5.0 dapat bervariasi secara signifikan.

Dari skenario uji coba tersebut ratio pembagian data yang digunakan pada penlitian ini hasil paling optimal didapatkan oleh skenario pengujian ketiga dengan ratio perbandingan sebesar 70:30. Ratio tersebut merupakan ratio paling optimal untuk dataset yang berukuran kecil (Raschka, 2018). Sehingga hasil evaluasi pada sistem yang dibuat dengan menggunakan konfusi matrik pada ratio perbandingan tersebut menunjukan bahwa pada algoritma *iterative dichotomiser-3* (ID3) sistem yang dibangun mendapatkan hasil nilai akurasi sebesar 73,5%, nilai *precision* sebesar 69,2%, nilai *recall* sebesar 64,2%, dan nilai F1 *score* sebesar 66,6%. Sedangkan hasil evaluasi pada sistem yang dibangun dengan menggunakan algoritma C5.0, menunjukan hasil bahwa sistem yang dibangun mampu melakukan proses klasifikasi terhadap kategori jenis kanker pada dataset kanker payudara dengan nilai akurasi sebesar 80%, nilai *precission* sebesar 84,2%, nilai *recall* sebesar 80% dan F1 *score* sebesar 82%.

Nilai klasifikasi pada sebuah sistem memiliki beberapa skala kelompok, dimana skala 90% - 100% merupakan kategori kelompok klasifikasi sangat bagus, skala 80% - 90% merupakan kategori kelompok klasifikasi baik, rentang skala dari 70% - 80% merupakan kelompok klasifikasi dengan kategori sedang, selanjutnya adalah nilai rentang skala 60% - 70% merupakan kategori kelompok klasifikasi buruk, sedangkan skala terakhir yakni rentang nilai 50% - 60%

dikategorikan sebagai kelompok klasifikasi yang gagal (Gorunescu, 2011). Dilihat dari skenario uji coba ketiga perolehan nilai akurasi, *precission, recall,* dan *micro* F1 model klasifikasi penyakit kanker payudara dengan menggunakan algoritma *iterative dichotomiser-3* (ID3) dikategorikan sebagai sistem yang cukup baik dalam melakukan klasfikasi. Sedangkan pada algoritma C5.0 berdasarkan perolehan nilai akurasi, *precission, recall,* dan *micro* F1 menunjukan bahwa algoritma tersebut dikategorikan sebagai sistem yang mampu dalam mengklasifikan penyakit kanker payudara dengan baik. Berdasarkan pemaparan hasil nilai akurasi tersebut dapat dinyatakan bahwa sistem yang dibangun dengan mengunakan algoritma C5.0 lebih baik dalam melakukan pengklasifikasian dibandingkan dengan algoritma *iterative dichotomiser-3* (ID3) sehingga model *decision tree* C5.0 dapat diterapkan untuk melakukan prediksi deteksi penyakit kanker payudara.

Dengan adanya sistem yang dibuat diharapkan dapat membantu mengklasifikan jenis penyakit kanker payudara kedalam kategori yang sudah ada, sehingga hal ini dapat djadikan sebagai pendukung hipotesis oleh pihak rumah sakit dengan tujuan agar dapat memperkuat hasil laboratorium yang sudah ada. Sistem yang dibangun diharapkan dapat bermanfaat dan menjadi model evaluasi bagi pihak rumah sakit Dalam islam memberikan manfaat sesama manusia serta tolong menolong merupakan sesuatu yang baik, sebagaimana firman Allah SWT dalam QS Al-Israa ayat 7 yang berbunyi:

اِنْ اَحْسَنْتُمْ اَحْسَنْتُمْ لِاَنْفُسِكُمْ ۙوَاِنْ اَسَأْتُمْ فَلَهَا ۗ

"*Jika kamu berbuat baik (berarti) kamu berbuat baik bagi diirmu sendiri dan jika kamu berbuaut jahat, maka (kejahatan) itu bagi dirimu sendiri…….*" (QS. Al-Israa: 7).

Menurut tafsir Jalalain maksud dari ayat pada potongan ayat QS. Al-Israa ayat 7 tersebut adalah kemudian kami katakan (apabila kalian berbuat baik) dengan mengerjakan ketaatan (artinya kamu berbuat baik untuk dirimu sendiri) karena sesungguhnya pahala dari kebaikan yang kalian lakukan itu untuk diri kalian sendiri (dan apabila kalian berbuat jahat) dengan menimbulkan kerugian (maka kerugian itu bagi diri kalian sendiri). Dari tafsir ayat tersebut dapat disimpulkan bahwa sudah selayaknya sebagai orang mukmin untuk berbuat baik dan bermanfaat kepada orang lain, siapapun dan apapun yang berada disekitar, karena dari setiap perbuatan baik tersebut tentunya akan kembali kepada diri sendiri. Hal ini selaras dengan sabda Nabi Muhammad SAW dalam hadist yang diriwayatkan oleh Ahmad, at-Thabrani, ad-Daruqutni dan dihasankan oleh imam al-Albani dalam Shahihul Jami' no:3289 yang berbunyi:

خَيْرُ النَّاسِ أَنْفَعُهُمْ لِلنَّاسِ

"*Sebaik-baik manusia adalah yang paling bemanfaat bagi manusia lainnya*" (HR. Ahmad, at-Thabrani, ad-Daruqutni dan dihasankan oleh imam al-Albani dalam Shahihul Jami' no:3289).

Dari ayat dan hadist tersebut sebagai mukmin yang baik sudah seaharusnya berbuat baik dan bermanfaat bagi sekitar. Oleh karenanyam peneliti melakukan penelitian sini serta memamaparkan hasil yang telah didapatkan mengenai tingkat akurasi dari penerapan metode *decision tree* dalam melakukan klasifikasi penyakit kanker payudara. Sehingga, diharapakan pembaca yang

menggunakan penelitian ini dapat dijadikan sebagai referensi atau melanjutkan kembali penelitian ini agar menjadi bermanfaat.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan dari penelitian yang telah dilakukan pada prediksi deteksi penyakit kanker payudara dengan menggunakan metode *decision tree.* Uji coba yang telah dilakukan pada bab sebelumnya bertujuan untuk mengetahui keberhasilan dari implementasi metode pada sistem yang dibuat serta untuk mengevaluasi hasil yang didapat dari uji coba untuk mendapatkan kesimpulan dan saran.

Berdasarkan hasil perbandingan akurasi dari dua algoritma pada metode *decision tree* yakni *iterative dichotomiser-3* (ID3) dan C5.0 secara umum diambil dari semua skenario pengujian yang telah dilakukan, maka algoritma C5.0 dengan perolehan nilai akurasi sebesar 80% mengunggungguli algortima *iterative dichotomiser-3* (ID3) dengan hasil nilai akurasi sebesar 77,14%. Algortima C5.0 memiliki keunggulan dalam kompleksitas model yan dihasilkan dengan menggunakan teknik pruning sehingga dapat meningkatkan kinerja dan akurasi model. Sementara untuk hasil lain dari *iterative dichotomiser-3* (ID3) nilai presisi yang didapat sebesar 87,5%, *recall* sebesar 70%, dan F1 *score* sebesar 78%. Sementara pada algoritma C5.0 nilai presisi sebesar 84,2%, *recall* sebesar 80%, dan nilai F1 *score* sebesar 82%. Hasil tersebut menunjukan bahwa nilai akurasi, presisi, *recall*, dan F1 *score* pada C5.0 termasuk kedalam kategori baik. Sehingga dapat ditarik kesimpulan bahwa pemodelan sistem dengan menggunakan

algoritma C5.0 memiliki tingkat akurasi yang lebih baik dibandingkan dengan algoritma *iterative dichotomiser-3* (ID3).

5.2 Saran

Berdasarkan hasil uji coba yang dilakukan pada penelitian ini, diharapkan pada penelitian selanjutnya hasil akurasi, presisi, *recall*, dan F1 *score* dapat ditingkatkan pada setiap model pengujian. Berikut merupakan saran yang diharapkan oleh peniliti sebagai pendukung untuk penelitian selanjutnya:

1. Jumlah *dataset* yang digunakan pada penelitian ini berukuran kecil, hasil yang didapat akan lebih presisi apabila menggunakan sumber data yang lebih banyak lagi.
2. Menerapkan metode optimasi agar hasil nilai akurasi yang didapatkan lebih optimal.
3. Metode yang digunakan dapat ditambah dengan metode lain, agar dapat menjadi perbandingan penelitian sehingga hasil yang didapat lebih bervariasi.

DAFTAR PUSTAKA

Adinugroho, S., & Sari, Y. A. (2018). *Implementasi Data Mining Menggunakan Weka* (1st Edition). Malang: UB Press.

Agarwal, S. (2014). Data mining: Data mining concepts and techniques. *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Austria, Y. D., Goh, M. L., Sta. Maria Jr., L., Lalata, J.-A., Goh, J. E., & Vicente, H. (2019). Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. *International Journal of Simulation: Systems, Science & Technology*, (March 2021). https://doi.org/10.5013/ijssst.a.20.s2.23

Defiyanti, S., & Pardede, D. L. C. (2008). *Perbandingan Kinerja Algoritma ID3 dan C4.5 dalam Klasifikasi Spam-Mail*.

Defriani, M., & Jaelani, I. (2020). Algoritma J48 Dan Logistic Model Tree Untuk Memprediksi Predikat Kelulusan Mahasiswa: Studi Kasus STT XYZ. *Journal of Information Technology and Computer Science*, *3*(2), 129–140. https://doi.org/10.31539/intecoms.v3i2.1478

Elmande, Y., & Widodo, P. P. (2016). Pemilihan Criteria Splitting dalam Algoritma Iterative Dichotomiser 3 (ID3) untuk Penentuan Kualitas Beras: Studi Kasus Pada Perum Bulog Divre Lampung. *Jurnal Telematika MKOM*, *4*(1), 73-82.

Freedman, D. A. (2010). *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press.

Ghani, M. U., Alam, T. M., & Jaskani, F. H. (2019). Comparison of Classification Models for Early Prediction of Breast Cancer. *3rd International Conference on Innovative Computing, ICIC 2019*, (January 2020). https://doi.org/10.1109/ICIC48496.2019.8966691

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756.

Han, J., Kamber, M., & Pei, J. (2012). Third Edition : Data Mining Concepts and Techniques. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Hazra, R., Banerjee, M., & Badia, L. (2020). Machine Learning for Breast Cancer Classification with ANN and Decision Tree. *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*, 522–527. https://doi.org/10.1109/IEMCON51383.2020.9284936

Higa, A. (2018). Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms. *International Journal of Computer Applications Technology and Research*, *7*(1), 23–27. https://doi.org/10.7753/ijcatr0701.1004

Imaduddin, H., Hermansyah, B. A., & Salsabilla B, F. A. (2021). Comparison of Support Vector Machine and Decision Tree Methods in the Classification of Breast Cancer. *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, *5*(1), 22. https://doi.org/10.22373/cj.v5i1.8805

Keerthika, J., Sruthi, D., Swathi, D., Swetha, S., & Vinupriya, R. (2021). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 1530–1535. https://doi.org/10.1109/ICACCS51430.2021.9442043

Mayadewi, P., & Rosely, E. (2015). Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining. *Seminar Nasional Sistem Informasi Indonesia*, (November), 329–334.

Musa, A. A., & Aliyu, U. M. (2018). Application of machine learning techniques in predicting of breast Cancer metastases using decision tree algorithm. *Sokoto Northwestern Nigeria. J Data Mining Genomics Proteomics*, *11*(220), 2153-0602.

Naveen, Sharma, R. K., & Ramachandran Nair, A. (2019). Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models. *2019 4th IEEE International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2019 - Proceedings*, 100–104. https://doi.org/10.1109/RTEICT46194.2019.9016968

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal SIMETRIS*, *10*(2), 421–432.

Osisanwo, Akinsola, O, A., J. O, H., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Pribadi, D., Athiry, S., Saputra, R. A., Supiandi, A., & Prayudi, D. (2018). Sistem Pakar Diagnosa Penyakit Demam Berdarah Dengue Menggunakan Algoritma

Iterative Dichotomiser 3 (ID3). *Seminar Nasional Inovasi Dan Tren (SNIT)*, *3*(1), 129–133.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. Retrieved from http://arxiv.org/abs/1811.12808

Sathiyanarayanan, P., Pavithra, S., Sai Saranya, M., & Makeswari, M. (2019). Identification of breast cancer using the decision tree algorithm. *2019 IEEE International Conference on System, Computation, Automation and Networking, ICSCAN 2019*, 1–6. https://doi.org/10.1109/ICSCAN.2019.8878757

Snow, C. P. (2012). *The Two Cultures*. Cambridge University Press.

Suhartono, Kurniawan, F., & Imran, B. (2018). Identification of virtual plants using bayesian networks based on parametric L-system. *International Journal of Advances in Intelligent Informatics*, *4*(1), 40–52. https://doi.org/10.26555/ijain.v4i1.157

Sunjana. (2010). Aplikasi Mining Data Mahasiswa dengan Metode Klasifikasi Decision Tree. *Seminar Nasional Aplikasi Teknologi Informasi*, *2010*(Snati), 1–6.

Tarawneh, O., Otair, M., Husni, M., Abuaddous, H. Y., Tarawneh, M., & Almomani, M. A. (2022). Breast Cancer Classification using Decision Tree Algorithms. *International Journal of Advanced Computer Science and Applications*, *13*(4), 676–680. https://doi.org/10.14569/IJACSA.2022.0130478

UCI. (2018). Breast Cancer Coimbra Dataset. Retrieved September 12, 2022, from UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Wahyudin. (2009). Metode Iterative Dichotomizer 3 (ID3) Untuk Penyeleksian Penerimaan Mahasiswa Baru. *Jurnal Pendidikan Teknologi Informasi Dan Komunikasi (PTIK)*, *2*(2), 5–15.

Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment: A Review. *JAMA - Journal of the American Medical Association*, *321*(3), 288–300. https://doi.org/10.1001/jama.2018.19323

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76-77.

Zhang, D., Wang, J., & Zhao, X. (2015). Estimating the uncertainty of average F1 scores. *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval*, 317–320. https://doi.org/10.1145/2808194.2809488.