KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *DECISION TREE* DAN *RANDOM FOREST*

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
Sarjana Sains Bidang Matematika

Oleh:

SITI KALIMAH
NIM. 08011181722063

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SRIWIJAYA
2022

LEMBAR PENGESAHAN

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *DECISION TREE* DAN *RANDOM FOREST*

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
Sarjana Sains Bidang Matematika

Oleh

SITI KALIMAH
NIM. 08011181722063

Indralaya, Juli 2022

Pembimbing Kedua

Novi Rustiana Dewi, M.Si
NIP.197011131996032002

Pembimbing Utama

Dr. Yulia Resti, M.Si
NIP.197307191997022001

Mengetahui,
Ketua Jurusan Matematika

Drs. Sugandi Yahdin,M.M
NIP. 195807271986031003

PERNYATAAN INTEGRITAS

Yang bertanda tangan di bawah ini :

Nama : Siti Kalimah

NIM : 08011181722063

Judul : Klasifikasi Penyakit Diabetes Menggunakan Metode *Decision Tree* Dan *Random Forest*

Menyatakan dengan sesungguhnya bahwa seluruh data dan informasi yang dimuat dalam hasil penelitian ini, kecauli yang disebutkan dengan jelas sumbernya adalah hasil pengamatan dan investigasi saya sendiri diabawah supervisi pembimbing dan belum pernah atau tidak sedang diajukan sebagai syarat untuk memperoleh gelar kesarjanaan lain. Apabila dikemudian hari ditemukan adanya unsur plagiasi, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya.

Demikian pernyataan ini saya buat dalam keadaan sadar dan tidak mendapat paksaan dari pihak manapun.

Indralaya, Juli 2022

Siti Kalimah

HALAMAN PERSEMBAHAN

MOTTO

"Berusaha lebih dari yang lain jika ingin mendapatkan hasil lebih dari yang lain dan ingat satu hal jangan pernah bandingkan dirimu dengan orang lain karena dirimu tetap dan akan selalu istimewa sebagaimana dirimu sendiri"

Skripsi ini ku persembahkan kepada:

1. **Allah SWT**
2. **Bapak dan Mamak**
3. **Saudaraku**
4. **Keluarga Besar**
5. **Dosen**
6. **Almamater**
7. **Sahabat dan temanku**

KATA PENGANTAR

Assalamu'alaikum warrahmatullahi wabarakatuh

Puji syukur atas kehadirat Allah SWT yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "**Klasifikasi Penyakit Diabetes Menggunakan Metode *Decision Tree* dan *Random Forest***". Shalawat beserta salam semoga tetap tercurahkan kepada bimbingan kita, nabi agung nabi Muhammad SAW, beserta keluarga dan para pengikutnya.

Ucapan terima kasih yang sebesar-besarnya penulis sampaikan kepada orang tua tercinta, yaitu **bapak Sukarji** dan **ibu Jumaroh** yang selalu mendoakan dan selalu memberikan dukungan serta semangat untuk penulis sehingga penulis mampu menyelesaikan masa studi kuliah ini. Penulisan ini tidak lepas dari bantuan beberapa pihak, oleh Karena itu penulis ingin menyampaikan rasa terima kasih kepada :

1. **Kemenristekdikti** selaku pemberi beasiswa bidikmisi selama penulis menempuh masa perkuliahan hingga penulis mampu merasakan menjadi seorang mahasiswa dan dapat menyelesaikan jenjang Strata Satu di Jurusan Matematika Universitas Sriwijaya.
2. **Bapak Drs. Sugandi Yahdin, M.M** selaku ketua jurusan Matematika FMIPA universitas sriwijaya untuk ilmu yang telah diberikan.

3. **Ibu Dr. Dian Cahyawati Sukanda, M.Si** selaku sekertaris jurusan yang telah membantu proses administrasi pendaftaran seminar, dan juga atas ilmu yang diberikan.
4. **Ibu Dr. Yulia Resti, M.Si** selaku dosen pembimbing utama yang telah memebrikan saran, kritikan, masukan, membimbing dan meluangkan waktu sehingga penulis mampu menyelsaikan skripsi dengan baik.
5. **Ibu Novi Rustiana Dewi, M.Si** selaku pembimbing kedua yang telah memberikan saran, masukan, dan membimbing penulis hingga mampu menyelesaikan skripsi dengan baik.
6. **Bapak Drs. Robinson Sitepu, M.Si** selaku ketua pelaksana seminar dan **Bapak Drs. Endro Setyo C, M.Si** selaku sekertaris pelaksana seminar yang telah membantu proses seminar sehingga seminar dapat dilaksanakan dengan baik.
7. **Ibu Dr.Ir. Herlina Hanum, M.Si** selaku pembahas 1 dan **Ibu Irmeilyana, M.Si** selaku pembahas 2 yang telah memberikan masukan, kritikan dan saran sehingga penulis mampu menyelesaikan skripsi dengan baik.
8. **Bapak Dr. Bambang Suprihatin M.Si** selaku pembimbing akademik yang telah memberikan arahan dan saran selama masa perkuliahan berlangsung.
9. **Seluruh Dosen di jurusan Matematika FMIPA UNSRI** yang telah memberikan ilmu nya kepada penulis selama perkuliahan di jurusan Matematika.

10. **Bapak Irwansyah dan ibu Hamida** yang telah membantu proses administrasi selama di jurusan Matematika FMIPA.
11. Tim hebat, tiada duanya **Abu, Agung, Yudha, Rendy, Wawan, Shohif, Oliv, Mega, Azizah, Tesya, Muflika** yang saling menguatkan satu sama lain, tidak pernah meninggalkan dan selalu siap untuk saling membantu satu sama lain.
12. **Teman-teman angkatan 17** jurusan Matematika FMIPA yang selalu memberikan tawa dan cerita yang tak pernah terlupakan dan akan menjadi sebuah memori kenangan abadi dimana kelak akan penulis ceritakan kepada semua orang bagaimana kisah kasih masa kuliah penulis dulu.
13. **Kakak tingkat 2015, 2016 serta adik tingkat 2018, 2019** atas semua cerita dan momen kebersamaanya.
14. **Semua pihak** yang terlibat dan terkait dalam hidup penulis yang tak dapat disebutkan satu persatu.

Indralaya, 2022

Siti kalimah
NIM. 08011181722063

CLASSIFICATION OF DIABETES USING DECISION TREE AND RANDOM FOREST METHOD

By:

Siti Kalimah

08011181722063

ABSTRACT

Diabetes is a metabolic disease characterized by hyperglycemia caused by defects in insulin secretion or reduced insulin production, slow insulin action or both. Diabetes is one of the deadliest diseases in the world. Accurate classification of people who have positive or negative laboratory test results have diabetes is important to get the right treatment. The purpose of this study is to classify the status of people who have positive or negative laboratory test results for diabetes using the Decision Tree C4.5 and Random Forest methods. In this study used data taken from kaggle.com. This data has a size of 520 and 17 variables. The variables are Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity, class. The results of this study indicate the level of accuracy, precision, recall, specificity, and F1 score on the Decision Tree C4.5 method respectively of 91.35%, 93.55%, 92.06%, 90.24%, and 92.80%. By using the Random Forest method, the accuracy, precision, recall, specificity, and F1 score levels respectively 98.08%, 100%, 96.88%, 100%, and 98.41%. Based on these measures, it is concluded that the Random Forest method is better than the Decision Tree C4.5 method in classifying the status of people who have positive or negative laboratory test results for diabetes.

Keywords: Diabetes, Decision Tree C4.5, Random Forest

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *DECISION TREE* DAN *RANDOM FOREST*

By:

Siti Kalimah

08011181722063

ABSTRAK

Diabetes merupakan suatu penyakit metabolik dengan karakteristik hiperglikemia yang disebabkan oleh kelainan sekresi insulin atau berkurangnya produksi insulin, kerja insulin yang lambat atau bisa karena kedua-duanya. Penyakit diabetes menjadi salah satu jenis penyakit yang mematikan di dunia. Pengklasifikasian secara tepat orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes penting dilakukan untuk memperoleh penanganan yang tepat. Tujuan peneitian ini adalah mengklasifikasi status orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes menggunakan metode *Decision Tree* C4.5 dan *Random Forest.* Pada penelitian ini digunakan data yang diambil dari *kaggle.com*. Data ini memiliki ukuran 520 dan 17 variabel. Variabel-variabel tersebut adalah *Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity, class.* Hasil penelitian ini menunjukkan tingkat akurasi, presisi, recall, specificity, dan F1 score pada metode *Decision Tree* C4.5 secara berturut-turut sebesar 91.35%, 93.55%, 92.06%, 90.24%, dan 92.80%. Dengan menggunakan metode *Random Forest* diperoleh tingkat akurasi, presisi, recall, specificity, dan F1 score secara berturut-turut sebesar 98.08%, 100%, 96.88%, 100%, dan 98.41%. Berdasarkan ukuran-ukuran ini disimpulkan bahwa metode *Random Forest* lebih baik daripada metode *Decision Tree* C4.5 dalam mengklasifikasi status orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes.

Kata kunci: Diabetes, *Decision Tree* C4.5, *Random Forest*

DAFTAR ISI

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Menurut *American Diabetes Association* dalam peneletian (Zhang & Tan, 2000) Diabetes merupakan suatu penyakit metabolik dengan karakteristik hiperglikemia yang disebabkan oleh kelainan sekresi insulin atau berkurangnya produksi insulin, kerja insulin yang lambat atau bisa karena kedua-duanya. Setiap tahun pasien positif diabetes mengalami peningkatan yang cukup signifikan. Menurut Organisasi Kesehatan Dunia pasien penderita diabetes terus mengalami peningkatan hingga mampu mencatat angka mencapai 422 juta orang di dunia yang mana hal ini menjadi empat kali lipat dari pada 30 tahun lalu (Najib et al., 2019).

Ciri-ciri umum orang yang memiliki penyakit diabetes adalah sering buang air kecil, gampang merasakan haus, merasa cepat lapar, mengalami penurunan berat badan secara drastis, kulit menjadi kering dan gatal, luka yang sulit untuk sembuh, mengalami gangguan penglihatan, mengalami kesemutan atau kebas, mudah terjadi pembengkakan pada kaki dan tangan, badan lemas serta mengalami sakit kepala yang berlebihan, obesitas, dan infeksi jamur atau bakteri. Pengklasifikasian secara tepat orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes penting dilakukan untuk memperoleh penanganan yang tepat.

Terdapat beberapa penelitian terkait dengan metode yang akan digunakan dalam penelitian ini.

Tabel 1.1 Tinjauan Studi Terdahulu

Sumber/judul	Deskripsi	Algoritma
(Apriyani & Kurniati, 2020) / Perbandingan Metode Naïve Bayes dan Support Vector Machine dalam Klasifikasi Penyakit Diabetes Militus	Data berasal dari rekam medik Rumah Sakit Siti Khadijah-Palembang yang berjumlah 613 record dengan atribut sebanyak 9 atribut dalam kurun waktu hampir 3 tahun. Hasil akurasi, presisi, recall, specificity dan juga F1 Score dari data tersebut dengan menggunakan metode naïve bayes yaitu 92.07%, 93.08%, 97.00%, 75.00%, 95.00%. jika menggunakan metode Support Vector Machine Kernel Polynomial hasil akurasi, presisi, recall, specificity dan juga F1 Score sebesar 96.72%, 99.42%, 96.10%, 97.14%, 97.73%. sedangkan jika menggunakan metode Support Vector Machine Kernel RBF hasil akurasi, presisi, recall, specificity dan juga F1 Score sebesar 80.89%, 100%, 80.89%, -, 89.43%	naïve bayes, Support Vector Machine Kernel Polynomial, dan Support Vector Machine Kernel RBF
(Putry et al., 2022) / Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus	Penelitian ini akan mengklasifikasikan diagnosis penyakit diabetes melitus. Dimana yang menjadi data masukan atau input dalam penelitian ini adalah data diagnosis penyakit diabetes melitus dengan variable-variabel penentunya yaitu *Outcome, Glucose, Blood Pressure, Insulin, BMI, Age, Diabetes Pedigree Function*. Hasil akurasi, presisi dan recall pada data ini dengan menggunakan metode *naïve bayes* sebesar 80%, 86%, dan 86%. Sedangkan jika menggunakan metode *K-Nearest Neighbor* hasil akurasi, presisi dan recall sebesar 75%, 80%, dan 86%.	naïve bayes, K-Nearest Neighbor (KNN)
(Naik & Patel, 2013) / Tumor Detection and Classification using Decision Tree in Brain MRI	Data ini menggunakan 124 gambar MRI. Dari 124 gambar tersebut dibagi menjadi 73 gambar *training* dan 51 gambar *testing*. Hasil akurasi presisi, recall dan specificity dengan menggunakan metode naïve bayes adalah 88.2%, 91%, 91%, dan 83%. Sedangkan jika menggunakan metode	naïve bayes, Decision Tree

	Decision Tree maka hasil akurasi, presisi, recall dan specificity sebesar 96%, 100%, 93% dan 100%.	
(Otok & Nidhomuddin, 2015) / Random forest dan Multivariatea daptive Regression spline (Mars) Binary response Untuk Klasifikasi Penderita hiv/Aids Di Surabaya	Data yang digunakan dalam penelitian ini adalah data sekunder berupa data kasus penderita HIV/AIDS di Kota Surabaya yang didapatkan dari skripsi S1 ITS Surabaya yang disusun oleh Romaiza Millah Hanifa pada tahun 2013. Banyaknya data yang digunakan pada penelitian ini sebanyak 218 sampel dengan 13 variabel yang terdiri dari klien dengan status HIV/AIDS negatif dan klien dengan status HIV/AIDS positif. Hasil akurasi, recall dan specificity dengan menggunakan metode MARS sebesar 80.28%, 94.12%, dan 31.25%. jika menggunakan metode Random Forest maka hasil akurasi, recall dan specificity sebesar 97.8%, 100%, dan 95.55%. sedangkan jika menggunakan metode Random Forest MARS hasil akurasi, recall dan specificity sebesar 91%, 98.82, dan 62.50%.	MARS, Random Forest, Random Forest MARS

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon dimana setiap node mempresentasikan atribut, dimana cabangnya mempresentasikan nilai dari variabel prediktor, dan daun mempresentasikan kelasnya. Menurut Eska (2016) pada *Decision tree* terdapat tiga simpul, yaitu simpul akar (*root node*), simpul percabangan/ internal (*branch/ internal node*) dan simpul daun (*leaf node*). Salah satu algoritma yang terdapat pada metode *decision tree* adalah algoritma C4.5. Algoritma C4.5 merupakan suatu algoritma yang dapat digunakan untuk membentuk pohon keputusan sehingga mampu membuat prediksi atau mengklasifikasi. Proses dalam algoritma ini yaitu dengan memilih salah satu atribut yang akan digunakan sebagai akar pohon, kemudian akan dibuat cabang di dalam akar tersebut. Langkah berikutnya

yaitu membagi setiap kasus yang ada dalam cabang, kemudian proses akan terus diulang sampai setiap kasus berada dalam tiap-tiap kelas yang telah ditentukan (Parung, 2018).

Metode klasifikasi lain yang digunakan dalam penelitian ini yaitu *random forest. Random forest* adalah suatu metode ensamble yang terdiri dari sekumpulan pohon keputusan, dimana pohon keputusan tersebut digunakan untuk mengklasifikasi data ke suatu kelas. Berdasarkan Paul et al., (2018) kinerja klasifikasi dari *random forest* meningkat dengan bertambahnya jumlah pohon yang dibentuk. Metode *decision tree* dan *random forest* memiliki peranan untuk mengambil keputusan dalam pengklasifikasian sebuah data.

Menurut Azhari et al., (2021) beberapa metode klasifikasi antara lain *Decision Tree, rule-based classifiers, Bayesian classifier Support Vector Machines, Artificial Neural Networks, Lazy Learners,* dan *ensemble methods.*

1.2 Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah :

1. Bagaimana klasifikasi penyakit diabetes dengan menggunakan metode *decision tree* C4.5 dan *random forest?*
2. Manakah metode yang memiliki tingkat ketepatan yang lebih baik dalam mengklasifikasikan penyakit diabetes?

1.3 Batasan Masalah

Batasan-batasan masalah dalam penelitian ini antara lain :

1. Tingkat ketepatan diukur menggunakan akurasi, presisi, recall, specificity dan F1 score.
2. Data di partisi menjadi 80% data latih (416 pengamatan) , dan 20% data uji (104 pengamatan).

1.4 Tujuan Penelitian

1. Mengklasifikasi penyakit diabetes menggunakan metode *decision tree* C4.5 dan *random forest.*

2. Membandingkan tingkat akurasi metode *decision tree* C4.5 dan *random forest*.

1.5 Manfaat Penelitian

1. Menjadi media belajar bagi penulis dan juga pembaca terkait penggunaan metode *decision tree* C4.5 dan *random forest* dalam melakukan klasifikasi.
2. Sebagai bahan referensi untuk penelitian yang membahas penyakit diabetes.

BAB II

TINJAUAN PUSTAKA

2.1 Diabetes

Diabetes adalah suatu penyakit yang memiliki berbagai macam gangguan metabolisme yang ditandai dengan tingginya kadar glukosa darah. Orang yang menderita penyakit diabetes memiliki peningkatan risiko masalah kesehatan serius hingga mengancam jiwa yang mengakibatkan biaya perawatan medis, penurunan kualitas hidup dan peningkatan kematian (Cho et al., 2018). Diabetes juga dapat dikatakan sebagai sebuah penyakit metabolik yang disebabkan oleh kurangnya hormon insulin atau ketidakmampuan tubuh dalam memanfaatkan insulin, sehingga kada glukosa atau kadar gula dalam darah tidak terkendali (Najib et al., 2019).

Dennedy et al., (2015) mengemukakan bahwa pasien yang sedang mengalami gejala diabetes dianjurkan untuk lebih mengkontrol diri karena hal ini sangat penting untuk mencegah komplikasi akut dan mengurangi risiko komplikasi. Perkiraan prevalensi global sebesar 9,3% pada tahun 2019, diabetes merupakan masalah kesehatan masyarakat global yang signifikan, bertanggung jawab atas mortalitas dan morbiditas yang cukup besar di seluruh dunia dan menyebabkan kerugian ekonomi yang substansial (Yezli et al., 2021).

Diabetes merupakan salah satu tantangan kesehatan masyarakat yang paling penting di abad-21 dan juga merupakan salah satu penyakit terbesar epidemi yang dihadapi dunia, baik dalam negara maju dan berkembang. Diabetes

memberikan kontribusi sebagai salah satu penyebab kematian utama pada penderita penyakit jantung dan pembuluh darah (Wulandari et al., 2015).

Menurut Punthakee et al., (2018) diabetes dapat dikelompokkan menjadi beberapa bagian yaitu:

1. Diabetes melitus tipe 1 meliputi diabetes disebabkan oleh penghancuran sel β pankreas baik oleh proses autoimun maupun idiopatik sehingga produksi insulin berkurang bahkan berhenti. Biasanya diabetes tipe ini terjadi pada usia kurang dari 20 tahun.
2. Diabetes melitus tipe 2 merupakan diabetes dengan kelainan metabolik yang ditandai dengan kadar glukosa darah yang tinggi dalam konteks resistensi insulin dan defisiensi insulin relatif. Diabetes tipe ini biasanya di derita pada usia lebih dari 20 tahun.
3. Diabetes melitus gestasional mengacu pada intoleransi glukosa pada masa pengenalan pertama selama kehamilan.
4. Jenis spesifik lainnya mencakup berbagai macam kondisi tidak umum, terutama bentuk diabetes yang ditentukan secara genetik atau diabetes yang terkait dengan penyakit lain atau penggunaan narkoba.

Menurut Siallagan dan Fitriyani (2021) ada beberapa gejala diabetes yang perlu kita waspadai, yaitu :

a. *Polydipsia* (cepat haus)
b. *Polyuria* (banyak buang air kecil)

c. *Polyphagia* (cepat lapar)

d. *Sudden Weight Loss* (penurunan berat badan)

e. *Weakness* (rasa lelah dam lemah yang tidak biasa)

f. *Visual Blurring* (pandangan kabur)

g. *Delayed Healing* (pemulihan luka yang lama atau sering infeksi)

h. *Acanthosis Nigricans* (warna kulit gelap)

2.2 Data Mining

Data mining atau disebut juga dengan *Knowledge Discovery in Database* (KDD) merupakan aktivitas yang berkaitan dengan pengumpulan data, pemakaian data historis untuk menemukan pengetahuan, informasi, keteraturan, pola atau hubungan dalam data yang berukuran besar (Buulolo, 2020). Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, ataupun algoritma yang ada dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan (Chan, 2018).

Kemampuan data mining dalam menggali informasi dalam cakupan yang sangat besar ini menjadi kelebihan yang tidak perlu diragukan. Teknologi seperti ini biasanya digunakan untuk memprediksi berbagai hal dalam kehidupan, dimana data mining mengotomatisasi proses pencarian informasi di dalam basis data yang besar dan menemukan pola-pola yang tidak diketahui sebelumnya.

Menurut (Mardi, 2017) data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Deskripsi adalah suatu cara yang digunakan untuk menggambarkan suatu pola dan kecenderungan dalam data.
2. Estimasi hampir sama dengan klasifikasi, tetapi bagian atribut target estimasi lebih mengarah pada numerik daripada kategori.
3. Prediksi hampir sama dengan klasifikasi dan estimasi, prediksi memberikan gambaran nilai dari hasil yang akan ada dimasa mendatang.
4. Klasifikasi merupakan pengelompokan data menjadi beberapa bagian atau kelas. Contohnya seperti penggolongan diabetes yaitu positif atau negatif, kemudian penggolongan berat badan yaitu gemuk, sedang atau kurus dll.
5. Pengklusteran merupakan pengelompokan hasil dari pengamatan yang memiliki objek kemiripan.
6. Asosiasi data mining memiliki tugas yaitu memberikan informasi tentang hubungan item dalam database serta menemukan atribut yang muncul dalam satu waktu.

2.3 Klasifikasi

Klasifikasi adalah proses pengkatagorian yang dilakukan dalam sekumpulan data kemudian membaginya dalam kelas kelas tertentu. Klasifikasi memberikan penilaian objek data untuk memasukannya kedalam kelas tertentu

dari jumlah kelas yang tersedia. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap set atribut ke satu jumlah label kelas yang tersedia (Utomo & Mesran, 2020). Menurut (NAsrullah, 2021) Klasifikasi adalah proses dari mencari suatu himpunan model (fungsi) yang dapat mendeskripsikan dan membedakan kelas-kelas data, dengan tujuan dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui.

Klasifikasi diberikan sejumlah *record* yang dinamakan langkah pelatihan, yang terdiri dari beberapa atribut, salah satu atribut menunjukkan kelas untuk *record* yang dapat digunakan untuk menemukan model dari langkah pelatihan sehingga dari hasil tersebut kita dapat membedakan *record* ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan *record* yang kelasnya belum diketahui sebelumnya (Sunjana, 2010). Proses klasifikasi biasanya dibagi menjadi dua tahapan : *learning* dan *test*. Pada tahap *learning*, sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan. Kemudian pada tahap *test* model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tersebut. Bila akurasinya mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui (Novianti et al., 2016).

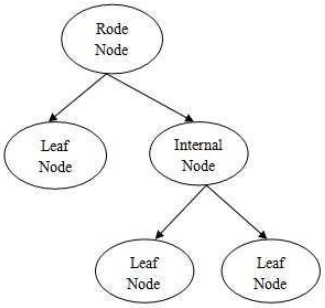
Menurut Azhari et al., (2021) beberapa metode klasifikasi antara lain *Decision Tree, rule-based classifiers, Bayesian classifier Support Vector Machines, Artificial Neural Networks, Lazy Learners,* dan *ensemble methods.*

2.4 Decision Tree

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon dimana setiap node mempresentasikan atribut, dimana cabangnya mempresentasikan nilai dari atribut, dan daun mempresentasikan kelasnya. *Decision tree* adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan-kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan (Mardi, 2017). *Decision tree* adalah metode klasifikasi yang mudah untuk dipahami atau diintrepetasikan oleh manusia, sehingga metode ini menjadi salah satu metode yang cukup popular. (Chandrasekar et al., 2017).

Menurut Alsalman et al., (2019) d*ecision tree* adalah teknik yang banyak digunakan untuk membangun model klasifikasi berdasarkan kumpulan data yang dikumpulkan. Ini adalah suatu teknik yang menciptakan pohon aturan dengan menghitung rasio keuntungan (*gain ratio*) yang memberikan bobot tertentu pada atribut-atribut yang terdapat dalam sebuah himpunan data.

Seorang peneliti mengatakan bahwa *decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon di mana setiap *node* merepresentasikan atribut, dimana cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. Akar dari *decision tree* disebut sebagai *root.* Pada *decision tree* terdapat 3 jenis *node,* yaitu *Root Node, Internal Node,* dan *Leaf node* (Saifullah et al., 2017). Setiap *leaf node* di pohon mewakili tes untuk nilai atribut, dan cabangnya mewakili setiap hasil tes (ya dan tidak berarti positif dan negatif) (Jiang et al., 2019). Berikut ini contoh gambar *decision tree.*



Gambar 2.1 Struktur *Decision Tree* (Zhou et al., 2020)

1. *Root Node* merupakan akar pohon atau *node* paling atas, pada *node* ini tidak ada input dan bisa tidak mempunyai output atau mempunyai output lebih dari satu.
2. *Internal Node* merupakan *node* percabangan, pada *node* ini hanya ada satu input dan setidaknya mempunyai dua output.
3. *Leaf Node* atau *Terminal Node*, merupakan *node* akhir atau hasil akhir pohon keputusan, pada *node* ini hanya terdapat satu input dan tidak mempunyai output (Widiyati et al., 2018).

Menurut Hamoud et al., (2018) *decision tree* memberikan banyak keuntungan untuk pengolahan data, beberapa di antaranya sebagai berikut:

1. *Decision tree* dapat dipahami dengan jelas oleh analis dan semua pihak pengguna.
2. *Decision tree* dapat menangani berbagai jenis data input, yaitu nominal, numerik, dan tekstual.
3. *Decision tree* dapat memproses kumpulan data yang salah atau hilang atau nilai-nilai yang belum selesai.

4. *Decision tree* memiliki tingkat kinerja yang tinggi dengan minimal jumlah data yang besar dan waktu yang singkat.
5. *Decision tree* dapat bekerja dalam aplikasi pengolahan data melalui berbagai platform atau software.

Selain kelebihan metode *decision tree* juga memiliki beberapa kekurangan, menurut Syamsu et al., (2019) kekurangan dari metode ini antara lain yaitu:

1. Ketika kelas-kelas dan kriteria yang digunakan jumlahnya sangat banyak maka akan terjadi kelebihan beban/*overlap*. Sehingga waktu pengambilan keputusan dan jumlah memori yang diperlukan pun juga akan mengalami peningkatan.
2. Pengakumulasian jumlah error dari setiap tingkat dalam sebuah pohon keputusan yang besar.
3. Kesulitan dalam mendesain pohon keputusan yang optimal.

Proses rekursif *decision tree* akan berhenti bila kondisi terpenuhi, yaitu: semua data sampel berada dalam kelas-kelas yang telah ditentukan, tidak ada lagi variabel prediktor yang akan dilakukan partisi, atau tidak ada data sampel lagi yang akan diuji (Kurniawan, 2020). Saat *decision tree* terus membelah data, pohon tumbuh lebih besar dan lebih besar, dan pohon menjadi lebih akurat untuk dataset pelatihan.

2.5 Algoritma C4.5

C4.5 merupakan salah satu algoritma pada *Decision Tree* berdasarkan informasi entropi. Algoritma ini menggunakan kriteria pemisahan yang dimodifikasi yang disebut dengan rasio *gain* (Nasution et al., 2018). Tujuan dari algoritma ini adalah untuk menemukan beberapa hubungan antara variabel prediktor dan kategori melalui pelatihan dan pembelajaran set pelatihan, kemudian menerapkan hubungan ini ke contoh, mengklasifikasikan data dan menyelesaikan pengambilan keputusan (Liu et al., 2019). Algoritma C4.5 merupakan pengembangan algoritma ID3 dimana kekurangan yang dimiliki algoritma ID3 ditutupi oleh algoritme C4.5. Empat hal yang membedakan algoritma C4.5 dengan ID3 antara lain: tahan (robust) terhadap data noise, mampu menangani variabel dengan tipe diskrit maupun kontinu, mampu menangani variabel yang memiliki missing value, dan dapat memangkas cabang dari pohon keputusan (Setio et al., 2020).

Kelebihan yang dimiliki oleh algoritma ini adalah dapat di pahami dengan mudah karena dapat digambarkan dalam bentuk pohon keputusan. Menurut Hana (2020) cara mencari nilai *entropy* adalah dengan persamaan dibawah ini:

$$Entropi(S) = -\sum_{i=1}^{k} P_i log_2 P_i \quad (2.1)$$

Keterangan:

S: Himpunan kasus

k: jumlah partisi S

P_i: probabilitas yang didapat dari jumlah (ya/tidak) dibagi total kasus.

Sedangkan untuk mencari nilai *gain* dapat menggunakan rumus sebagai berikut:

$$Gain(S, X) = Entropi(S) - \sum_{i=1}^{n} - \frac{|s_i|}{|s|} * Entropi\ S_i \quad (2.2)$$

Keterangan :

S: Himpunan kasus

X: variabel prediktor

n: Jumlah partisi variabel prediktor X

$|s_i|$: Jumlah kasus pada partisi ke-i

$|s|$: Jumlah kasus dalam S

2.6 Random Forest

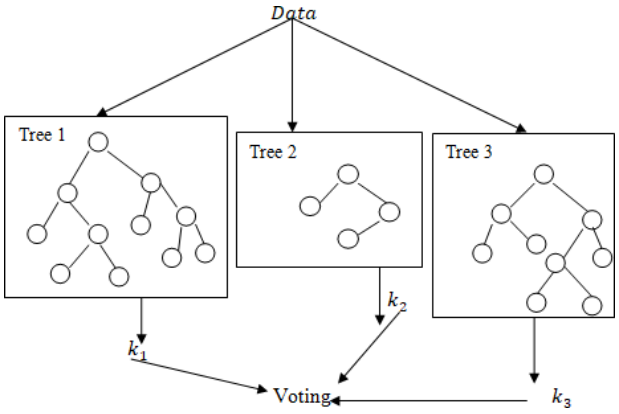
Random forest adalah salah satu metode di dalam *machine learning* yang digunakan untuk proses mengklasifikasikan data dalam jumlah yang besar. *Random forest* adalah pengembangan dari metode *Classification and Regression Tree* (CART) dengan menerapkan agregasi *bootstrap* dan metode pemilihan fitur secara acak (Nugroho et al., 2019).

Random Forest melakukan klasifikasi dengan cara melakukan pendekatan metode ensemble dari berbagai pohon melalui banyaknya kemunculan untuk mencapai keputusan akhir (Religia et al., 2021). *Random forest* diawali dengan teknik dasar data mining yaitu *decision tree*. Dengan kata lain *random forest* terdiri dari sekumpulan *decision tree*, dimana kumpulan decision tree tersebut digunakan untuk mengklasifikasi data ke suatu kelas. Metode ini menciptakan

berbagai pohon (*tree*) dalam sebuah hutan (*forest*) sehingga akan membuat hutan ini menjadi semakin kuat. Sama halnya dengan *decision tree* yang menggunakan *entropy* dan *gain* untuk perhitungan dalam membangun satu pohon. Masing-masing pohon yang akan dibentuk menggunakan data set yang diambil secara acak dari data latih. Selama proses klasifikasi maka setiap pohon akan memberikan *voting* kelas yang paling popular.

Cara kerja dari *random forest* dapat dianalogikan dengan kasus berikut ini. Seorang mahasiswa ingin membeli makanan yang enak namun tidak tahu makanan mana yang enak menurutnya. Mahasiswa tersebut memutuskan untuk bertanya pada seorang temanya. Kemudian temanya memberikan beberapa pertanyaan untuk memutuskan rekomendasi makanan enak yang ingin di beli oleh mahasiswa tersebut. Sejauh ini kasus tersebut menggambarkan metode *decision tree*, dimana seorang mahasiswa menggambarkan pohon yang dibangun untuk memutuskan rekomendasi makanan enak. Kemudian mahasiswa tersebut bertanya pada beberapa teman lainya. Teman lainya mengajukan beberapa pertanyaan yang berbeda secara acak, setiap teman memberikan rekomendasi yang berbeda-beda dan juga ada yang sama. Lalu mahasiswa tersebut akan memutuskan untuk membeli makanan enak yang paling banyak direkomendasikan oleh teman-temanya. Inilah yang menggambarkan proses dan cara kerja metode *random forest*, dimana akan membentuk pohon yang banyak untuk memutuskan suatu keputusan, lalu keputusan akhir akan ditentukan oleh hasil keputusan terbanyak dari pohon yang telah dibangun. Konsep inilah yang disebut dengan *majority*

voting. Berikut ini adalah gambar dari metode *Random Forest* (Nugroho et al., 2019)



Gambar 2.2 Struktur *Random Forest*

2.8 *Confusion Matrix*

Confusion Matrix adalah satu istilah mendasar dalam *machine learning*. *Confusion Matrix* digunakan untuk mengukur akurasi model dengan cara membandingkan nilai prediksi dan juga nilai aktual (Zeng, 2020). Berikut ini tabel *Confusion Matrix* menurut (Tanujayaa et al., 2020).

Tabel 2.1 *Confusion Matrix*

kelas	Aktual Positif	Negatif
Prediksi positif	Positif benar TP	Positif palsu FP
Negatif	Negatif palsu FN	Negatif benar TN

Rumus untuk menghitung nilai akurasi:

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

Rumus untuk menghitung nilai presisi:

$$presisi = \frac{TP}{TP + FP} \quad (2.4)$$

Rumus untuk menghitung nilai recall:

$$recall = \frac{TP}{TP + FN} \quad (2.5)$$

Rumus untuk menghitung nilai specificity:

$$specificity = \frac{TN}{TN + FP} \quad (2.6)$$

Rumus untuk menghitung nilai F1 Score:

$$F1\ Score = 2\ (\frac{recall * presisi}{recall + presisi}\ ) \quad (2.7)$$

Dimana:

TP : jumlah kelas positif yang diklasifikasi sebagai positif

FP : jumlah kelas negatif yang diklasifikasi sebagai positif

TN : jumlah kelas negatif yang diklasifikasi sebagai negatif

FN : jumlah kelas positif yang klasifikasi sebagai negative

BAB III

METODOLOGI PENELITIAN

3.1 Tempat

Penelitian ini dilakukan di jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam dan Perpustakaan Universitas Sriwijaya.

3.2 Waktu

Waku yang dibutuhkan dalam penelitian ini mulai dari bulan Agustus 2021 sampai dengan Juli 2022.

3.3 Data Penelitian

Data yang akan digunakan dalam penelitian ini diperoleh dari *https://archive.ics.uci.edu/ml/machine-learning-databases/00529/*. Data ini berukuran 520 dan memiliki 16 variabel independen dengan variabel dependen berupa status penyakit diabetes yaitu positif atau negatif diabetes. Keenam belas variabel tersebut adalah :

a. Age : umur dari orang-orang yang ikut serta dalam kuisioner penelitian.
b. Gender : jenis kelamin
c. Polyuria : sering buang air kecil dengan frekuensi lebih dari 20 kali/hari
d. Polydipsia : haus yang berlebihan hingga mencapai lebih dari 3 L/hari
e. Sudden weight loss : penurunan berat badan secara drastis
f. Weakness : daya tahan tubuh mulai menurun

g. Pholypagia : peningkatan nafsu makan yang berlebihan
h. Genital trhush : infeksi jamur dan bakteri
i. Visual blurring : sakit kepala dan penglihatan yang mulai buram
j. Itching : rasa gatal yang tidak berkesudahan
k. Irritability : sifat cepat memerah pada badan dan pembengkakan pada tangan serta kaki
l. Delayed healing : luka yang sulit untuk sembuh
m. Partial paresis : lemahnya gerak badan atau lemas pada badan
n. Muscle stiffness : sulit bergerak atau melemahnya kinerja otot
o. Alopecia : rambut rontok bpada area tertentu sehingga membuat pitak
p. Obesity : obesitas
q. Class : kelas yang akan diklasifikasikan yaitu positif atau negatif diabetes

3.4 Metode Penelitian

Langkah- langkah yang digunakan dalam penelitian ini yaitu :

1. Mendeskripsikan data penelitian, *preprocessing*, dan mempartisi data menjadi 80% data latih dan 20% data uji.
2. Mengklasifikasi status penyakit diabetes menggunakan metode *Decision Tree*

a. Menghitung *entropy* menggunakan rumus pada persaman (2.1).
b. Menghitung nilai *gain* menggunakan rumus pada persaman (2.2).
c. *Gain* tertinggi akan dipilih untuk menjadi *root node*/akar pohon.

d. Ulangi langkah perhitungan *entropy d*an *gain* sampai seluruh variabel predictor masuk dalam kelas. variabel prediktor yang sudah terpilih sebelumnya maka tidak diikutkan untuk perhitungan selanjutnya.

e. Menghitung tingkat ketepatan klasifikasi.

3. Mengklasifikasi status penyakit diabetes menggunakan metode *Random Forest* :

a. Melakukan *bootstrap sampling* untuk mengambil sampel.

b. Menentukan banyaknya m variabel prediktor secara acak untuk setiap *node*, dimana dalam hal ini kita menggunakan $m = |\sqrt{(16\ )}| = 4$.

c. Menghitung nilai *entropy* dan *gain* dengan rumus pada persamaan (2.1) dan (2.2).

d. Nilai *Gain* tertinggi dipilih untuk menjadi *root node*.

e. Ulangi proses di atas sampai membentuk pohon yang banyak.

f. Menentukan klasifikasi dengan cara *majority vote*.

g. Menghitung tingkat ketepatan klasifikasi.

4. Membandingkan tingkat ketepatan klasifikasi metode *decision tree* dan *random forest*

BAB IV

HASIL DAN PEMBAHASAN

4.1 Deskripsi Data, *Prepocessing*, Partisi Data

Deskripsi variabel penelitian dapat dilihat pada tabel 4.1 dibawah ini.

Tabel 4. 1 Deskripsi variabel

No	Variabel	Data	Frekuensi
1	X_1(*Age*)	1 : <20 tahun	1
		2 : 20-40 tahun	167
		3 : >40 tahun	352
2	X_2(*Gender)*	1: Male	328
		0: Female	192
3	X_3(*Polyuria*)	1: Yes	258
		0: No	262
4	X_4(*Polydipsia)*	1: Yes	233
		0: No	287
5	X_5(*sudden_weight_loss)*	1: Yes	217
		0: No	303
6	X_6(*weakness*)	1: Yes	305
		0: No	215
7	X_7(*Polyphagia)*	1: Yes	237
		0: No	283
8	X_8(*Genital_thrush*)	1: Yes	116
		0: No	404
9	X_9(*visual_blurring)*	1: Yes	233
		0: No	287
10	X_{10}(*Itching)*	1: Yes	253
		0: No	267
11	X_{11}(*Irritability)*	1: Yes	126
		0: No	394
12	X_{12}(*delayed_healing)*	1: Yes	239
		0: No	281
13	X_{13}(*partial_paresis)*	1: Yes	224
		0: No	296
14	X_{14}(*muscle_stiffness)*	1: Yes	195
		0: No	325
15	X_{15}(*Alopecia)*	1: Yes	179
		0: No	341

16	X_{16}(*Obesity)*	1: Yes 0: No	88 432
17	*Y* (*class)*	1: Positive 0: Negative	320 200

Preprosessing data dilakukan dengan mendiskritisasi variabel kontinu menjadi varibel katagorik. Pada data diabetes diatas terdapat variabel kontinu yaitu X_1 yang merupakan variabel umur dengan interval 16-90 tahun. Jika umur pasien kurang dari 20 tahun makan dikatagorikan sebagai *young* (1), jika umur pasien berada pada interval 20 tahun sampai 40 tahun dikatagorikan sebagai *medium/*dewasa (2), dan jika umur pasien berada diatas 40 tahun maka dikatagorikan *old* (3) (Anggraeni & Ramadhani, 2018).

Partisi data dilakukan untuk membagi data ke dalam dua bagian secara acak, dimana dalam penelitian ini yang digunakan adalah *train/test* split. Data dipartisi menjadi 80% data *train* (416 pengamatan) dan 20% data *test* (104 pengamatan).

Tabel 4.2 Data *train* 80 %

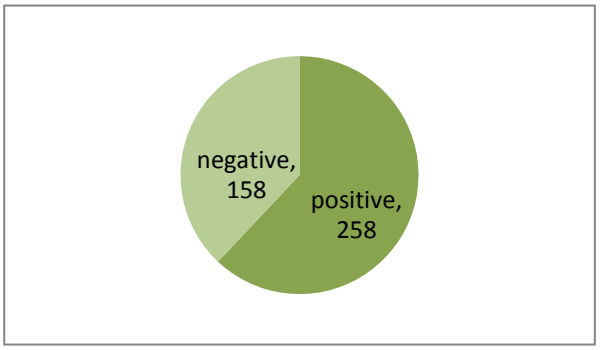
No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y
1	2	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1
2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	0	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1
4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	2	1	1	0	0	0	1	0	1	1	0	1	1	1	1	0	0
.																	
.																	
.																	
416	2	1	1	0	0	0	0	1	0	1	0	0	0	1	0	1	1

Tabel 4.3 Data *test* 20%

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y
1	3	1	1	1	0	1	0	1	0	0	1	1	0	1	1	0	1
2	3	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
3	3	1	1	0	1	0	0	1	0	1	1	0	0	0	1	0	1
4	3	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1
5	3	1	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1
.																	
.																	
.																	
104	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

4.2 Mengklasfikasi Status Penyakit Diabetes dengan Metode *Decision Tree*

Langkah pertama yang dilakukan adalah dengan mencari nilai *entropy* dan juga nilai *gain*. Nilai *gain* tertinggi pertama akan dijadikan sebagai akar pohon. Banyaknya sampel yang digunakan sama dengan jumlah data *train* yaitu sebanyak 416. Berdasarkan banyaknya data sampel yang digunakan terdapat 258 orang yang positif diabetes dan 158 orang yang negative diabetes.



Gambar 4.1 Grafik jumlah data *train* positif dan negaitif

Maka peluang dari variabel respon untuk katagori positif dan negatif yaitu :

$$P(Y = 0) = \frac{n(Y = 0)}{n_{total}} = \frac{158}{416} = 0.3798$$

$$P(Y=1) = \frac{n(Y=1)}{n_{total}} = \frac{258}{416} = 0.6202$$

Selanjutnya akan dibentuk node ke-1 dengan mencari nilai e*ntropy* setiap variabel, yaitu sebagi berikut:

Tabel 4.4 perhitungan *entropy* dan *gain* node 1

Variabel	Label	Jumlah	0	1	p_0	p_1	Entropy	Gain
Total		416	258	158	0.6201	0.3798	0.9579	
X_1	1	0	0	1	0	1	0	0.1271
	2	348	191	157	0.5488	0.4511	0.9931	
	3	67	67	0	1	0	0	
X_2	0	152	136	16	0.8947	0.1052	0.4854	0.1485
	1	264	122	142	0.4621	0.5378	0.9958	
X_3	0	216	69	147	0.3194	0.6805	0.9037	0.3409
	1	200	189	11	0.945	0.055	0.3072	
X_4	0	226	75	151	0.3318	0.6681	0.9168	0.3558
	1	190	183	7	0.9631	0.0368	0.2276	
X_5	0	249	111	138	0.4457	0.5542	0.9915	0.1522
	1	167	147	20	0.8802	0.1197	0.5286	
X_6	0	176	85	91	0.4829	0.5170	0.9991	0.0423
	1	240	173	67	0.7208	0.2791	0.8543	
X_7	0	226	104	122	0.4601	0.5398	0.9954	0.0972
	1	190	154	36	0.8105	0.1894	0.7003	
X_8	0	327	197	130	0.6024	0.3975	0.9695	0.0036
	1	89	61	28	0.6853	0.3146	0.8984	
X_9	0	228	115	113	0.5043	0.4956	0.9999	0.0510
	1	188	143	45	0.7606	0.2393	0.7939	
X_{10}	0	220	136	84	0.6181	0.3818	0.9593	1.38951E-05
	1	196	122	74	0.6224	0.3775	0.9562	
X_{11}	0	318	174	144	0.5471	0.4528	0.9935	0.0590
	1	98	84	14	0.8571	0.1428	0.5916	
X_{12}	0	234	138	96	0.5897	0.4102	0.9766	0.0036
	1	182	120	62	0.6593	0.3406	0.9254	
X_{13}	0	237	101	136	0.4261	0.5738	0.9842	0.1658
	1	179	157	22	0.8770	0.1229	0.5376	
X_{14}	0	264	152	112	0.5757	0.4242	0.9833	0.0106
	1	152	106	46	0.6973	0.3026	0.8844	
X_{15}	0	283	199	84	0.7031	0.2968	0.8773	0.0442
	1	133	59	74	0.4436	0.5563	0.9908	
X_{16}	0	345	208	137	0.6028	0.3971	0.9692	0.0045
	1	71	50	21	0.7042	0.2957	0.8760	

$$Entropi(Y) = -\sum_{i=1}^{k} P_i log_2 P_i$$

$$= -(P(Y=0)log_2 P(Y=0) + (P(Y=1)log_2 P(Y=1)$$

$$= -((0.3798)log_2(0.3798) + (0.6202)log_2(0.6202))$$

$$= 0.9579$$

$$Entropi(X_1 = 1) = -(P(Y=0|X_1=1)log_2 P(Y=0|X_1=1) +$$

$$P(Y=1|X_1=1)log_2 P(Y=1|X_1=1))$$

$$= -((1)log_2(1) + (0)log_2(0))$$

$$= 0$$

$$Entropi(X_1 = 2) = -(P(Y=0|X_1=2)log_2 P(Y=0|X_1=2) +$$

$$P(Y=1|X_1=2)log_2 P(Y=1|X_1=2))$$

$$= -((0.4511)log_2(0.4511) + (0.5488)log_2(0.5488))$$

$$= 0.9931$$

$$Entropi(X_1 = 3) = -(P(Y=0|X_1=3)log_2 P(Y=0|X_1=3) +$$

$$P(Y=1|X_1=3)log_2 P(Y=1|X_1=3))$$

$$= -((0)log_2(0) + (1)log_2(1))$$

$$= 0$$

$$Entropi(X_2 = 0) = -(P(Y=0|X_2=0)log_2 P(Y=0|X_2=0) +$$

$$P(Y=1|X_2=0)log_2 P(Y=1|X_2=0))$$

$$= -((0.1053)log_2(0.1053) + (0.8947)log_2(0.8947))$$

$$= 0.4854$$

$$Entropi(X_2 = 1) = -(P(Y = 0|X_2 = 1)log_2P(Y = 0|X_2 = 1) + P(Y = 1|X_2 = 1)log_2P(Y = 1|X_2 = 1))$$
$$= -((0.5379)log_2(0.5379) + (0.4621)log_2(0.4621))$$
$$= 0.9958$$
$$Entropi(X_3 = 0) = -(P(Y = 0|X_3 = 0)log_2P(Y = 0|X_3 = 0) + P(Y = 1|X_3 = 0)log_2P(Y = 1|X_3 = 0))$$
$$= -((0.6805)log_2(0.6805) + (0.3194)log_2(0.3194))$$
$$= 0.9037$$

$$Entropi(X_3 = 1) = -(P(Y = 0|X_3 = 1)log_2P(Y = 0|X_3 = 1) + P(Y = 1|X_3 = 1)log_2P(Y = 1|X_3 = 1))$$
$$= -((0.055)log_2(0.055) + (0.945)log_2(0.945))$$
$$= 0.3072$$
$$Entropi(X_4 = 0) = -(P(Y = 0|X_4 = 0)log_2P(Y = 0|X_4 = 0) + P(Y = 1|X_4 = 0)log_2P(Y = 1|X_4 = 0))$$
$$= -((0.6681)log_2(0.6681) + (0.3318)log_2(0.3318))$$
$$= 0.9168$$

$$Entropi(X_4 = 1) = -(P(Y = 0|X_4 = 1)log_2P(Y = 0|X_4 = 1) + P(Y = 1|X_4 = 1)log_2P(Y = 1|X_4 = 1))$$
$$= -((0.0368)log_2(0.0368) + (0.9632)log_2(0.9632))$$

$$= 0.2276$$

Perhitungan *entropy* dilanjutkan sampai semua variabel terhitung *entropy* nya, dengan cara yang sama seperti yang dilakukan di atas. Setelah semua variabel terhitung *entropy* nya, langkah selanjutnya yaitu menghitung nilai *gain.*

$$Gain(y,x_1) = Entropi(y) - \sum_{i=1}^{3} - \frac{x_{1i}}{n_{total}} * Entropi\,(x_{1i})$$

$$= 0.9579 - ((\frac{1}{416} * 0) + (\frac{348}{416} * 0.9931) + (\frac{67}{416} * 0))$$

$$= 0.1271$$

$$Gain(y, x_2) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{2i}}{n_{total}} * Entropi\,(x_{2i})$$

$$= 0.9579 - ((\frac{264}{416} * 0.9958) + (\frac{152}{416} * 0.4854))$$

$$= 0.1485$$

$$Gain(y, x_3) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{3i}}{n_{total}} * Entropi\,(x_{3i})$$

$$= 0.9579 - ((\frac{200}{416} * 0.3072) + (\frac{216}{416} * 0.9037))$$

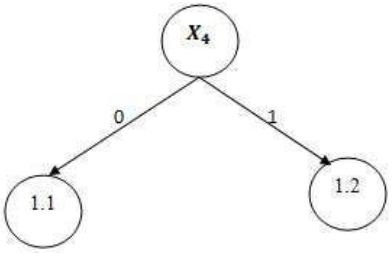
$$= 0.3409$$

$$Gain(y, x_4) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{4i}}{n_{total}} * Entropi\,(x_{4i})$$

$$= 0.9579 - ((\frac{190}{416} * 0.2276) + (\frac{226}{416} * 0.9168))$$

$$= 0.3558$$

Perhitungan *gain* dilakukan sampai semua variabel terhitung. Dari perhitungan tersebut hasil *entropy* dan juga *gain* diatas terlihat bahwa nilai *gain* tertinggi terletak pada X_4 yaitu sebesar 0.3558, ini berarti X_4 menjadi *root*/ akar dari node ke-1. Pohon keputusan untuk node 1 dapat dilihat pada gambar berikut.



Gambar 4.2 pohon keputusan *root node*.

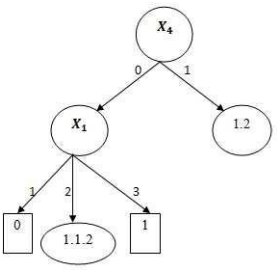
Selanjutnya akan dilakukan perhitungan *entropy* dan juga *gain* untuk semua cabang dari variabel X_4 sebagai akar pohon yang mana memiliki dua katagori yaitu katagori *Yes* (1) memiliki 190 kasus dan katagori *No* (0) memiliki 226 kasus. Berikut ini tabel perhitungan *entropy* dan *gain* untuk node 1.1 katagori *No* (0).

Tabel 4.5 perhitungan *entropy* dan *gain* node 1.1

Variabel	Label	Jumlah	0	1	p_0	p_1	Entropy	Gain
X_1	1	1	0	1	0	1	0	0.1474
	2	206	56	150	0.2718	0.7281	0.8440	
	3	19	19	0	1	0	0	
X_2	0	55	39	16	0.7090	0.2909	0.8698	0.1433

	1	171	36	135	0.2105	0.7894	0.7424	
X_3	0	181	41	140	0.2265	0.7734	0.7718	0.1388
	1	45	34	11	0.7555	0.2444	0.8023	
X_5	0	177	44	133	0.2485	0.7514	0.8090	0.0775
	1	49	31	18	0.6326	0.3673	0.9486	
X_6	0	131	44	87	0.3358	0.6641	0.9208	7.25357E-05
	1	95	31	64	0.3263	0.6736	0.9111	
X_7	0	158	42	116	0.2658	0.7341	0.8354	0.0320
	1	68	33	35	0.4852	0.5147	0.9993	
X_8	0	180	57	123	0.3166	0.6833	0.9007	0.0028
	1	46	18	28	0.3913	0.6086	0.9656	
X_9	0	162	50	112	0.3086	0.6913	0.8915	0.0043
	1	64	25	39	0.3906	0.6093	0.9652	
X_{10}	0	138	54	84	0.3913	0.6086	0.9656	0.0184
	1	88	21	67	0.2386	0.7613	0.7927	
X_{11}	0	190	53	137	0.2789	0.7210	0.8540	0.0452
	1	36	22	14	0.6111	0.3888	0.9640	
X_{12}	0	142	52	90	0.3661	0.6338	0.9477	0.0065
	1	84	23	61	0.2738	0.7261	0.8468	
X_{13}	0	176	44	132	0.25	0.75	0.8112	0.0730
	1	50	31	19	0.62	0.38	0.9580	
X_{14}	0	161	50	111	0.3105	0.6894	0.8938	0.0036
	1	65	25	40	0.3846	0.6153	0.9612	
X_{15}	0	126	48	78	0.3809	0.73	0.9587	0.0099
	1	100	27	73	0.27	0.6190	0.8414	
X_{16}	0	196	65	131	0.3316	0.6683	0.9165	1.08257E-06
	1	30	10	20	0.333	0.6666	0.9182	

Dari tabel perhitungan diatas terlihat bahwa yang memiliki nilai *gain* tertinggi adalah X_1 dengan nilai *gain* sebesar 0.1474 . Maka X_1 menjadi cabang pohon pertama dari katagori 0 (*No*). Pada variabel X_1 memiliki tiga katagori dimana dari ketiga katagori tersebut dua diantaranya memiliki hasil akhir untuk klasifikasi penyakit diabetes. Dua katagori tersebut yaitu pada katagori 1 (umur kurang dari 20 tahun (*young*)) masuk dalam klasifikasi negatif diabetes dengan jumlah 1 orang dan katagori 3 (umur lebih dari 40 tahun (*old*)) masuk dalam klasifikasi positif diabetes dengan jumlah 19 orang. Pohon keputusan dengan akar X_4 dan katagori 0 (*No*) adalah X_1dapat dilihat pada gambar berikut.



Gambar 4.3 pohon keputusan node 1.1

Selanjutnya melakukan perhitungan *entropy* dan juga *gain* dengan cara yang sama untuk seluruh cabang pohon keputusan. Proses perhitungan pohon keputusan dihentikan jika semua data sampel berada dalam kelas yang sama, tidak ada lagi atribut yang akan dilakukan partisi, atau tidak ada data sampel lagi yang akan diuji. Dari perhitungan yang dilakukan maka akan terbentuk sebuah model pengkondisian untuk menentukan klasifikasi pada penyakit diabetes. Model pengkondisian dari metode *decision tree* dapat dilihat pada gambar dibawah ini.

```
Polydipsia = 0
|   Age = 1: 0 {1=0, 0=1}
|   Age = 2
|   |   Polyuria = 0
|   |   |   Gender = 0
|   |   |   |   Alopecia = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   muscle_stiffness = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   Irritability = 0: 1 {1=6, 0=2}
|   |   |   |   |   |   |   |   Irritability = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   |   Polyphagia = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   muscle_stiffness = 1: 1 {1=4, 0=0}
|   |   |   |   |   visual_blurring = 1: 1 {1=11, 0=0}
|   |   |   |   Alopecia = 1: 0 {1=0, 0=12}
|   |   |   Gender = 1
|   |   |   |   Irritability = 0
|   |   |   |   |   Alopecia = 0: 0 {1=0, 0=68}
|   |   |   |   |   Alopecia = 1
|   |   |   |   |   |   Itching = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   weakness = 0: 0 {1=0, 0=6}
|   |   |   |   |   |   |   |   weakness = 1: 1 {1=2, 0=2}
|   |   |   |   |   |   |   Polyphagia = 1: 1 {1=1, 0=0}
|   |   |   |   |   |   Itching = 1: 0 {1=0, 0=38}
|   |   |   |   Irritability = 1
|   |   |   |   |   Genital_thrush = 0
|   |   |   |   |   |   weakness = 0: 0 {1=0, 0=5}
|   |   |   |   |   |   weakness = 1
|   |   |   |   |   |   |   visual_blurring = 0: 1 {1=1, 0=0}
|   |   |   |   |   |   |   visual_blurring = 1: 0 {1=0, 0=4}
|   |   |   |   |   Genital_thrush = 1: 1 {1=3, 0=0}
|   |   Polyuria = 1
|   |   |   Itching = 0: 1 {1=20, 0=0}
|   |   |   Itching = 1
|   |   |   |   delayed_healing = 0: 1 {1=7, 0=0}
|   |   |   |   delayed_healing = 1
|   |   |   |   |   Gender = 0: 1 {1=1, 0=0}
|   |   |   |   |   Gender = 1: 0 {1=0, 0=11}
|   Age = 3: 1 {1=19, 0=0}
```

Gambar 4.4 Pengkondisian pohon keputusan metode *decision tree* menggunanakan software *rapid miner*

Dari gambar diatas terlihat sebagian pengkondisian dari pohon keputusan dengan metode *decision tree. Polydipsia* ($\boldsymbol{X_4}$) terpilih sebagai akar pohon. Jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) dan umur dengan katagori 1 (umur kurang dari 20 tahun) maka terklasifikasi kedalam katagori negatif diabetes. Jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) dan umur dengan katagori 3 (umur lebih dari 40 tahun) maka terklasifikasi kedalam katagori positif diabetes. Sedangkan jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) kemudian umur dengan katagori 2 (umur antara 20-

40 tahun), *polyuria* (kelainan produksi air seni) dengan katagori *No* (0) , jenis kelamin dengan katagori perempuan (0), *alopecia* (Kerontokan Rambut/kebotakan) dengan katagori *No* (0), *visual blurring* (Penglihatan tidak jelas) dengan katagori *No* (0), *muscle stiffness* (Kekakuan otot) dengan katagori *No* (0), *polyphagia* (peningkatan nafsu makan berlebih) dengan katagori *No* (0), dan *irritability* (kepekaan terhadap rangsangan) dengan katagori *No* (0) maka terklasifikasi kedalam katagori positif diabetes, sedangakan jika *irritability* (kepekaan terhadap rangsangan) dengan katagori *Yes* (1) maka terklasifikasi kedalam katagori negatif diabetes. Selanjutnya mencari setiap cabang lain yang belum terklasifikasi.

Model klasifikasi pada *decision tree* secara lengkap dapat dilihat pada **Lampiran**. Hasil klasifikasi dengan menggunakan metode *decision tree* dapat dilihat pada tabel 4.6 dibawah ini.

Tabel 4.6 *confusion matrix* metode *decision tree*

	True 1	*True* 0
Predict 1	58	4
Predict 0	5	37

Berdasarkan tabel yang ada diatas dapat dilihat bahwa 58 orang di prediksi benar masuk dalam klasifikasi positif diabetes, sedangkan 37 orang diprediksi benar masuk dalam klasifikasi negatif diabetes. Terdapat 5 orang di prediksi masuk dalam katagori positif diabetes, namun ternyata masuk dalam katagori negatif diabetes pada data sebenarnya. Sedangkan 4 orang yang diprediksi masuk

dalam katagori negatif diabetes ternyata masuk dalam katagori positif diabetes pada data sebenarnya. Kita juga dapat melihat tingkat akurasi, presisi, recall, specificity dan F1 score dari tabel diatas berikut ini.

$$akurasi = \frac{58 + 37}{104} = 0.9135$$

$$presisi = \frac{58}{58 + 4} = 0.9355$$

$$recall = \frac{58}{58 + 5} = 0.9206$$

$$specificity = \frac{37}{37 + 4} = 0.9024$$

$$F1\ score = 2\,(\frac{0.9206 * 0.9355}{0.9206 + 0.9355}) = 0.9280$$

Jadi, dengan menggunakan metode *decision tree* akurasi yang didapatkan sebesar 0.9135, presisi sebesar 0.9355, recall sebesar 0.9206, specificity sebesar 0.9024, dan F1 score sebesar 0.9280, hal ini menunjukkan bahwa ketepatan akurasi dalam memprediksi klasifikasi data penyakit diabetes dengan menggunakan metode *decision tree* adalah 91.35%.

4.3 Mengklasfikasi Status Penyakit Diabetes dengan Metode *Random Forest*

Langkah awal yang dilakukan dalam *Random Forest* yaitu menentukan banyaknya pohon yang akan di bentuk (n_{pohon}) dan juga banyaknya varibel yang dipilih untuk membangun pohon keputusan. Dalam penelitian akan dibangun 11 pohon dengan 4 variabel terpilih. Selanjutnya mengambil sampel data *train*,

dimana banyaknya data *train* ini yaitu sebesar 416 untuk kemudian dilakukan *bootstrap sampling* atau pengambilan sampel secara acak dengan pengembalian. Proses ini dilakukan setiap akan membangun suatu pohon keputusan, sehingga setiap pohon akan memiliki sampel yang berbeda-beda. Hasil dari proses *bootstrap sampling* untuk pohon pertama dapat dilihat pada tabel dibawah ini.

Tabel 4.7 *bootstrap sampling* pohon pertama

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	Y
72	2	1	1	0	0	0	1	0	1	1	0	1	1	0	1	0	0
180	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
162	2	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1
159	2	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
92	2	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
.																	
.																	
.																	
251	2	0	1	1	0	`1	0	0	0	0	1	0	0	1	0	0	1

Berdasarkan banyaknya data sampel yang digunakan pada tabel diatas terdapat 250 orang yang positif diabetes dan 166 orang yang negatif diabetes. Maka peluang dari variabel respon untuk katagori positif dan negatif yaitu :

$$P(Y = 0) = \frac{n(Y = 0)}{n_{total}} = \frac{166}{416} = 0.3991$$

$$P(Y = 1) = \frac{n(Y = 1)}{n_{total}} = \frac{250}{416} = 0.6009$$

Selanjutnya akan dibentuk node ke-1 untuk pohon pertama dengan mencari nilai *entropy* dan juga g*ain* untuk variabel yang ditentukan secara acak. Pada node selanjutnya juga akan dilakukan pengacakan variabel. Pada node pertama ini didapatkan variabel X_2, X_3, X_4, X_5 yang akan dihitung nilai *entropy* dan juga g*ain*, yaitu sebagai berikut:

Tabel 4.8 Perhitungan *entropy* dan juga *gain* untuk variabel X_2,X_3,X_4,X_5

Atribut	Label	Jumlah	1	0	p_1	p_0	Entropy	Gain
Total		416	250	166	0.6009	0.3991	0.9704	
$\boldsymbol{X_2}$	0	150	132	18	0.88	0.12	0.5293	0.1459
	1	266	118	148	0.4436	0.5563	0.9908	
$\boldsymbol{X_3}$	0	216	59	157	0.2731	0.7268	0.8459	0.4038
	1	200	191	9	0.955	0.045	0.2648	
$\boldsymbol{X_4}$	0	228	72	156	0.3158	0.6842	0.8997	0.3418
	1	188	178	10	0.9468	0.0532	0.2998	
$\boldsymbol{X_5}$	0	250	105	145	0.42	0.58	0.9814	0.1619
	1	166	145	21	0.8735	0.1265	0.5477	

$$Entropi(Y) = -\sum_{i=1}^{2} P_i log_2 P_i$$

$$= -(P(Y=0)log_2 P(Y=0) + (P(Y=1)log_2 P(Y=1)$$

$$= -((0.3991)log_2(0.3991) + (0.6009)log_2(0.6009))$$

$$= 0.9703$$

$$Entropi(X_2 = 0) = -(P(Y=0|X_2=0)log_2 P(Y=0|X_2=0) +$$

$$P(Y=1|X_2=0)log_2 P(Y=1|X_2=0))$$

$$= -((0.12)log_2(0.12) + (0.88)log_2(0.88))$$

$$= 0.5293$$

$$Entropi(X_2 = 1) = -(P(Y=0|X_2=1)log_2 P(Y=0|X_2=1) +$$

$$P(Y=1|X_2=1)log_2 P(Y=1|X_2=1))$$

$$= -((0.5563)log_2(0.5563) + (0.4436)log_2(0.4436))$$

$$= 0.9908$$

$$Entropi(X_3 = 0) = -(P(Y=0|X_3=0)log_2 P(Y=0|X_3=0) +$$

$$P(Y = 1|X_3 = 0)log_2P(Y = 1|X_3 = 0))$$

$$= -((0.7268)log_2(0.7268) + (0.2731)log_2(0.2731))$$

$$= 0.8459$$

$$Entropi(X_3 = 1) = -(P(Y = 0|X_3 = 1)log_2P(Y = 0|X_3 = 1) +$$

$$P(Y = 1|X_3 = 1)log_2P(Y = 1|X_3 = 1))$$

$$= -((0.054)log_2(0.054) + (0.955)log_2(0.955))$$

$$= 0.2647$$

$$Entropi(X_4 = 0) = -(P(Y = 0|X_4 = 0)log_2P(Y = 0|X_4 = 0) +$$

$$P(Y = 1|X_4 = 0)log_2P(Y = 1|X_4 = 0))$$

$$= -((0.6842)log_2(0.6842) + (0.3157)log_2(0.3157))$$

$$= 0.8997$$

$$Entropi(X_4 = 1) = -(P(Y = 0|X_4 = 1)log_2P(Y = 0|X_4 = 1) +$$

$$P(Y = 1|X_4 = 1)log_2P(Y = 1|X_4 = 1))$$

$$= -((0.0532)log_2(0.0532) + (0.9468)log_2(0.9468))$$

$$= 0.2998$$

$$Entropi(X_5 = 0) = -(P(Y = 0|X_5 = 0)log_2P(Y = 0|X_5 = 0) +$$

$$P(Y = 1|X_5 = 0)log_2P(Y = 1|X_5 = 0))$$

$$= -((0.58)log_2(0.58) + (0.42)log_2(0.42))$$

$$= 0.9814$$

$$Entropi(X_5 = 1) = -(P(Y = 0|X_5 = 1)log_2P(Y = 0|X_5 = 1) +$$

$$P(Y = 1|X_5 = 1)log_2P(Y = 1|X_5 = 1))$$

$$= -((0.1265)log_2(0.1265) + (0.8734)log_2(0.8734))$$

$$= 0.5477$$

$$Gain(y, x_2) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{2i}}{n_{total}} * Entropi\ (x_{2i})$$

$$= 0.9703 - ((\frac{150}{416} * 0.5293) + (\frac{266}{416} * 0.9990))$$

$$= 0.1459$$

$$Gain(y, x_3) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{3i}}{n_{total}} * Entropi\ (x_{3i})$$

$$= 0.9703 - ((\frac{216}{416} * 0.8459) + (\frac{200}{416} * 0.2647))$$

$$= 0.4038$$

$$Gain(y, x_4) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{4i}}{n_{total}} * Entropi\ (x_{4i})$$

$$= 0.9703 - ((\frac{228}{416} * 0.8997) + (\frac{188}{416} * 0.2998))$$

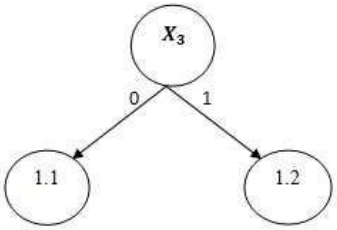
$$= 0.3417$$

$$Gain(y, x_5) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{5i}}{n_{total}} * Entropi\ (x_{5i})$$

$$= 0.9703 - ((\frac{250}{416} * 0.9814) + (\frac{166}{416} * 0.5477))$$

$= 0.1619$

Berdasarkan hasil diatas, maka didapatkan nilai gain tertinggi yaitu pada X_3 (kelainan produksi air seni atau *polyuria)* sebesar 0.4038. Dengan hasil ini, maka X_3 menjadi akar pohon pertama. Selanjutnya akan di cari cabang dari pohon X_3 yang memiliki dua katagori. Cara perhitungan sama seperti yang dilakukan di atas. Pohon untuk *root node* dapat dilihat pada gambar di bawah ini.



Gambar 4.5 Pohon keputusan *root node*

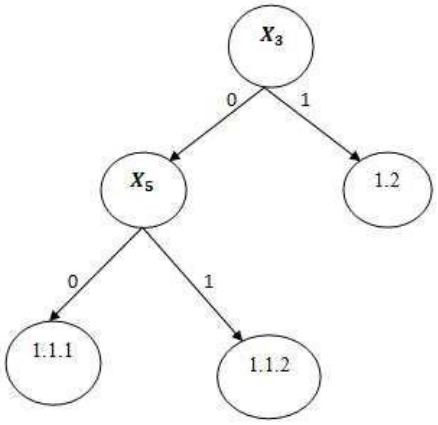
Selanjutnya melakukan perhitungan untuk cabang X_3. Dimana perhitungan ini untuk katagori *No* (0) terlebih dahulu. Pada variabel X_3 memiliki 216 kasus dengan 59 orang termasuk penderita diabetes dan 157 orang termasuk yang tidak menderita diabetes. Sebelum melakukan perhitungan *entropy* dan *gain*, dilakukan pengacakan variabel terlebih dahulu. Pilih 4 variabel untuk dilakukan pengacakan, kecuali X_3 karena variebl ini sudah menjadi akar pohon atau *root node.* Variabel yang terpilih yaitu X_5,X_6,X_{10},X_{14}. Perhitungan untuk *entropy* dan *gain* pada variabel tersebut dapat dilihat pada tabel berikut ini.

Tabel 4.9 Perhitungan *entropy* dan juga *gain* untuk variabel X_5,X_6,X_{10},X_{14}

Atribut	Label	Jumlah	1	0	p_1	p_0	Entropy	Gain
Total		216	59	127	0.2731	0.7268	0.8459	
X_5								0.0512
	0	177	38	139	0.2146	0.7853	0.7503	

	1	39	21	18	0.5384	0.4615	0.9957	
X_6	0	116	33	83	0.2844	0.7155	0.8267	0.0005
	1	100	26	74	0.26	0.74	0.8614	
X_{10}	0	120	34	86	0.2833	0.7166	0.8273	0.0004
	1	96	25	71	0.2604	0.7395	0.8599	
X_{14}	0	151	39	112	0.2582	0.7417	0.8241	0.0018
	1	65	20	45	0.3076	0.6923	0.8904	

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu X_5 (kehilangan berat badan secara drastis/*sudden weight lost*) dengan nilai *gain* sebesar 0.0512. Variabel X_5 memiliki 2 katagori yaitu katagori *No* (0) dan *Yes* (1). Pohon untuk cabang pertama pada variabel X_3 dapat dilihat berikut ini.



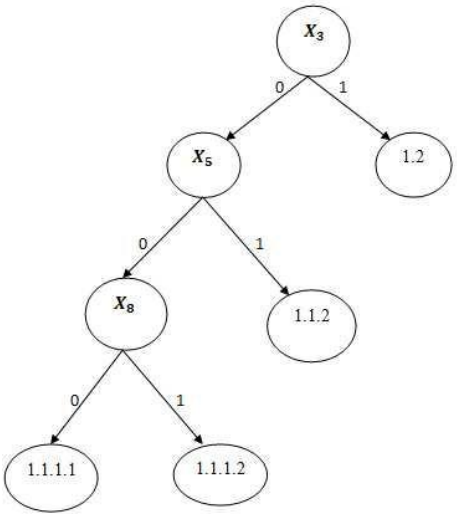
Gambar 4.6 Pohon keputusan *node* 1.1

Selanjutnya mencari cabang dari variabel X_5 dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari variabel yang telah digunakan sebagai *node*. Pada node 1.1.1 variabel yang terpilih untuk digunakan yaitu X_2, X_8 , X_{11}, X_{16} . Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.8 berikut ini.

Tabel 4.10 Perhitungan *entropy* dan juga *gain* untuk variabel X_2,X_8,X_{11},X_{16}

Atribut	Label	Jumlah	1	0	p_1	p_0	Entropy	Gain
Total		177	38	139	0.2146	0.7853	0.7503	
X_2	0	38	22	16	0.5789	0.4210	0.9819	0.1349
	1	139	16	123	0.1151	0.8848	0.5151	
X_8	0	142	30	112	0.2112	0.7887	0.7438	1.1938
	1	35	8	27	0.2285	0.7714	0.7755	
X_{11}	0	157	28	129	0.1783	0.8216	0.6764	0.0373
	1	20	10	10	0.5	0.5	1	
X_{14}	0	162	32	130	0.1975	0.8024	0.7169	0.0118
	1	15	6	9	0.4	0.6	0.9709	

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu X_8 (infeksi jamur/*Genital trush*) dengan nilai *gain* sebesar 1.1938. Variabel X_8 memiliki 2 katagori yaitu katagori *No* (0) dengan 142 orang dan *Yes* (1) dengan 35 orang. Pohon untuk cabang node 1.1.1 pada variabel X_8 dapat dilihat berikut ini.



Gambar 4.7 Pohon keputusan *node* 1.1.1

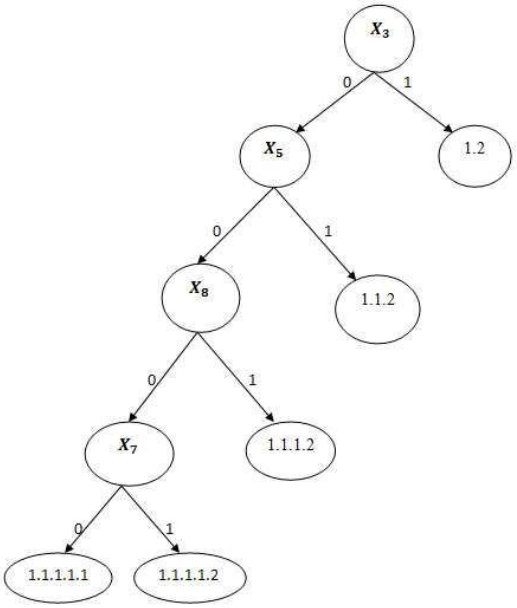
Selanjutnya mencari cabang dari variabel X_8 dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari

variabel yang telah digunakan sebagai *node*. Pada node 1.1.1.1 variabel yang terpilih untuk digunakan dalam perhitungan cabang selanjutnya yaitu $\boldsymbol{X_7}$, $\boldsymbol{X_9}$, $\boldsymbol{X_{11}}$, $\boldsymbol{X_{15}}$ **.** Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.9 berikut ini.

Tabel 4.11 Perhitungan *entropy* dan juga *gain* untuk variabel X_7,X_9,X_{11},X_{15}

Atribut	Label	Jumlah	1	0	p_1	p_0	Entropy	Gain
Total		142	30	112	0.2112	0.7887	0.7438	
$\boldsymbol{X_7}$	0	109	18	91	0.1651	0.8348	0.6464	0.0279
	1	33	12	21	0.3636	0.63636	0.9456	
$\boldsymbol{X_9}$	0	88	13	75	0.1477	0.8522	0.6041	0.0277
	1	54	17	37	0.3148	0.6851	0.8986	
$\boldsymbol{X_{11}}$	0	127	25	102	0.1968	0.8031	0.7155	0.0069
	1	15	5	10	0.3333	0.6667	0.9182	
$\boldsymbol{X_{15}}$	0	89	24	65	0.2696	0.7303	0.8409	0.0266
	1	53	6	47	0.1132	0.8867	0.5095	

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu $\boldsymbol{X_7}$ (peningkatan nafsu makan berlebih/*Polyphagia*) dengan nilai *gain* sebesar 0.0279. Variabel $\boldsymbol{X_7}$ memiliki 2 katagori yaitu katagori *No* (0) dengan 109 orang dan *Yes* (1) dengan 33 orang. Pohon untuk cabang node 1.1.1.1 pada variabel $\boldsymbol{X_7}$ dapat dilihat berikut ini.



Gambar 4.8 Pohon keputusan *node* 1.1.1.1

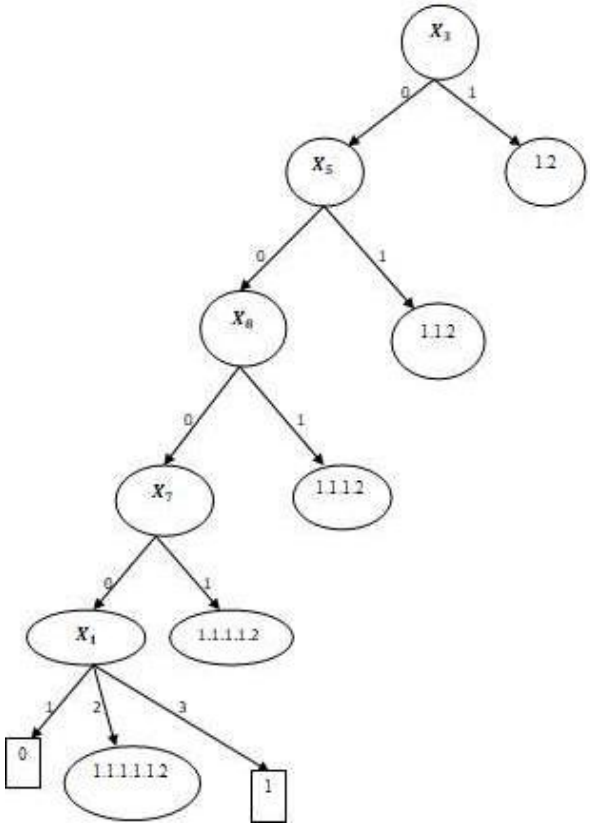
Selanjutnya mencari cabang dari variabel $\boldsymbol{X_7}$ dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari variabel yang telah digunakan sebagai *node*. Pada node 1.1.1.1.1 variabel yang terpilih untuk digunakan dalam perhitungan cabang selanjutnya yaitu $\boldsymbol{X_1}$, $\boldsymbol{X_6}$, $\boldsymbol{X_{12}}$, $\boldsymbol{X_{13}}$. Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.10 berikut ini.

Tabel 4.12 Perhitungan *entropy* dan juga *gain* untuk variabel X_1,X_6,X_{12},X_{13}

Atribut	Label	Jumlah	1	0	p_1	p_0	Entropy	Gain
Total		109	18	91	0.1651	0.8348	0.6464	
$\boldsymbol{X_1}$	1	1	0	1	0	1	0	0.1574
	2	102	12	90	0.1176	0.8823	0.5225	
	3	6	6	0	1	0	0	
$\boldsymbol{X_6}$	0	72	17	55	0.2361	0.7638	0.7885	0.0647
	1	37	1	36	0.0270	0.9729	0.1792	
$\boldsymbol{X_{12}}$	0	82	16	66	0.1951	0.8048	0.7120	0.0164

	1	27	2	25	0.0740	0.9259	0.3809	
X_{13}	0	89	10	79	0.1123	0.8876	0.5069	0.0543
	1	20	8	12	0.4	0.6	0.9709	

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu X_1 (umur) dengan nilai *gain* sebesar 0.1574. Variabel X_1 memiliki 3 katagori yaitu katagori 1 (umur kurang dari 20 tahun) dengan jumlah 1 orang yang telah masuk dalam klasifikasi tidak menderita diabetes atau negatif, katagori 2 (umur 20 tahun sampai dengan 40 tahun) dengan jumlah 102 orang dan katagori 3 (umur diatas 40 tahun) dengan jumlah 6 orang yang masuk dalam katagori penderita diabetes atau positif diabetes. Katagori 2 pada variabel X_1 belum memberikan hasil seperti 2 katagori lainya, maka akan dilakukan perhitungan untuk cabang selanjutnya. Pohon untuk cabang node 1.1.1.1.1 pada variabel X_1 dapat dilihat berikut ini.

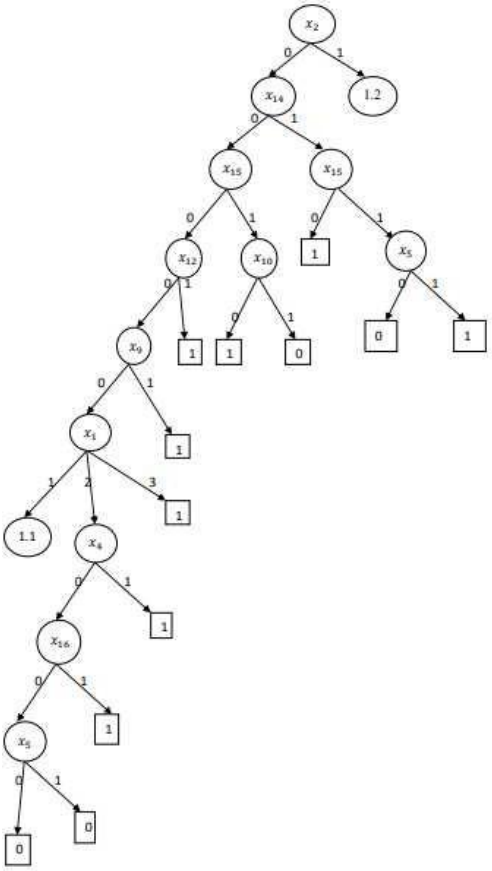


Gambar 4.9 Pohon keputusan pertama

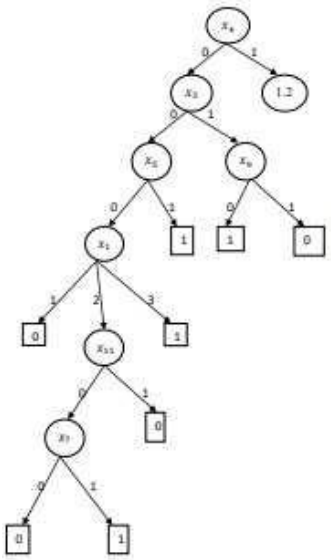
Proses perhitungan dilakukan sampai semua cabang sudah pada *terminal node* atau tidak ada lagi cabang yang tesisa. Setelah membentuk 1 pohon maka akan terlihat model pengkondisian yang akan digunakan sebagai penentu prediksi klasifikasi pada data *test.* Model pengkondisian dapat dilihat pada **Lampiran.**

Langkah selanjutnya yaitu mencari pohon keputusan dari pohon ke 2 sampai dengan pohon ke 1820 dengan cara perhitungan yang sama. Pertama, melakukan *bootstrap* pada data *training* untuk setiap pohon. kemudian melakukan pengambilan 4 variabel secara acak untuk menentukan node. Dari ke 4 variabel tersebut akan dilakukan perhitungan *entropy* dan juga *gain*. Variabel yang

memiliki nilai *gain* tertinggi akan menjadi root node/ akar pohon. Ulangi proses perhitungan sampai semua cabang pohon selesai dicari atau telah mencapai *terminal node.* Beberapa pohon keputusan dari *random forest* dapat dilihat berikut ini :



Gambar 4.10 Pohon keputusan kedua



Gambar 4.11 Pohon keputusan ketiga

Semua pohon yang telah dibangun akan dibuat pengkodisian sehingga dapat dilakukan proses klasifikasi dengan cara *majorty voting*. Setiap data *test* memiliki 1820 klasifikasi berdasarkan pengkondisian dari 1820 pohon. Penggabungan dari masing-masing pohon keputusan pada setiap data *test* merupkan hasil klasifikasi akhir. Hasil klasifikasi akhir dari 1820 pohon dengan menggunakan *Random Forest* dapat dilihat pada berikut ini.

Tabel 4.13 *confusion matrix* metode *Random Forest*

	True 1	*True* 0
Predict 1	62	0
Predict 0	2	40

Berdasarkan tabel yang ada diatas dapat dilihat bahwa 62 orang di prediksi benar masuk dalam klasifikasi positif diabetes, sedangkan 40 orang diprediksi

benar masuk dalam klasifikasi negatif diabetes. Terdapat 2 orang di prediksi masuk dalam katagori negatif diabetes, namun ternyata masuk dalam katogori positif diabetes pada data sebenarnya. Sedangkan tidak ada orang yang diprediksi masuk dalam katagori positif diabetes dan negatif diabetes pada data sebenarnya. Kita juga dapat melihat tingkat akurasi, presisi, recall, specificity, dan F1 score dari tabel diatas berikut ini.

$$akurasi = \frac{62 + 40}{104} = 0.9808$$

$$presisi = \frac{62}{62 + 0} = 1$$

$$recall = \frac{62}{62 + 2} = 0.9688$$

$$specificity = \frac{40}{40 + 0} = 1$$

$$F1\ score = 2\,(\frac{0.9687 * 1}{0.9687 + 1}) = 0.9841$$

Jadi, dengan menggunakan metode *random forest* akurasi yang didapatkan sebesar 0.9808, presisi sebesar 1. Recall sebesar 0.9687, specificity sebesar 1 dan F1 score sebesar 0.9841, hal ini menunjukkan bahwa ketepatan akurasi dalam memprediksi klasifikasi data penyakit diabetes dengan menggunakan metode *random forest* adalah 98.08%.

4.4 Perbandingan Tingkat Ketepatan Klasifikasi

Berdasarkan perhitungan yang telah dilakukan maka didapatkan hasil perbandingan akurasi dari kedua metode yang dapat dilihat pada tabel berikut ini.

Tabel 4.14 perbandingan akurasi 2 metode

Tingkat ketepatan	Metode Decision Tree C4.5	Metode Random Forest
Akurasi	91.35%	98.08%.
Presisi	93.55%	100%
Recall	92.06%	96.88%
Specificity	90.24%	100%
F1 score	92.80%	98.41%

Dari perbandingan tingkat akurasi tersebut, terlihat bahwa metode *decision tree* C4.5 memiliki akurasi sebesar 91.35% sedangkan pada metode *random forest* akurasi yang didapatkan lebih tinggi yaitu sebesar 98.08%. Begitupun dengan nilai dari presisi, recall, specificity dan juga Fi score pada *metode random forest* lebih tinggi dibandingkan dengan nilai dari metode *decision tree* C4.5. Jika variabel umur pada data penyakit diabetes tidak di diskritisasi pada metode *decision tree* maka nilai akurasi, presisi, recall, specificity dan F1 score yang didapat secara berturut-turut sebesar 85.58%, 86.18%, 85.71%, 85.58%, dan 85.77%. Sedangkan pada metode *random forest* nilai akurasi, presisi, recall, specificity, dan F1 score untuk variabel umur tanpa diskritisasi secara berturut-turut sebesar 75.56%, 76.55%, 78.92%, 77.65%, dan 76.11%. Untuk data

prediksi penyakit diabetes dari pasien *Diabetes Sylhet* Rumah sakit di Sylhet, Bangladesh metode *random forest* dinilai lebih baik dibandingkan dengan metode *decision tree*. Metode *Random Forest* memiliki tingkat akurasi yang lebih baik karena menggabungkan hasil dari klasifikasi masing-masing pohon kemudian dilakukan *majorty voting* sehingga dapat mengurangi tingkat kesalahan pada proses prediksi klasifikasi.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Metode *decision tree* memprediksi 58 orang benar masuk dalam klasifikasi positif diabetes, sedangkan 37 orang diprediksi benar masuk dalam klasifikasi negatif diabetes. Terdapat 5 orang di prediksi masuk dalam katagori positif diabetes, namun ternyata masuk dalam katagori negatif diabetes pada data sebenarnya. Sedangkan 4 orang yang diprediksi masuk dalam katagori negatif diabetes ternyata masuk dalam katagori positif diabetes pada data sebenarnya. Untuk metode *random forest* memprediksi 62 orang benar masuk dalam klasifikasi positif diabetes, sedangkan 40 orang diprediksi benar masuk dalam klasifikasi negatif diabetes. Terdapat 2 orang di prediksi masuk dalam katagori negatif diabetes, namun ternyata masuk dalam katogori positif diabetes pada data sebenarnya. Sedangkan tidak ada orang yang diprediksi masuk dalam katagori positif diabetes dan negatif diabetes pada data sebenarnya. Pada metode *random forest* membentuk 1820 pohon keputusan dimana hasil akhir dari klasifikasi pada setiap data test berdasarkan gabungan klasifikasi yang paling banyak dari 1820 pohon.
2. Tingkat akurasi, presisi, recall, specificity dan F1 score pada metode *decision tree* secara berturut-turut sebesar 91.35%, 93.55%, 92.06%, 90.24%, dan 92.80% , sedangkan pada metode *random forest* nilai akurasi,

presisi, recall, specificity dan F1 score secara berturut-turut adalah 98.08%, 100%, 96.88%, 100%, dan 98.41%. Pada penelitian penyakit diabetes ini metode *random forest* dinilai lebih baik dibandingkan dengan metode *decision tree.*

5.2 Saran

Pada penelitian ini penulis hanya menggunakan 2 metode untuk melakukan klasifikasi. Dalam statistika dan data mining terdapat banyak metode yang bisa digunakan untuk pengklasifikasian. Sehingga penulis menyarankan untuk penelitian selanjutnya dapat menggunakan lebih banyak lagi metode untuk mencari nilai akurasi terbaik dari setiap metode yang digunakan.

DAFTAR PUSTAKA

Alsalman, Y. S., Khamees Abu Halemah, N., Alnagi, E. S., & Salameh, W. (2019). Using Decision Tree and Artificial Neural Network to Predict Students Academic Performance. *2019 10th International Conference on Information and Communication Systems, ICICS 2019*, 104–109. https://doi.org/10.1109/IACS.2019.8809106

Anggraeni, D., & Ramadhani. (2018). *Analisa Decision Tree Untuk Prediksi Diagnosa Diabetes Mellitus*. *9986*(September).

Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. *Journal of Information Technology Ampera*, *1*(3), 133–143. https://doi.org/10.51519/journalita.volume1.isssue3.year2020.page133-143

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Media Informatika Budidarma*, *5*(2), 640–651. https://doi.org/10.30865/mib.v5i2.2937

Chan, A. S. (2018). Prediksi Kedatangan Wisatawan Pada Pariwisata Kota Batam Dengan Menggunakan Teknik Knowledge Data Discovery. *Jurnal Ilmiah Informatika*, *6*(01), 11. https://doi.org/10.33884/jif.v6i01.432

Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. (2017). Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing. *Proceedings - International Computer Software and Applications Conference*, *2*, 481–484. https://doi.org/10.1109/COMPSAC.2017.146

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, *138*, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Dennedy, M. C., Rizza, R. A., & Dinneen, S. F. (2015). Classification and Diagnosis of Diabetes Mellitus. *Endocrinology: Adult and Pediatric*, *1–2*(January), 662-671.e2. https://doi.org/10.1016/B978-0-323-18907-1.00038-X

Eska, J. (2016). Penerapan Data Mining Untuk Prediksi Penjualan Wallpaper Menggunakan Algoritma C4.5. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *2*, 9 – 13. https://doi.org/10.31227/osf.io/x6svc

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26. https://doi.org/10.9781/ijimai.2018.02.004

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *4*(1), 32–39. https://doi.org/10.47970/siskom-kb.v4i1.173

Jiang, W., Liu, G., Zhao, X., & Yang, F. (2019). Cross-Subject Emotion Recognition with a Decision Tree Classifier Based on Sequential Backward Selection. *Proceedings - 2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2019*, *1*, 309–313. https://doi.org/10.1109/IHMSC.2019.00078

Kurniawan, H. (2020). Deteksi Twitter Bot Menggunakan Klasifikasi Decision Tree. *Jurnal Sustainable: Jurnal Hasil Penelitian Dan Industri Terapan*, *09*(2), 31–37. http://ejournal.bsi.ac.id/ejurnal/index.php/cakrawala/article/view/3680/2624%0Ahttp://j-ptiik.ub.ac.id

Liu, J., Ning, B., & Shi, D. (2019). Application of Improved Decision Tree C4.5 Algorithms in the Judgment of Diabetes Diagnostic Effectiveness. *Journal of Physics: Conference Series*, *1237*(2). https://doi.org/10.1088/1742-6596/1237/2/022116

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, *2*(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

Naik, J., & Patel, P. S. (2013). *Tumor Detection and Classification using Decision Tree in Brain MRI*. 49–53.

Najib, A., Nurcahyono, D., & Setiawan, R. P. P. (2019). Klasifikasi Diagnosa Penyakit Diabetes Mellitus (Dm) Menggunakan Algoritma C4.4. *Just TI (Jurnal Sains Terapan Teknologi Informasi)*, *11*(2), 47. https://doi.org/10.46964/justti.v11i2.153

NAsrullah, A. H. (2021). *IMPLEMENTASI ALGORITMA DECISION TREE UNTUK*. *7*(2), 45–51.

Nasution, M. Z. F., Sitompul, O. S., & Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification. *Journal of Physics: Conference Series*, *978*(1). https://doi.org/10.1088/1742-6596/978/1/012058

Novianti, B., Rismawan, T., & Bahri, S. (2016). Implementasi Data Mining

Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak). *Jurnal Coding, Sistem Komputer Untan*, *04*(3), 75–84.

Nugroho, K., Noersasongko, E., Purwanto, Muljono, Fanani, A. Z., Affandy, & Basuki, R. S. (2019). Improving random forest method to detect hatespeech and offensive word. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 514–518. https://doi.org/10.1109/ICOIACT46704.2019.8938451

Otok, B. W., & Nidhomuddin. (2015). Random Forest Dan Multivariate Adaptive Regression Spline (Mars) Binary Response Untuk Klasifikasi Penderita Hiv / Aids Di Surabaya. *Statistika Fakultas Matematika Dan Ilmu Pengetahuan Alam Institut Teknologi Sepuluh November*, *1*(3), 50–57.

Parung, F. (2018). Penerapan algoritma decision tree c4.5 dalam penerimaan guru pada smk sirajul falah parung. *CKI On SPOT*, *11*(2), 192–198.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved Random Forest for Classification. *IEEE Transactions on Image Processing*, *27*(8), 4012–4024. https://doi.org/10.1109/TIP.2018.2834830

Punthakee, Z., Goldenberg, R., & Katz, P. (2018). Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome. *Canadian Journal of Diabetes*, *42*, S10–S15. https://doi.org/10.1016/j.jcjd.2017.10.003

Putry, N. M., Sari, B. N., Kom, M., Informatika, T., & Karawang, U. S. (2022). *KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS*. *10*(1).

Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk. *JURNAL RESTI*, *1*(10), 187–192.

Saifullah, S., Zarlis, M., Zakaria, Z., & Sembiring, R. W. (2017). Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *1*(2), 180. https://doi.org/10.30645/j-sakti.v1i2.41

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 64–71.

Siallagan, R. A., & Fitriyani. (2021). Prediksi Penyakit Diabetes Mellitus. *Jurnal Responsif*, *3*(1), 45–46.

Sunjana. (2010). Aplikasi mining data mahasiswa dengan metode klasifikasi decision tree. *Snati*.

Syamsu, S., Muhajirin, M., & Wijaya, N. S. (2019). Rules Generation Untuk Klasifikasi Data Bakat dan Minat Berdasarkan Rumpun Ilmu Dengan Decision Tree. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *9*(1), 40. https://doi.org/10.35585/inspir.v9i1.2495

Tanujayaa, L. B. C., Susanto, B., & Saragiha, A. (2020). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify. *Indonesian Journal of Data and Science (IJODAS)*, *1*(2715–9930), 68–78.

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 440. https://doi.org/10.30865/mib.v4i2.2080

Widiyati, D. K., Wati, M., & Pakpahan, H. S. (2018). Penerapan Algoritma ID3 Decision Tree Pada Penentuan Penerima Program Bantuan Pemerintah Daerah di Kabupaten Kutai Kartanegara. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, *2*(2), 126. https://doi.org/10.30872/jurti.v2i2.1864

Wulandari, P., Sugiyanto, Z., & Kresnowati, L. (2015). Ananlisis Faktor Penyebab Kadar Gula darah Pada Penderita Diabetes Melitus (DM) Tipe-2 Di RSUD Tugurejo Semarang. *Jurnal Kesehatan*, *14*(2), 353–360.

Yezli, S., Yassin, Y., Mushi, A., Balkhi, B., & Khan, A. (2021). Insulin Knowledge , Handling , and Storage among Diabetic Pilgrims during the Hajj Mass Gathering. *Journal of Diabetes Research*, *2021*.

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, *49*(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485

Zhang, X. F., & Tan, B. K. H. (2000). Effects of an ethanollic extract of Gynura procumbens on serum glucose, cholesterol and triglyceride levels in normal and streptozotocin-induced diabetic rats. *Singapore Medical Journal*, *41*(1), 9–13.

Zhou, X., Lu, P., Zheng, Z., Tolliver, D., & Keramati, A. (2020). Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree. *Reliability Engineering and System Safety*, *200*, 9. https://doi.org/10.1016/j.ress.2020.106931

LAMPIRAN

1. Pengkondisian *Decision Tree* menggunakan software rapid miner

```
Polydipsia = 0
|   Age = 1: 0 {1=0, 0=1}
|   Age = 2
|   |   Polyuria = 0
|   |   |   Gender = 0
|   |   |   |   Alopecia = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   muscle_stiffness = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   Irritability = 0: 1 {1=6, 0=2}
|   |   |   |   |   |   |   |   Irritability = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   |   Polyphagia = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   muscle_stiffness = 1: 1 {1=4, 0=0}
|   |   |   |   |   visual_blurring = 1: 1 {1=11, 0=0}
|   |   |   |   Alopecia = 1: 0 {1=0, 0=12}
|   |   |   Gender = 1
|   |   |   |   Irritability = 0
|   |   |   |   |   Alopecia = 0: 0 {1=0, 0=68}
|   |   |   |   |   Alopecia = 1
|   |   |   |   |   |   Itching = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   weakness = 0: 0 {1=0, 0=6}
|   |   |   |   |   |   |   |   weakness = 1: 1 {1=2, 0=2}
|   |   |   |   |   |   |   Polyphagia = 1: 1 {1=1, 0=0}
|   |   |   |   |   |   Itching = 1: 0 {1=0, 0=38}
|   |   |   |   Irritability = 1
|   |   |   |   |   Genital_thrush = 0
|   |   |   |   |   |   weakness = 0: 0 {1=0, 0=5}
|   |   |   |   |   |   weakness = 1
|   |   |   |   |   |   |   visual_blurring = 0: 1 {1=1, 0=0}
|   |   |   |   |   |   |   visual_blurring = 1: 0 {1=0, 0=4}
|   |   |   |   |   Genital_thrush = 1: 1 {1=3, 0=0}
|   |   Polyuria = 1
|   |   |   Itching = 0: 1 {1=20, 0=0}
|   |   |   Itching = 1
|   |   |   |   delayed_healing = 0: 1 {1=7, 0=0}
|   |   |   |   delayed_healing = 1
|   |   |   |   |   Gender = 0: 1 {1=1, 0=0}
|   |   |   |   |   Gender = 1: 0 {1=0, 0=11}
|   Age = 3: 1 {1=19, 0=0}
Polydipsia = 1
|   Polyuria = 0
|   |   Irritability = 0
|   |   |   Gender = 0: 1 {1=5, 0=0}
|   |   |   Gender = 1
|   |   |   |   muscle_stiffness = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   Itching = 0: 1 {1=2, 0=0}
|   |   |   |   |   |   Itching = 1: 0 {1=0, 0=1}
|   |   |   |   visual_blurring = 1: 1 {1=3, 0=0}
```

```
|   |   |   |   muscle_stiffness = 1
|   |   |   |   |   Age = 2: 0 {1=0, 0=6}
|   |   |   |   |   Age = 3: 1 {1=1, 0=0}
|   |   Irritability = 1: 1 {1=17, 0=0}
|   Polyuria = 1: 1 {1=155, 0=0}
```

Lampiran 2. Pengkondisian pohon keputusan pertama *random forest* menggunakan software rapid miner

```
Gender = 0
|   Alopecia = 0
|   |   Polydipsia = 0
|   |   |   weakness = 0
|   |   |   |   muscle_stiffness = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   partial_paresis = 0
|   |   |   |   |   |   |   Obesity = 0: 1 {1=6, 0=1}
|   |   |   |   |   |   |   Obesity = 1: 1 {1=3, 0=0}
|   |   |   |   |   |   partial_paresis = 1: 0 {1=0, 0=1}
|   |   |   |   |   visual_blurring = 1: 1 {1=14, 0=0}
|   |   |   |   muscle_stiffness = 1: 1 {1=8, 0=0}
|   |   |   weakness = 1: 1 {1=15, 0=0}
|   |   Polydipsia = 1: 1 {1=92, 0=0}
|   Alopecia = 1
|   |   Age = 2
|   |   |   visual_blurring = 0: 0 {1=0, 0=2}
|   |   |   visual_blurring = 1
|   |   |   |   sudden_weight_loss = 0: 0 {1=0, 0=5}
|   |   |   |   sudden_weight_loss = 1: 1 {1=4, 0=0}
|   |   Age = 3: 1 {1=2, 0=0}
Gender = 1
|   Age = 2
|   |   Obesity = 0
|   |   |   Polydipsia = 0
|   |   |   |   Irritability = 0
|   |   |   |   |   delayed_healing = 0
|   |   |   |   |   |   visual_blurring = 0: 0 {1=0, 0=76}
|   |   |   |   |   |   visual_blurring = 1
|   |   |   |   |   |   |   partial_paresis = 0: 0 {1=0, 0=4}
|   |   |   |   |   |   |   partial_paresis = 1: 1 {1=1, 0=0}
|   |   |   |   |   delayed_healing = 1
|   |   |   |   |   |   Polyuria = 0
|   |   |   |   |   |   |   muscle_stiffness = 0
|   |   |   |   |   |   |   |   Genital_thrush = 0: 0 {1=1, 0=11}
|   |   |   |   |   |   |   |   Genital_thrush = 1: 0 {1=0, 0=11}
|   |   |   |   |   |   |   muscle_stiffness = 1: 0 {1=0, 0=16}
|   |   |   |   |   |   Polyuria = 1
|   |   |   |   |   |   |   sudden_weight_loss = 0: 0 {1=0, 0=10}
|   |   |   |   |   |   |   sudden_weight_loss = 1: 1 {1=3, 0=0}
|   |   |   |   Irritability = 1
|   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   weakness = 0: 0 {1=0, 0=1}
|   |   |   |   |   |   weakness = 1: 1 {1=5, 0=0}
|   |   |   |   |   Polyphagia = 1
|   |   |   |   |   |   partial_paresis = 0
|   |   |   |   |   |   |   visual_blurring = 0: 1 {1=3, 0=0}
|   |   |   |   |   |   |   visual_blurring = 1
```

```
|   |   |   |   |   |   |   |   sudden_weight_loss = 0: 0 {1=0,
0=5}
|   |   |   |   |   |   |   |   sudden_weight_loss = 1: 1 {1=1,
0=0}
|   |   |   |   |   |   partial_paresis = 1: 1 {1=4, 0=0}
|   |   |   Polydipsia = 1
|   |   |   |   Genital_thrush = 0
|   |   |   |   |   Polyuria = 0
|   |   |   |   |   |   Irritability = 0: 0 {1=0, 0=4}
|   |   |   |   |   |   Irritability = 1: 1 {1=2, 0=0}
|   |   |   |   |   Polyuria = 1: 1 {1=21, 0=0}
|   |   |   |   Genital_thrush = 1: 1 {1=22, 0=0}
|   |   Obesity = 1
|   |   |   Alopecia = 0
|   |   |   |   muscle_stiffness = 0
|   |   |   |   |   Polydipsia = 0: 0 {1=0, 0=8}
|   |   |   |   |   Polydipsia = 1
|   |   |   |   |   |   partial_paresis = 0: 1 {1=1, 0=0}
|   |   |   |   |   |   partial_paresis = 1: 0 {1=0, 0=1}
|   |   |   |   muscle_stiffness = 1
|   |   |   |   |   Polyuria = 0: 0 {1=0, 0=1}
|   |   |   |   |   Polyuria = 1: 1 {1=5, 0=0}
|   |   |   Alopecia = 1
|   |   |   |   delayed_healing = 0: 1 {1=8, 0=0}
|   |   |   |   delayed_healing = 1
|   |   |   |   |   partial_paresis = 0: 0 {1=0, 0=8}
|   |   |   |   |   partial_paresis = 1: 1 {1=3, 0=0}
|   Age = 3: 1 {1=27, 0=0}
```

Lampiran 3. Pengkondisian pohon keputusan kedua *random forest* menggunakan software rapid miner

```
Gender = 0
|   muscle_stiffness = 0
|   |   Alopecia = 0
|   |   |   delayed_healing = 0
|   |   |   |   visual_blurring = 0
|   |   |   |   |   Age = 2
|   |   |   |   |   |   Polydipsia = 0
|   |   |   |   |   |   |   Obesity = 0
|   |   |   |   |   |   |   |   sudden_weight_loss = 0: 0 {1=3,
0=6}
|   |   |   |   |   |   |   |   sudden_weight_loss = 1: 0 {1=0,
0=2}
|   |   |   |   |   |   |   Obesity = 1: 1 {1=1, 0=0}
|   |   |   |   |   |   Polydipsia = 1: 1 {1=6, 0=0}
|   |   |   |   |   Age = 3: 1 {1=7, 0=0}
|   |   |   |   visual_blurring = 1: 1 {1=21, 0=0}
|   |   |   delayed_healing = 1: 1 {1=42, 0=0}
|   |   Alopecia = 1
|   |   |   Itching = 0: 1 {1=7, 0=0}
|   |   |   Itching = 1: 0 {1=0, 0=11}
|   muscle_stiffness = 1
|   |   Alopecia = 0: 1 {1=58, 0=0}
|   |   Alopecia = 1
|   |   |   sudden_weight_loss = 0: 0 {1=0, 0=2}
|   |   |   sudden_weight_loss = 1: 1 {1=2, 0=0}
Gender = 1
|   Age = 1: 0 {1=0, 0=1}
|   Age = 2
|   |   Polydipsia = 0
|   |   |   Irritability = 0
|   |   |   |   Polyuria = 0
|   |   |   |   |   weakness = 0: 0 {1=0, 0=65}
|   |   |   |   |   weakness = 1
|   |   |   |   |   |   Itching = 0
|   |   |   |   |   |   |   Alopecia = 0: 0 {1=0, 0=8}
|   |   |   |   |   |   |   Alopecia = 1
|   |   |   |   |   |   |   |   sudden_weight_loss = 0: 1 {1=2,
0=0}
|   |   |   |   |   |   |   |   sudden_weight_loss = 1: 0 {1=0,
0=1}
|   |   |   |   |   |   Itching = 1: 0 {1=0, 0=32}
|   |   |   |   Polyuria = 1
|   |   |   |   |   Itching = 0: 1 {1=8, 0=0}
|   |   |   |   |   Itching = 1
|   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   sudden_weight_loss = 0: 1 {1=3, 0=0}
|   |   |   |   |   |   |   sudden_weight_loss = 1: 0 {1=0, 0=2}
|   |   |   |   |   |   Polyphagia = 1: 0 {1=0, 0=6}
|   |   |   Irritability = 1
```

```
|   |   |   |   visual_blurring = 0
|   |   |   |   |   Obesity = 0
|   |   |   |   |   |   weakness = 0
|   |   |   |   |   |   |   Polyphagia = 0: 0 {1=0, 0=1}
|   |   |   |   |   |   |   Polyphagia = 1: 1 {1=2, 0=0}
|   |   |   |   |   |   weakness = 1: 1 {1=9, 0=0}
|   |   |   |   |   Obesity = 1: 0 {1=1, 0=5}
|   |   |   |   visual_blurring = 1
|   |   |   |   |   Genital_thrush = 0: 0 {1=0, 0=6}
|   |   |   |   |   Genital_thrush = 1: 1 {1=3, 0=0}
|   |   Polydipsia = 1
|   |   |   visual_blurring = 0
|   |   |   |   Polyuria = 0
|   |   |   |   |   partial_paresis = 0: 1 {1=6, 0=0}
|   |   |   |   |   partial_paresis = 1
|   |   |   |   |   |   Irritability = 0: 0 {1=0, 0=1}
|   |   |   |   |   |   Irritability = 1: 1 {1=2, 0=0}
|   |   |   |   Polyuria = 1: 1 {1=16, 0=0}
|   |   |   visual_blurring = 1
|   |   |   |   muscle_stiffness = 0: 1 {1=23, 0=0}
|   |   |   |   muscle_stiffness = 1
|   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   Genital_thrush = 0
|   |   |   |   |   |   |   partial_paresis = 0: 0 {1=0, 0=3}
|   |   |   |   |   |   |   partial_paresis = 1
|   |   |   |   |   |   |   |   Irritability = 0: 0 {1=0, 0=4}
|   |   |   |   |   |   |   |   Irritability = 1: 1 {1=2, 0=0}
|   |   |   |   |   |   Genital_thrush = 1: 1 {1=3, 0=0}
|   |   |   |   |   Polyphagia = 1
|   |   |   |   |   |   Alopecia = 0: 1 {1=6, 0=0}
|   |   |   |   |   |   Alopecia = 1
|   |   |   |   |   |   |   Obesity = 0: 0 {1=0, 0=2}
|   |   |   |   |   |   |   Obesity = 1: 1 {1=1, 0=0}
|   Age = 3: 1 {1=24, 0=0}
```

Lampiran 4. Pengkondisian pohon keputusan ketiga *random forest* menggunakan software rapid miner

```
Polydipsia = 0
|   Polyuria = 0
|   |   sudden_weight_loss = 0
|   |   |   Age = 1: 0 {1=0, 0=2}
|   |   |   Age = 2
|   |   |   |   Irritability = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   Genital_thrush = 0
|   |   |   |   |   |   |   delayed_healing = 0
|   |   |   |   |   |   |   |   Itching = 0: 0 {1=2, 0=41}
|   |   |   |   |   |   |   |   Itching = 1: 0 {1=0, 0=6}
|   |   |   |   |   |   |   delayed_healing = 1
|   |   |   |   |   |   |   |   Itching = 0: 1 {1=2, 0=0}
|   |   |   |   |   |   |   |   Itching = 1: 0 {1=0, 0=12}
|   |   |   |   |   |   Genital_thrush = 1: 0 {1=0, 0=28}
|   |   |   |   |   visual_blurring = 1
|   |   |   |   |   |   partial_paresis = 0: 0 {1=0, 0=20}
|   |   |   |   |   |   partial_paresis = 1
|   |   |   |   |   |   |   weakness = 0: 1 {1=6, 0=0}
|   |   |   |   |   |   |   weakness = 1: 0 {1=0, 0=1}
|   |   |   |   Irritability = 1
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   partial_paresis = 0
|   |   |   |   |   |   |   Obesity = 0
|   |   |   |   |   |   |   |   Polyphagia = 0: 0 {1=0, 0=1}
|   |   |   |   |   |   |   |   Polyphagia = 1: 1 {1=1, 0=0}
|   |   |   |   |   |   |   Obesity = 1: 1 {1=3, 0=0}
|   |   |   |   |   |   partial_paresis = 1: 0 {1=0, 0=4}
|   |   |   |   |   visual_blurring = 1: 0 {1=0, 0=5}
|   |   |   Age = 3: 1 {1=15, 0=0}
|   |   sudden_weight_loss = 1
|   |   |   Gender = 0: 1 {1=9, 0=0}
|   |   |   Gender = 1: 0 {1=0, 0=10}
|   Polyuria = 1
|   |   visual_blurring = 0: 1 {1=30, 0=0}
|   |   visual_blurring = 1
|   |   |   delayed_healing = 0: 1 {1=4, 0=0}
|   |   |   delayed_healing = 1: 0 {1=0, 0=12}
Polydipsia = 1
|   Polyphagia = 0
|   |   partial_paresis = 0: 1 {1=39, 0=0}
|   |   partial_paresis = 1
|   |   |   delayed_healing = 0
|   |   |   |   Gender = 0: 1 {1=4, 0=0}
|   |   |   |   Gender = 1
|   |   |   |   |   weakness = 0: 0 {1=0, 0=2}
```

```
|   |   |   |   |   weakness = 1
|   |   |   |   |   |   Polyuria = 0: 0 {1=0, 0=3}
|   |   |   |   |   |   Polyuria = 1: 1 {1=2, 0=0}
|   |   |   delayed_healing = 1: 1 {1=14, 0=0}
|   Polyphagia = 1: 1 {1=138, 0=0}
```