Klasifikasi Penyakit Batu Ginjal Menggunakan Algoritma Decision Tree C4.5 Dengan Membandingkan Hasil Uji Akurasi

Yuni Widiastiwi[1], Iin Ernawati[2]

[1,2]Universitas Pembagunan Nasional Veteran Jakarta
JL. RS. Fatmawati Pondok Labu Jakarta Selatan
E-mail : widiastiwi@upnvj.ac.id[1], iin_ernawati@yahoo.com[2]

ABSTRAK

Decision tree merupakan algoritma klasifikasi yang dapat digunakan untuk mengidentifikasi faktor-faktor dari sebuah kejadian, salah satunya adalah penyakit batu ginjal. Batu ginjal merupakan salah satu penyakit penyebab terbanyak kelainan pada saluran kemih. Terbentuknya batu ginjal secara garis besar dipengaruhi oleh faktor intrinsik dan ekstrinsik. Faktor intrinsik adalah faktor yang berasal dari dalam individu sendiri yaitu umur, jenis kelamin, keturunan, atau riwayat keluarga. Faktor ekstrinsik adalah faktor yang berasal dari lingkungan luar individu seperti kebiasaan minum dan makan. Penelitian ini dilakukan untuk mengklasifikasikan dan membandingkan hasil uji akurasi terhadap dataset rekam medis penyakit batu ginjal. Pendekatan metode penelitian yang dilakukan adalah dengan menggunakan model data mining *decision tree* C4.5 untuk melakukan klasifikasi dan mendapatkan hasil akurasi terbaik dengan pendekatan pembagian data latih dan data uji dengan tiga kategori pendekatan. Hasil yang diharapkan dalam penelitian ini adalah berupa informasi pembentukan pohon keputusan dan hasil uji akurasi terbaik menggunakan data latih sebanyak 70% menghasilkan tingkat akurasi sebesar 95,71%.

Kata Kunci : Batu Ginjal, Klasifikasi, Decision Tree, Akurasi

ABSTRACT

The decision tree is one of the classification algorithms that can use to identify factors that cause an event, one of which is kidney stone disease. Kidney stones are one of the most common causes of urinary tract disorders. The formation of kidney stones commonly influenced by intrinsic and extrinsic factors. Intrinsic factors are factors that come from within the individual, namely age, gender, heredity, or family history. Extrinsic factors are factors that come from the environment outside the individual like drinking and eating habits. The aim of This study conducted to classify and compare the results of the accuracy-test against a dataset of kidney stone disease medical records. The research method approach used is to use the decision tree C4.5. To classify and get the best accuracy results with the training data and test data with three categories. The results expected in this study are in the form of decision tree formation information also the best accuracy test results using training data as much as 70% produces an accuracy rate of 95,71%.

Keywords: Kidney Stones, Classification, Decision Tree, Accuracy

1. PENDAHULUAN

Kesehatan merupakan salah satu hal penting yang menjadi prioritas utama bagi manusia untuk dapat menghadapi kehidupan dalam pekerjaan, sehingga dapat berjalan dengan baik dan optimal. Berbagai macam cara dilakukan oleh manusia dalam menjaga kesehatan, namun saat

ini akibat pola hidup yang tidak teratur, tekanan pekerjaan dan stres berkepanjangan menyebabkan munculnya berbagai penyakit. Secara perlahan namun pasti.

Batu ginjal adalah salah satu jenis penyakit yang berada pada bagian saluran kemih, penyakit batu ini dapat teratasi dengan baik apabila dapat dideteksi sedini mungkin, oleh karena itu diperlukan pengetahuan yang baik mengenai faktor-faktor penyebab seseorang terkena penyakit munculnya sejenis batu di ginjal, karena apabila faktor penyebab dapat diidentifikasikan dengan baik maka tindakan pencegahan akan lebih mudah dilakukan.

Salah satu teknik data mining semisal *decision tree* (pohon keputusan) dapat digunakan untuk melakukan klasifikasi terhadap data, proses pembuatan pohon keputusan dipengaruhi oleh penentuan atribut sebagai node terpilih (Wirdasari, 2013), dalam penelitian ini metode penentuan atribut yang dipilih adalah dengan menggunakan perhitungan gain ratio.

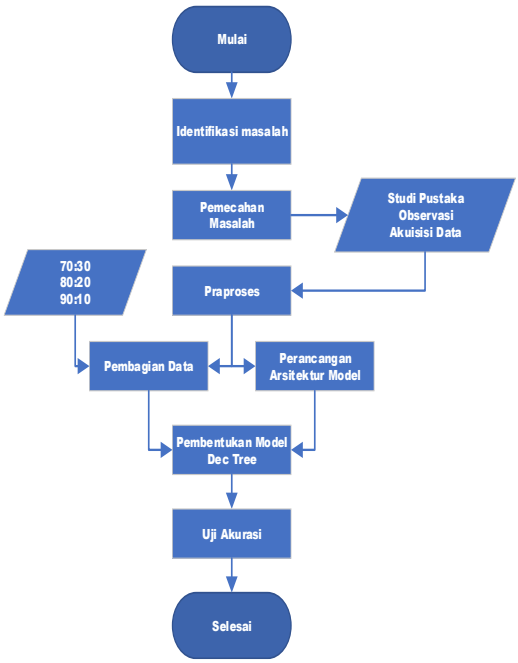
Penelitian ini berusaha untuk mencari atribut yang penting dalam pembentukan aturan/ *rule* dari pohon keputusan dan juga membandingkan hasil uji akurasi dengan membuat tiga skenario pembagian data latih mulai dari pembahgian data latih sebesar 70%, 80% dan 90%. Hal ini dilakukan untuk melihat nilai akurasi terbaik yang dapat dihasilkan.

Penggunaan algoritma *decision tree* saat ini sudah memenuhi kebutuhan untuk dapat memetakan faktor-faktor pendukung apa saja, sehingga dapat dikategorikan sebagai suspek penyakit batu ginjal pada manusia.

Hasil akhir yang diharapkan adalah dengan diketahuinya faktor-faktor penyebab suspek batu ginjal dapat dilakukan tindakan pencegahan untuk menghindari tingginya peluang faktor resiko kejadian terjadi.

2. METODOLOGI

Untuk mempermudah pada tahap pelaksanaan kegiatan penelitian, dibuatlah sebuah metodelogi yang mampu mengakomodir kebutuhan dari ruang lingkup penelitian yang dikerjakan, adapun metodologi yang digunakan dalam desain rencana penelitian tergambar berikut ini :



Gambar 1. Metodologi Penelitian

Mengacu pada Gambar 1 terdeskripsikan dengan jelas bahwa pelaksanaan penelitian dikerjakan dengan beberapa tahapan penelitian, dimana tahapan ini dapat dijadikan panduan dalam mempermudah proses penelitian.

1. Tahapan awal penelitian dimulai dari identifikasi masalah bahwa tingkat kejadian orang menderita penyakit batu ginjal cukup banyak tanpa pengetahuan yang lebih jelas terkait faktor penyebab munculnya penyakit tersebut.
2. Pemecahan masalah didekati dengan melakukan studi pustaka, studi literatur, observasi data di lapangan dan juga proses akuisisi data yang dibutuhkan untuk proses komputasi menggunakan algoritma *decision tree*. Akusisi dataset dilakukan dengan mengambil data hasil rekam medis pasien baik yang suspek batu ginjal maupun yang tidak.
3. Tahap pra proses merupakan tahapan pembersihan data yang tidak lengkap atau pemilihan atribut yang tidak dibutuhkan, sehingga dapat menghasilkan proses klasifikasi data yang lebih optimal.
4. Tahapan pembentukan model arsitektur *decision tree* merupakan tahapan awal desain aplikasi atau model dari aplikasi data mining yang dilakukan, model ini mengacu kepada kebutuhan untuk melakukan klasifikasi dan juga melihat tingkat akurasi dari model yang dibuat.
5. Tahap selanjutnya adalah penentuan pembagian data baik untuk data latih dan data uji dengan menggunakan 3(tiga) pendekatan data latih yaitu 70%, 80% dan 90%.
6. Setelah tahap pembuatan model arsitektur dan pembagian data telah selesai, maka tahapan berikutnya adalah proses komputasi dari model arsitektur yang sudah dibuat, pada tahap ini terbentuknya pohon keputusan.
7. Tahap ini merupakan tahapan akhir, dimana dapat terlihat proses hasil uji akurasi dari data yang telah dilatih berdasarkan model yang dibuat, untuk melihat hasil akurasi yang terbaik dari skenario pembagian data latih yang telah ditentukan.

3. LANDASAN TEORI

3.1 Data Mining

Data mining adalah suatu algoritma didalam menggali informasi berharga yang terpendam atau tersembunyi pada suatu koleksi data/ database yang sangat besar sehingga ditemukan suatu pola yang menarik yang sebelumnya tidak diketahui. Analisa data mining berjalan pada data yang cenderung terus membesar dan teknik terbaik yang digunakan kemudian berorientasi kepada data berukuran sangat besar untuk mendapatkan kesimpulan dan keputusan paling layak.

Data mining memiliki beberapa sebutan atau nama lain yaitu: *Knowledge discovery* (mining) in *databases* (KDD), ekstraksi pengetahuan (*knowledge extraction*), analisa data/pola, kecerdasan bisnis (*business intelligence*), dll (Elmande & Widodo, 2012).

3.2 Klasifikasi

Klasifikasi dapat digambarkan sebagai berikut. Data input, disebut juga training set, terdiri atas banyak contoh (*record*), yang masing-masing memiliki beberapa atribut. Selanjutnya, tiap contoh diberi sebuah label kelas khusus. Tujuannya untuk menganalisa data input dan mengembangkan deskripsi atau model

akurat untuk tiap class menggunakan fitur-fitur pada data.

Deskripsi kelas ini digunakan untuk mengklasifikasikan data pengujian lainnya dengan label kelas yang tidak diketahui. Deskripsi tersebut juga dapat digunakan untuk memahami tiap kelas dalam data. Aplikasi aplikasi klasifikasi antara lain berupa persetujuan kredit, target pemasaran, diagnosa medis, keefektifan perawatan, lokasi toko, dll (Nugraha et al., 2016).

Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011) :

1. Kelas Variabel dependen berupa kategori yang merepresentasikan "label" yang terdapat pada objek. Contohnya: risiko penyakit jantung, risiko kredit, dan jenis gempa.
2. Pemprediksi variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok atau tidak, besar tekanan darah, jumlah tabungan, jumlah aset, jumlah gaji.
3. Pelatihan dataset Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan pemprediksi.
4. Pelatihan dataset Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal, akurasi adalah fraksi prediksi model kita yang benar. Secara formal, akurasi memiliki definisi berikut:

$$Akurasi = \frac{Jumlah\ prediksi\ benar}{Jumlah\ total\ prediksi}$$

3.3 Algoritma C4.5

Algoritma C4.5 Algoritma C4.5 adalah salah satu metode untuk membuat pohon keputusan berdasarkan *training* data yang telah disediakan.

Beberapa pengembangan yang dilakukan pada C45 antara lain bisa mengatasi *missing value*, bisa mengatasi data kontinu, dan *prunning*.

Secara umum algoritma C4.5 dalam membangun pohon keputusan mengacu kepada tahapan sebagai berikut:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama. Untuk memilih atribut akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada (Elisa, 2017).

4. HASIL DAN PEMBAHASAN

4.1 Akuisisi Data

Tahapan awal yang dilakukan pada proses kegiatan penelitian ini berupa kegiatan akuisisi data. Akusisi data itu sendiri merupakan sebuah kegiatan untuk mengkoleksi data, data yang dimaksud disini adalah data set yang dibutuhkan dalam konteks sesuai dengan topik penelitian yang diangkat. Dataset yang diambil berasal dari sebuah rumah sakit di daerah Tangerang.

Data yang dibutuhkan adalah berupa kumpulan data rekam medis pasien baik pasien positif batu ginjal maupun pasien negatif batu ginjal. Data tersebut merupakan data rekam medis pasien batu ginjal yang terhitung dari tahun 2016-2019, yang

kemudian terkumpul sebanyak 283 data pasien terdiri atas pasien suspek batu ginjal dan bukan pasien suspek batu ginjal, dengan jumlah atribut sebanyak 18 (delapan belas) atribut.

4.2 Praproses Data

Tahapan praproses data merupakan tahapan dalam pembersihan data, karena dataset yang berasal dari data rekam medis pasien masih memiliki data yang tidak lengkap, terdapat *missing value* ataupun atribut dari data tersebut tidak relevan dengan kasus yang sedang dianalisis.

Pada tahap pra proses ini setelah dilakukan proses identifikasi terdapat sejumlah data yang memiliki *missing value* jumlahnya sebanyak 50 baris (*record*) data, sehingga data yang akan digunakan sebagai dataset menjadi sebesar 231 data dari keseluruhan data rekam medik yang telah diakuisisi.

Sedangkan atribut terpilih yang digunakan mengacu kepada hasil konsultasi dengan dokter penyakit dalam hanya sebanyak 12(dua belas) atribut yang merepresentasikan munculnya penyakit batu ginjal.

4.3 Pembagian Data

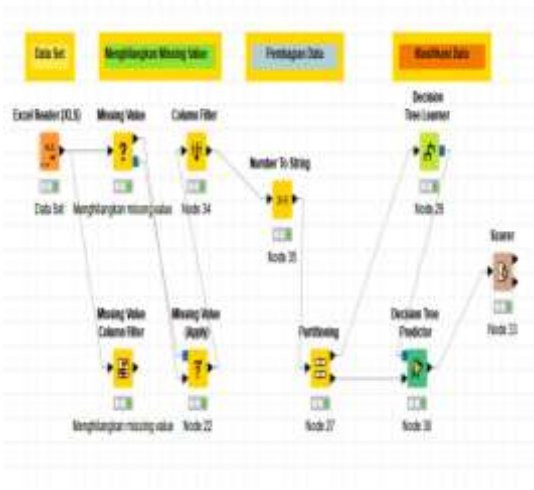
Dalam penelitian ini akan menggunakan tiga pendekatan pembagian jumlah data baik untuk data latih maupun data uji, adapun deskripsi pembagian datanya sebagai berikut :

Tabel 1. Skenario Pembagian Data

No	Pembagian Data		Jumlah Data	
	Latih	Uji	Latih	Uji
1	70	30	162	69
2	80	20	185	46
3	90	10	208	23

4.4 Perancangan Model Pohon Keputusan

Tahapan perancangan model pohon keputusan dibuat sebagai langkah awal dalam menerapkan komputasi algoritma *decision tree*, model yang dirancang harus dapat merepresentasikan kebutuhan sistem untuk mampu membentuk prinsip dasar desain model berbasis *decision tree*, adapaun dalam penelitian ini desain model arsitektur yang dibuat adalah sebagai berikut :



Gambar 2. Arsitektur Model Decision Tree

Arsitektur model *decision tree* yang dibuat pada gambar 2 menggunakan 10 node dari inisiasi awal penyiapan dataset sampai dengan proses komputasi algoritmanya, adapun rincian model arsitekturnya dapat dijelaskan sebagai berikut :

1. Penyiapan dataset merupakan tahapan awal dalam desain aristektur, pada tahap ini dataset yang telah diakuisisi kemudian dihubungkan untuk dapat diakses dalam pengolahan data di tahap berikutnya.
2. Tahap kedua merupakan tahapan dilakukannya praproses, untuk menghilangkan baris yang masih

terdapat data yang tidak lengkap, ataupun menghilangkan kolom yang memiliki atribut yang tidak merepresentasikan suspek batu ginjal.

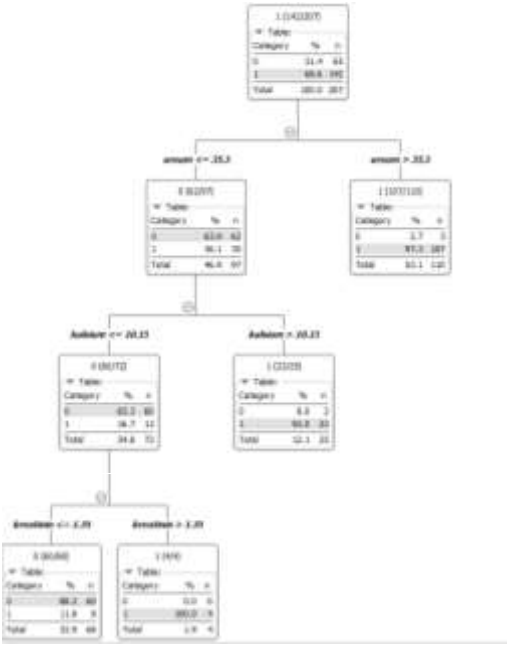
3. Tahap ketiga merupakan tahapan pembagian data untuk pelatihan dan pengujian, pada tahap ini pemisahan data latih akan dilakukan sebanyak 3(tiga) kali.
4. Tahap keempat merupakan tahapan akhir dalam melakukan klasifikasi dan prediksi, pada tahap ini akan terbentuk pohon keputusan dan juga aturan berdasarkan dataset yang telah dilatih. Pada tahap ini juga akan terlihat hasil uji akurasi dari data yang telah diuji.

4.5 Pembentukan Pohon Keputusan

Dalam mengubah data menjadi pohon terlebih dahulu data dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Dalam data sampel tentukan dulu *node* terpilih, yaitu dengan menghitung nilai informasi gain masing-masing atribut untuk menentukan node terpilih, gunakan nilai informasi gain yang paling besar (Azmi & Dahria, 2013).

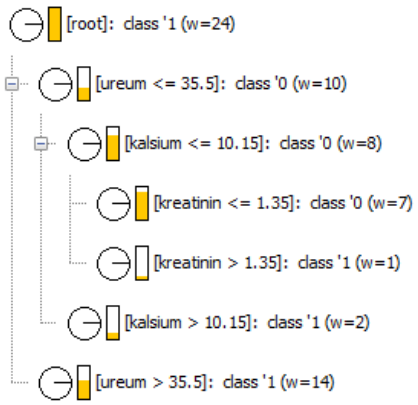
Pembentukan pohon keputusan pada penelitian ini menggunakan algoritma data mining C4.5 dimana untuk pemilihan atribut menggunakan gain ratio yang memiliki kemampuan untuk mengurangi bias dari atribut yang memiliki banyak cabang.

Berdasarkan arsitektur model pohon keputusan yang telah didesain, maka pohon keputusan yang terbentuk dengan menggunakan pemilihan atribut berdasarkan gain ratio adalah sebagai berikut ini :



Gambar 3. Bentuk Pohon Keputusan

Pada gambar 3, terlihat pohon keputusan yang terbentuk dengan menggunakan pemangkasan pada pohon keputusan agar pohon keputusan yang didapatkan lebih sederhana namun memiliki informasi yang lebih baik. Pohon keputusan yang terbentuk menghasilkan 4(empat) daun, dengan 4(empat) level *node* dari *root* yang terbentuk.



Gambar 4. Pohon Keputusan Sederhana

Pohon keputusan yang terbentuk pada gambar 3, dapat disederhanakan dalam bentuk pohon keputusan yang lebih sederhana seperti terlihat dalam gambar 4, hal ini sangat membantu sekali untuk mempermudah merepresentasikan banyaknya aturan yang terbentuk atau informasi berupa faktor-faktor penyebab munculnya penyakit batu ginjal.

Berikut ini merupakan aturan (*rule*) yang terbentuk berdasarkan pohon keputusan yang telah dibuat :

R1: Jika ureum > 35,5 Maka suspek batu ginjal

R2: Jika ureum ≤ 35,5 Dan kalsium > 10,15 Maka suspek batu ginjal

R3: Jika ureum ≤ 35,5 Dan kalsium ≤ 10,15 Dan kreatin > 1,35 Maka suspek batu ginjal

R4: Jika ureum ≤ 35,5 Dan Kalsium ≤ 10,15 Dan Kreatin ≤ 1,35 Maka negatif batu ginjal

4.6 Uji Akurasi

Uji akurasi dilakukan untuk melihat seberapa baik sistem mampu melakukan klasifikasikan data secara benar, adapun hasil klasifikasi data dengan menggunakan tiga pendekatan pembagian data pelatihan yaitu sebesar 70%, 80% serta pembagian data pelatihan sebesar 90%, memberikan informasi sebagai terdeskripsi pada tabel 2 berikut ini:

Tabel 2. Uji Akurasi

No	Pembagian Data Latih	Hasil Akurasi
1	70	95,71
2	80	78,72
3	90	87,5

Mengacu pada tabel 2, untuk dapat membantu mempermudah melihat data yang diperoleh, maka dapat dibuat sebuah grafik yang memberikan informasi yang lebih efektif mengenai perbandingan hasil uji akurasi sebagai berikut :

Gambar 5. Grafik Perbandingan Hasil Uji Akurasi

Berdasarkan informasi yang didapatkan dari tabel 2 dan gambar 5, tercermin bahwa akurasi yang terbaik berasal dari model yang dibuat pada skenario pertama, dengan pembagian jumlah data pelatihan sebanyak 70% dan data pengujian sebanyak 30%.

Tingkat akurasi dengan menggunakan skenario pertama didapatkan hasil nilai pengujian data latih tertinggi dengan nilai akurasi sebesar 95,71%, selisih sebesar 16,99% dari skenario kedua, selisih 8,21% dari skenario ketiga.

Berdasarkan hasil yang klasifikasi dan juga uji akurasi yang telah dilakukan, maka dapat disimpulkan bahwa model arsitektur yang dibangun dapat merepresentasikan kebutuhan dalam melakukan klasifikasi terhadap data set penyakit batu ginjal dengan baik, sehingga mampu menghasilkan bebrapa aturan yang dapat digunakan untuk dapat mengientifikasi faktor-faktor penyebab munculnya penyakit batu ginjal dan juga melihat tingkat akurasi dalam proses klasifikasi.

5. KESIMPULAN

Penggunaan algoritma *decision tree* untuk melakukan identifikasi dan klasifikasi terhadap dataset rekam medis pasien penyakit batu ginjal sangat diperlukan untuk dapat melihat faktor-faktor penyebab munculnya penyakit batu ginjal.

Dalam penelitian ini terlihat bentukan pohon keputusan yang memberikan informasi mengenai *rule* yang terbentuk merepresentasikan faktor penyebab terjadinya batu ginjal.

Penelitian ini menggunakan pendekatan tiga skenario pembagian data untuk melihat hasil uji akurasi data yang terbaik. Pada penelitian ini hasil uji akurasi yang terbaik ada di skenario pembagian data latih sebanyak 70% dengan tingkat akurasi 95,71%.

DAFTAR PUSTAKA

Azmi, Z., & Dahria, M. (2013). Decision Tree Berbasis Algoritma Untuk Pengambilan Keputusan. *Saintikom*, *12*, 157–164. http://demo.pohonkeputusan.com/files/Decision Tree Berbasis Algoritma Untuk Pengambilan Keputusan.pdf?i=1

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

Elmande, Y., & Widodo, P. (2012). 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), 10. http://demo.pohonkeputusan.com/files/Pemilihan Criteria Splitting Dalam Algoritma Iterative Dichotomiser 3 (Id3) Untuk Penentuan Kualitas Beras.pdf

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques*.

Nugraha, P. G. S. C., Aribawa, I. W., Priyana, I. P. O., & Indrawan, G. (2016). Penerapan Metode Decision Tree(Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1 Kintamani. *Seminar Nasional Vokasi Dan Teknologi (SEMNASVOKTEK)*, 35–44.

Wirdasari, D. (2013). *Analisa Teknik Penentuan Atribut Dalam Membuat Pohon Keputusan Pada Penambangan Data*. https://doi.org/10.1017/CBO9781107415324.004