Implementasi Algoritme Klasifikasi Naïve Bayes, Decision Tree J48, dan Multilayer Perceptron Menggunakan Weka

Rifqi Fauzi Rahmadzani
Magister Teknologi Informasi
Universitas Gajah Mada
Yogyakarta, Indonesia
Email: rifqifauzi@mail.ugm.ac.id

Abstrak - Kegiatan evaluasi dan pengambilan keputusan akan dapat dilakukan dengan baik jika suatu masalah memiliki informasi yang lengkap, cepat, tepat, dan akurat. Penelitian ini mengkaji untuk menguraikan kinerja terbaik dari beberapa algoritme klasifikasi dalam *data mining* yaitu *naïve bayes*, *decision tree J48*, dan *multilayer perceptron*. Beberapa aspek yang dilihat dalam penelitian adalah dari sisi keakuratan prediksi dan kecepatan/efisiensi. Adapun software yang digunakan untuk mengevaluasi beberapa algoritme klasifikasi tersebut adalah Weka versi 3.8. Hasil pengujian menunjukkan bahwa kinerja model *multilayer perceptron* memiliki akurasi terbaik sebesar 61.59% dengan menggunakan mode tes *precentage split*. Namun algoritme *naïve bayes* memiliki kecepatan yang terbaik untuk semua mode tes yang digunakan dalam penelitian ini.

Kata kunci—*Data mining, Naïve Bayes, Decision Tree J48, Multilayer Perceptron, Weka*

I. PENDAHULAN

Keberadaan dan pertumbuhan jumlah data yang begitu besar, akan menjadi suatu elemen penting dalam perkembangan masyarakat saat ini dan masa depan. Pemanfaatan data dalam sistem informasi untuk menunjang kegiatan pengambilan keputusan tidak cukup hanya mengandalkan data operasional saja, tetapi diperlukan suatu analisis data untuk menggali potensi-potensi informasi yang ada [1]. Untuk menangani masalah tersebut, diperlukan teknik dalam menggali suatu data yaitu *data mining*. *Data mining* berfokus pada penyelesaian dengan melakukan analisis pada data atau dapat didefinisikan sebagai proses menemukan pola makna dalam data dengan tujuan untuk memperoleh pada beberapa keuntungan [2]. Hal ini merupakan metode di mana data mentah ditransformasikan menjadi hubungan data pola antara variabel sebagai ekstraksi dan visualisasi sebagai pencapaiannya. *Data mining* mampu memproses data dalam jumlah besar, dengan demikian komputer dapat memperoleh pengetahuan yang diperlukan untuk melakukan tugas yang tidak dapat dilakukan oleh program.

Pada proses *data mining* terdapat beberapa metode pengolahan data, salah satunya adalah klasifikasi. Tujuan dari penelitian ini adalah untuk mengetahui perbandingan kinerja dari beberapa algoritme yang terdapat dalam metode klasifikasi sehingga dapat diketahui algoritme mana yang mempunyai keunggulan dalam hal keakuratan prediksi dan kecepatan/efisiensi.

Beberapa algoritme yang akan dibandingkan dalam penelitian ini adalah *naïve bayes*, *decision tree J48*, dan *multilayer perceptron* dengan masing-masing menggunakan mode tes *cross validation* dan *precentage split*. Hasil dari pembangunan model diukur berdasarkan tingkat akurasi klasifikasi data, *mean absolute* error, dan waktu pembangunan model. Penelitian ini menggunakan software WEKA versi 3.8 sebagai alat bantu untuk mengevaluasi kinerja empat algoritme tersebut.

II. ANALISA DAN KEBUTUHAN

Analisa Kebutuhan dilakukan untuk menganalisa kebutuhan apa saja yang diperlukan dalam mendefinisikan data yang akan diolah diantaranya.

1. Alat

Alat dan bahan yang digunakan untuk melakukan proses pengolahan data ditunjukkan seperti pada tabel 1 berikut.

TABEL 1. PERANGKAT YANG DIGUNAKAN

Perangkat keras	Perangkat lunak
a. Processor Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHZ	a. Sistem operasi Windows 10 Pro 64-bit
b. Hard Disk 1 TB, SSD 250 GB	b. Weka 3.8
c. RAM 8 GB	c. Library Naïve Bayes
d. GPU NVIDIA Geforce 930MX	d. Library J48 Trees
	e. Library Multilayer Percepton

2. Bahan

Percobaan ini menggunakan *dataset* dari *yeast data set* yang merupakan data untuk memprediksi situs lokalisasi seluler dari protein di dalam ragi [3]. Detail keterangan dari *dataset* tersebut ditunjukkan pada tabel 2.

Hasil yang tertera pada jendela *classifier output* setelah melalui proses pembangunan model akan dicatat dan dari pencatatan tersebut akan dbandingkan nilainya, sehingga dapat diketahui algoritme mana yang kinerjanya paling baik.

TABEL 2. INFORMASI DETAIL DATASET

Informasi	Detail
Dataset	Yeast
Tipe File	ARFF
Banyak Atribut	9
Banyak record	1484
Karakteristik Atribut	Real
Karakteristik Dataset	Multivariate
Missing Value	Tidak ada

Pada *dataset* Yeast memiliki 1448 baris data dan memiliki 9 atribut diantaranya:

a. SequenceName: Nomor akses untuk *database* SWISS-PROT.
b. Mcg: Metode McGeoch untuk pengenalan urutan sinyal.

c. Gvh: Metode von Heijne untuk pengenalan urutan sinyal.
d. Alm: Skor dari program prediksi wilayah span membran ALOM.
e. Mit: Skor analisis diskriminan dari kandungan asam amino wilayah N-terminal (20 residu lamanya) dari protein mitokondria dan non-mitokondria.
f. Erl: Adanya substrat "HDEL" (dianggap bertindak sebagai sinyal untuk retensi dalam lumen retikulum endoplasma). Atribut biner.
g. Pox: Sinyal penargetan peroxisomal dalam terminal-C.
h. Vac: Skor analisis diskriminan dari kandungan asam amino protein vakuolar dan ekstraseluler.
i. Nuc: Skor analisis diskriminan sinyal lokalisasi nuklir protein nuklir dan non-nuklir.

Satu kolom kelas ada di kolom paling terakhri dari kedua data set tersebut.

III. SPESIFIKASI DESAIN

1. Model Pelatihan

Pada penilian ini menggunakan tiga model untuk melakukan proses analisis yaitu *naïve bayes, decision tree,* dan *multilayer perceptron*.

a. *Naïve Bayes*

Naive bayes classifier (NBC) merupakan salah satu metode pada teknik klasifikasi dan termasuk dalam classifier statistik yang dapat memprediksi probabilitas keanggotaan class. NBC berprinsip pada teori bayes. NBC mengasumsikan bahwa nilai atribut pada sebuah class adalah independen terhadap nilai pada atribut yang lain [4].

$$P(Ci \mid X) = \frac{P(X \mid Ci)\, P(Ci)}{P(X)}$$

Class Ci adalah nilai terbesar, sedangkan P(X) adalah konstanta untuk semua class. P merupakan posterior probability.

b. *Decision Tree – J48*

Proses klasifikasi dapat dianggap sebagai penugasan suatu objek ke kelas tertentu, berdasarkan kesamaannya dengan contoh elemen sebelumnya, dalam kumpulan data tertentu [5].

J48 Decision Tree yang digunakan dalam penelitian ini, merupakan ekstensi dari ID3 dan implementasi algoritme C4.5. Algoritme ini awalnya didasarkan pada pendekatan *divide* and *conquer* dan bekerja secara optimal pada atribut nominal, bukan yang numerik.

Algoritme ID3 memperluas pendekatan *divide* dan *conquer*, menggunakan teori informasi untuk menyelesaikan masalah dengan memisahkan variabel yang memiliki ukuran kemurnian terbesar, yang dikenal sebagai informasi. Informasi dapat dianggap sebagai ukuran ketidakpastian dan didasarkan pada probabilitas sesuatu terjadi. Rumus matematika untuk menghitung informasi adalah $I(e) = -log_2(p_e)$, di mana p_e adalah probabilitas suatu peristiwa yang terjadi. Informasi rata-rata tertimbang dari semua nilai yang mungkin dari suatu atribut disebut nilai informasi atau entropi dan dievaluasi sebagai jumlah total probabilitas dari setiap peristiwa yang dikalikan dengan nilai informasinya. Entropi dihitung menggunakan rumus berikut ini, di mana H(X) dianggap sebagai ukuran ketidakmurnian dan ketidakpastian dalam variabel X.

$$H(X) = \sum P\,(x_i) I(x_i) = -\sum P\,(x_i) log_2(P(x_i))$$

Semakin merata distribusi X, semakin tinggi entropi didapat. Untuk mengukur jumlah informasi yang diperlukan dalam mengevaluasi hasil dari variabel X, diberikan input yang diketahui, entropi kondisional perlu dihitung. Rumus entropi bersyarat $H(Outcome \mid Known\ Input)$ pada dasarnya digunakan untuk menghitung entropi setelah melakukan pemisahan variabel. Selanjutnya, membandingkan entropi yang dihitung sebelum dan sesudah pemisahan, hal ini digunakan untuk mengukur Penguatan Informasi Input. Penguatan informasi didasarkan pada penurunan entropi setelah dataset dibagi pada atribut. Membangun *decision tree* merupakan semua tentang atribut temuan yang mengembalikan perolehan informasi tertinggi [6].

$$Information\ Gain = H(Outcome) - H(Outcome \mid Known\ Input)$$

Oleh karena itu, algoritme ID3 mengambil simpul akar dengan menghitung perolehan informasi dari kelas output untuk setiap variabel input, memilih salah satu yang menghilangkan ketidakpastian paling banyak. Selain itu, cabang dibuat untuk setiap nilai yang dapat diambil atribut yang dipilih, dan perhitungan perolehan informasi yang sama secara berulang-ulang untuk setiap node internal sampai tidak diperlukan lagi percabangan. Ketika entropi cabang mencapai 0, simpul daun telah tercapai. Sengan demikian tidak diperlukan lagi pemisahan. Algoritme juga berhenti ketika semua data telah diklasifikasikan.

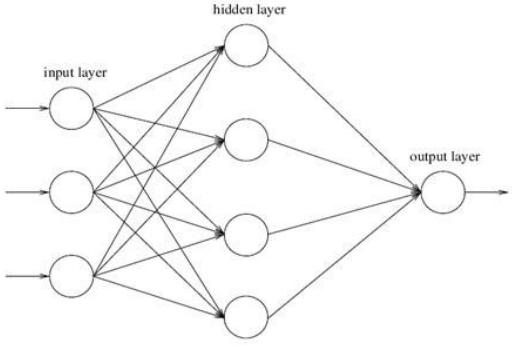
c. *Multilayer Perceptron – Neural Network*

Jaringan Saraf Tiruan (JST) disusun dengan struktur dan fungsi otak manusia sebagai model untuk ditiru. Pada sebuah jaringan saraf tiruan terdapat sejumlah neuron. Satu neuron bisa terhubung ke banyak neuron lain, dan setiap koneksi (link) tersebut mempunyai bobot (weight). Tabel 3 merupakan analogi bagian-bagian dari jaringan saraf tiruan terhadap jaringan saraf biologis. Pembelajaran merupakan karakteristik dasar dari jaringan saraf biologis. Jaringan saraf tiruan melakukan proses pembelajaran melalui penyesuaian bobot pada koneksi antar neuronnya.

TABEL 3. ANALOGI JARINGAN SARAF BIOLOGIS DAN JST

Jaringan Saraf biologis	Jaringan saraf tiruan
Soma	Neuron
Dendrite	Input (Masukan)
Axon	Output (Luaran)
Synapse	Weight (Bobot)

Multilayer Perceptron adalah topologi paling umum dari JST, di mana perceptronperceptron terhubung membentuk beberapa lapisan (layer). Sebuah MLP mempunyai lapisan masukan (input layer), minimal satu lapisan tersembunyi (hidden layer), dan lapisan luaran (output layer). Arsitektur JST ditunjukkan pada gambar 3.1 [7].



Gambar 3.1. Arsitektur JST

Jenis jaringan saraf ini merupakan sebuah *finite acyclic graph* yang berarti bahwa semua neuron dapat diatur dalam lapisan. Setiap *node* adalah neuron dengan aktivasi logistik. *Node* yang memproses semua koneksi lain disebut neuron input. Sebaliknya, *node* yang tidak ada sumber koneksi apa pun disebut *output* neuron. *Neuron output* juga dapat terdiri lebih dari satu untuk MLP yang diberikan. Namun, dalam penelitian ini hanya ada satu *output* saja. Sisa dari *node*, yang bukan

bagian dari lapisan *input*, atau lapisan *output*, disebut neuron tersembunyi.

2. Parameter

Parameter yang digunakan untuk membandingkan kinerja dari beberapa algoritme klasifikasi adalah:

a. *Test Mode*: Mendefinisikan mode tes yang digunakan adalah cross-validation test dan percentage split test mode untuk teknik evaluasi.
b. *Time to build model*: merupakan istilah untuk menerangkan berapa waktu yang dibutuhkan untuk membangun model klasifikasi untuk masing-masing algoritme
c. *Correctly classified instances*: berapa banyak baris data yang terklasifikasikan dengan benar.
d. *Incorrectly classified instances*: berapa banyak baris data yang terklasifikasikan tidak benar.

IV. IMPLEMENTASI

Implementasi dilakukan dengan cara melakukan klasifikasi dari dataset yang digunakan kemudian dilakukan evaluasi terhadap hasil dari beberapa model. Hasil evaluasi dari kinerja algoritme *naïve bayes, decision tree J48*, dan *multilayer perceptron* dapat dilihat pada tabel 4-6. Informasi yang didapat dari tabel 4-6 terdiri dari mode tes yang digunakan untuk masing-masing dataset yang terdiri dari mode tes *cross validation* dan *percentage split*. Sementara untuk konfigurasi parameter lainnya diterapkan secara default. Informasi ukuran akurasi juga bisa didapatkan dari tabel 4-6 pada kolom *correctly classified instances* yang berarti nilai akurasi klasifikasi data benar dan *incorrectly classified instances* yang berarti nilai akurasi data salah, sementara untuk *mean absolute error* merepresentasikan rata-rata kesalahan (*error*) absolut antara hasil prediksi dengan nilai sebenarnya. Semakin kecil nilai dari *mean absolute error* maka akan membuktikan bahwa algoritme tersebut semakin baik untuk digunakan [8].

TABEL 4. HASIL EVALUASI ALGORITME BERDASARKAN TINGKAT KEBENARAN KLASIFIKASI

Algoritme	Mode Tes	*Correctly Classified Instances*	
		Jumlah	Persentase
Naïve Bayes	*Cross Validation*	855	57.61 %
	Precentage Split	304	60.20 %
Tree J48	*Cross Validation*	830	55.93 %
	Precentage Split	297	58.81 %
Multilayer Perceptron	*Cross Validation*	882	59.44 %
	Precentage Split	311	61.59 %

TABEL 5. HASIL EVALUASI ALGORITME BERDASARKAN TINGKAT KESALAHAN KLASIFIKASI

Algoritme	Mode Tes	*Incorrectly Classified Instances*	
		Jumlah	Persentase
Naïve Bayes	*Cross Validation*	629	42.39 %
	Precentage Split	201	39.80 %
Tree J48	*Cross Validation*	654	44.07 %
	Precentage Split	208	41.18 %
Multilayer Perceptron	*Cross Validation*	602	40.56 %
	Precentage Split	194	38.41 %

TABEL 6. HASIL MEAN ABSOLUTE ERROR MASING-MASING ALGORITME

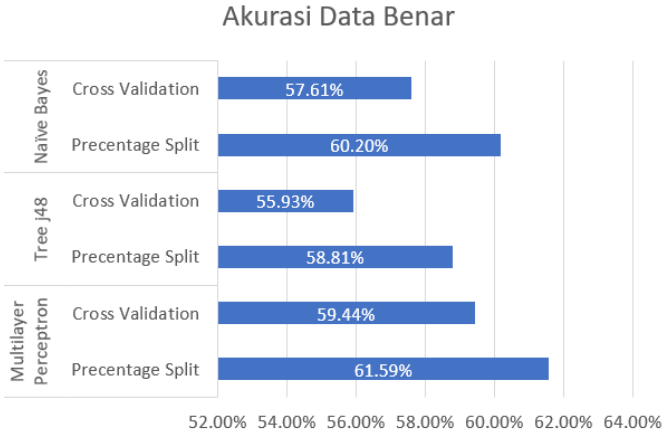
Algoritme	Mode Tes	Mean Absolute Error
Naïve Bayes	*Cross Validation*	0.1045
	Precentage Split	0.1035
Tree J48	*Cross Validation*	0.1017
	Precentage Split	0.0994
Multilayer Perceptron	*Cross Validation*	0.1018
	Precentage Split	0.1013

Mode tes *cross validation* menggunakan 10 *folds cross validation*, sementara untuk mode tes *percentage split* dalam penelitian ini menggunakan nilai pembagian jumlah data training dan tes sesuai dengan nilai *default* yang disediakan yaitu sebesar 34% untuk data *training* dan 66% untuk data *testing*. Nilai pada kolom persentase didapatkan dari hasil nilai pada kolom angka dibagi dengan total baris data pada *dataset* kemudian dikalikan dengan 100.

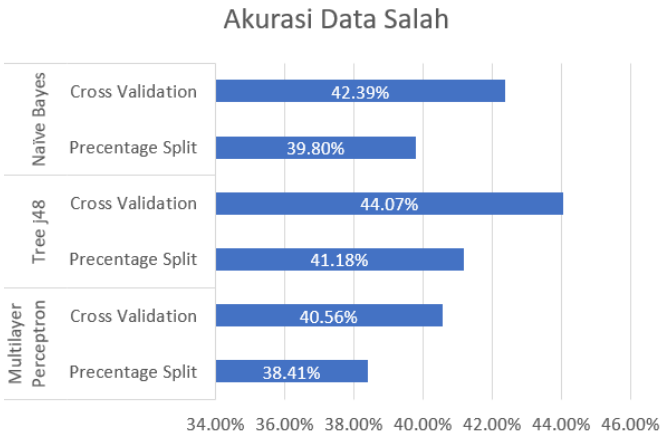
V. EVALUASI

Jika dilihat secara keseluruhan pada tabel 4-6, nilai akurasi tertinggi diraih pada algoritme *multilayer perceptron* dengan mode tes *precentage split* yang mencapai 61,59% pada kolom *correctly classified instances*. Dalam hal ini, implementasi algoritme *multilayer perceptron* dengan menggunakan mode tes *precentage split* yaitu terdiri dari 311 *instance* yang terklasifikasi benar dari total 505 jumlah data keseluruhan dengan nilai *mean absolute error* sebesar 0.1013.

Begitu sebaliknya algoritme *multilayer perceptron* yang menggunakan mode tes *percentage split* memiliki nilai terendah untuk *incorrectly classified instance* yaitu 38,41% di mana klasifikasi data salah hanya sebesar 194 *instances* dari total keseluruhan data sebanyak 505 *instances*.



Gambar 5.1. Hasil perbandingan nilai akurasi klasifikasi data benar



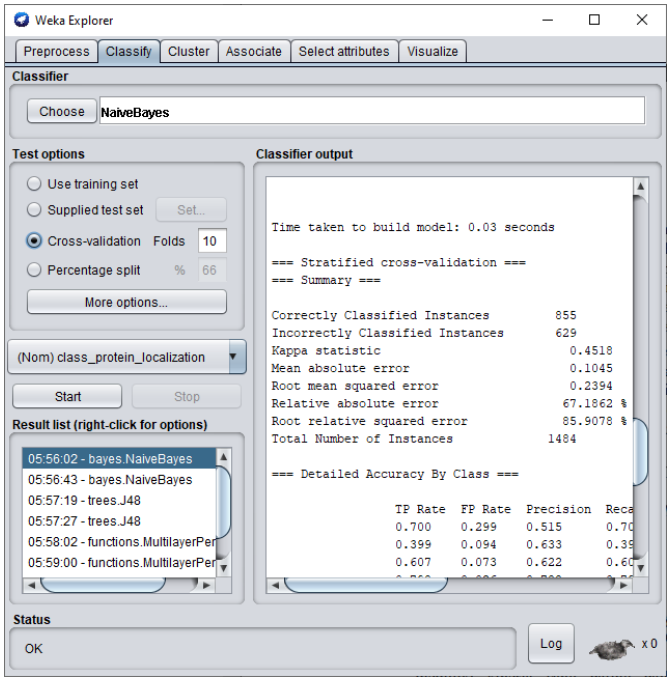
Gambar 5.2. Hasil perbandingan nilai akurasi klasifikasi data salah

Dari dua grafik yang terdapat pada gambar 5.1 dan gambar 5.2 memperlihatkan perbandingan nilai akurasi dari beberapa algoritme dan mode tes yang digunakan. Pada gambar 5.1 memperlihatkan bahwa mode tes *percentage split* di keseluruhan algoritme yang digunakan meraih hasil akurasi benar lebih tinggi dibanding menggunakan mode tes *cross validation*. Begitupun pada gambar 5.2 menunjukkan bahwa di keseluruhan algoritme dari hasil akurasi data salah terlihat

pada mode tes *percentage split* memiliki persentase yang lebih rendah dibandingkan menggunakan mode tes *cross validation*. Hal ini menunjukkan bahwa mode tes *percentage split* lebih efektif dibanding dengan mode tes *cross validation* dari implementasi beberapa algoritme khususnya untuk menangani klasifikasi terhadap *dataset* yeast yang digunakan ini.

TABEL 7. WAKTU YANG DIBUTUHKAN UNTUK MEMBANGUN MODEL

Algoritme	Mode Tes	Waktu (Detik)
Naïve Bayes	*Cross Validation*	0.03
	Precentage Split	0.03
Tree J48	*Cross Validation*	0.13
	Precentage Split	0.04
Multilayer Perceptron	*Cross Validation*	3.25
	Precentage Split	3.25



Gambar 5.3. Hasil output waktu yang digunakan untuk membangun model

Dari data tabel 7 dapat diketahui informasi mengenai waktu yang dibutuhkan untuk membangun model pada beberapa algoritme klasifikasi. Satuan waktu yang digunakan adalah detik. Mode tes yang digunakan juga terdiri dari dua yaitu *cross validation* dan *precentage split*.

Gambar 5.3 menunjukkan salah satu hasil evaluasi untuk algoritme *naïve bayes* yang menggunakan mode tes *cross validation*. Output dari jendela *classifier output* memberikan catatan waktu selama 0.03 detik untuk kriteria waktu yang dibutuhkan dalam membangun model.

Algoritme *naïve bayes* memiliki waktu *time to build* model yang sangat cepat untuk kedua mode tes yang digunakan yaitu selama 0.03 detik. Sebaliknya algoritme *multilayer perceptron* memakan waktu yang lebih lama dari kedua algoritme lainnya yaitu selama 3.25 detik.

VI. KESIMPULAN DAN SARAN

Berdasarkan data hasil evaluasi kinerja dari beberapa algoritme klasifikasi yaitu: *naïve bayes*, *decision tree J48*, dan *multilayer perceptron* dapat disimpulkan bahwa *multilayer perceptron* memiliki kinerja yang paling baik dalam hal akurasi. Hal tersebut dapat dibuktikan dari nilai *correctly classified instance*s yang mencapai angka presentase tertinggi sebesar 61,59% pada mode tes *precentage split*. Begitu juga pada mode tes *cross validation*, *multilayer perceptron* mencapai prosentase tertinggi sebesar 59,44%. Untuk kategori Time to build model, algoritme *naïve bayes* memilik waktu tercepat selama 0,03 detik untuk kedua jenis mode tes baik cross-validation maupun percentage-split.

Namun dalam penelitian ini masih memiliki akurasi yang cukup rendah terhadap dataset yang digunakan. Untuk penelitian kedepan dapat menggunakan konfigurasi parameter lebih lanjut dengan mengatur nilai *batch size*, *numfolds*, atau jumlah *hidden layer* dalam suatu algoritme, agar model mencapai nilai akurasi yang lebih baik.

REFERENSI

[1] Andri, Y.N. Kunang, dan S. Murniati, "Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang," Seminar Nasional Informatika 2013, ISSN: 1979-2328.

[2] H. Witten, E. Frank, M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," San Francisco, Elsevier, 2005.

[3] Yeast Data Set, 9 Desember 2019, https://archive.ics.uci.edu/ml/datasets/Yeast

[4] M. K. and J. P. Jiawei Han, *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers, 2012.

[5] P. N. Tang, M. Steinbach, V. Kumar, "Introduction to Data Mining (First Edition)," Boston MA, USA, Addison-Wesley Longman Publishing Co. Inc., 2005.

[6] Decision Tree - Classification, 11 Desember 2019, https://www.saedsayad.com/decision_tree.htm

[7] M. Negnevitsky, "Artificial Intelligence: A Guide to Intelligent Systems (3rd Edition)," Pearson Education Canada, 2011.

[8] P. Mittal and N. S. Gill, "A Comparative Analysis of Classification Techniques on Medical Data Sets," 2014.

Video Presentasi tersedia di
https://www.youtube.com/watch?v=_x4FPiQf6CM