Komparasi Algoritma *Decision Tree*, *K-Nearest Neighbors* (KNN) dan *Naïve Bayes* pada Pengobatan Penyakit Kutil Menggunakan *Cryotherapy*

Bayu Nur Angga
Magister Teknik Informatika, Universitas Pamulang
e-mail: bayunurangga21@gmail.com

***Abstrak*—Pengobatan penyakit kutil menggunakan *Cryotherapy* merupakan salah satu jenis pegobatan penyakit kutil yang direkomendasikan oleh beberapa pakar kesehatan. Metode yang digunakan dengan menggunakan nitrogen cair untuk pembekuan pada penyakit kutil. Dalam penelitian ini dilakukan komparasi pengujian model dengan menggunakan Decision Tree, K-Nearest Neighbors dan Naïve Bayes untuk prediksi pengobatan penyakit kutil. Dalam proses pengujiannya, peneliti menggunakan aplikasi orange untuk mengolah data dan melakukan pengujian. Hasil pengujian yang telah dilakukan menunjukkan pengujian menggunakan model K-Nearest Neighbors (K-NN) didapat nilai akurasi terbaik adalah 90,00% dengan nilai AUC sebesar 0,902, hasil pengujian menggunakan model Naïve Bayes didapat nilai akurasi lebih kecil dibandingkan dengan model K-NN yaitu 88,33% dengan nilai AUC sebesar 0,950, sedangkan model Decision Tree paling rendah didapat nilai akurasi 86,67% dengan nilai AUC 0,890. Berdasarkan pengujian yang sudah dilakukan dapat disimpulkan bahwa model K-Nearest Neigbor memiliki tingkat akurasi lebih baik dibandingkan dengan model Decision Tree dan Naïve Bayes dalam prediksi pengobatan penyakit kutil menggunakan *Cryotherapy*.**

***Kata Kunci—Decision Tree, K-Nearest Neighbors, Naïve Bayes*, Penyakit Kutil, *Cryotherapy*.**

I. PENDAHULUAN

Penyakit kutil adalah masalah kesehatan kulit yang biasanya ditandai dengan munculnya benjolan kecil pada permukaan kulit, penyakit ini disebabkan oleh virus yaitu human papilloma virus (HPV). Penyebaran virus penyebab kutil bisa terjadi dengan mudah, contohnya hanya bersentuhan langsung dengan seseorang penderita kutil, tetapi tidak semua orang yang bersentuhan dengan virus hpv akan menimbulkan penyakit kutil. kekebalan tubuh (imunitas) masing- masing orang berbeda dan sangat berpengaruh dalam penularan penyakit ini, seseorang yang memiliki imunitas yang bermasalah akan rentan untuk terserang penyakit kutil [1].

Virus HPS masuk ke kulit melalui mikroabrasi dan menginfeksi sel basal. Siklus hidup HPV berkaitan dengan diferensi keratinosit baik pada fase produktif atau non produktif, fase non produktif meliputi pembentukan genome viral dalam jumlah sedikit sesuai dengan tingkat pembelahan sel basal. Fase produktif mengikuti proses diferensiasi keratinosit dan virus mengalami replikasi dalam jumlah besar, mengepresikan late gen serta menghasilkan viral progeny. Terdapat berbagai cara untuk mengobati penyakit kutil (virus hpv) yaitu dengan melakukan vaksin, *Cryotherapy*, dan imunoteraphy. Salah satu contoh untuk mengobati penyakit kutil dengan metode *Cryotherapy* adalah teknik pengobatan terapi dengan berendam di dalam es atau air yang dingin selama kurang lebih 30 menit [2].

Dengan menggunakan es atau *Cold Bath* adalah beberapa cara yang dapat dipakai saat terapi, pada terapi ini menggunakan modalitas yang dapat menyerap jaringan suhu yang mengakibatkan penurunan penurunan suhu ringan yang melewati mekanisme konduksi. Efek yang terjadi pada pendinginan ini dapat terjadi tergantung jenis aplikasi terapi dingin yang dipakai, kalori yang diserap pada area local cedera sehingga terjadinya penurunan suhu adalah inti dari terapi dingin ini. Cryotherapy menyebabkan kerusakan jaringan melalui pembentukan Kristal es ekstra dan intraseluler, distrupsi membrane sel dan perubahan sirkulasi bagian kulit. Cryotherapy tidak menyebabkan keterlibatan sistemik, sehingga sesuai untuk wanita hamil. Terapi jenis ini sering tidak sesuai untuk lesi luas dan efek local yang sering terjadi adalah nekrosis, nyeri, pembentukan bula, edem dan hipopigmentasi. Pemakaian cryotherapy pada anak tidak dianjurkan. *Cryotherapy* tunggal mempunyai clearance rate sebesar 54-88% dengan tingkat rekuensi 21-40% [3].

Untuk menentukan klasifikasi pengobatan penyakit kutil dengan menggunakan *cryotherapy* terdapat banyak metode yang dapat digunakan seperti Naïve bayes, K- Nearest Neighbors, Decision Tree, K-Means, dan lain sebagainya, Namun untuk memilih metode yang paling cocok, dapat dilakukan komparasi antara beberapa metode.

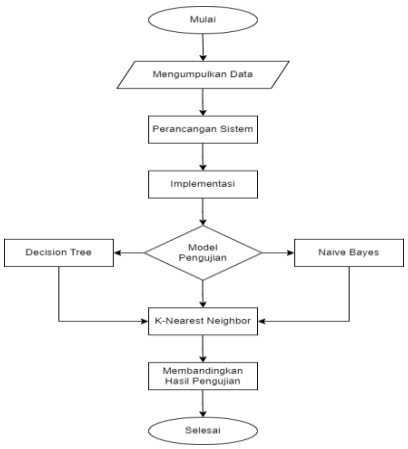
Machine learning dan algoritma data mining digunakan untuk menganalisa dataset dalam jumlah yang banyak, menemukan dan mengekstrak pengetahuan dari dataset tersebut. Algoritma data mining dapat menganalisa kumpulan fakta dan informasi (data) untuk menentukan pola yang tidak diketahui dalam basis data (*database*) dalam jumlah banyak dari beberapa instansi seperti manufacturing, perbankan, asuransi, marketing, dan kesehatan. Umumnya algoritma data mining diterapkan untuk tujuan

mengurangi biaya, meningkatkan kualitas penelitian, dan meningkatkan jumlah *income* bisnis [4]. Decision Tree, K-Nearest Neighbors dan Naïve bayes merupakan tiga algoritma data mining yang digunakan untuk melakukan klasifikasi. Penelitian ini bertujuan untuk melakukan pengolahan dataset tentang penyembuhan penyakit kutil dengan metode *cryotherapy* dan menggunakan algoritma Decision Tree, K-Nearest Neighbors dan Naïve Bayes dalam memprediksi keberhasilan pengobatan penyakit kutil dengan menggunakan metode cryoteraphy.

Untuk itu penelitian ini akan melakukan komparasi antara metode algoritma Decision Tree, K-Nearest Neighbors dan Naïve Bayes untuk mengetahui metode mana yang lebih baik dalam membantu melakukan klasifikasi terhadap keberhasilan pengobatan penyakit kutil dengan menggunakan *cryotherapy*. Berdasarkan uraian deskripsi pendahuluan, maka judul yang digunakan untuk jurnal ini adalah "Komparasi Algoritma *Decision Tree*, *K-Nearest Neighbors* (KNN*)* dan *Naïve Bayes* pada Pengobatan Penyakit Kutil Menggunakan *Cryotherapy*.

II. METODE PENELITIAN

Metode penelitian yang digunakan komparasi antara tiga perbandingan model atau metode klasifikasi dengan menggunakan metode algoritma data mining yaitu metode Decision Tree, K-Nearest Neighbors (K-NN) dan Naïve Bayes, tiga jenis metode ini adalah bagian dari model atau metode supervised learning. Decision Tree adalah metode yang termasuk dalam kategori metode yang relatif sederhana dan mudah di interpretasi [5]. K-Nearest Neighbors (K-NN) adalah metode yang dikenal paling sederhana [6]. Naïve Bayes adalah sebuah metode yang dapat menampilkan menggunakan label kelas terkait walaupun dengan data *training* yang sedikit [7]. Hasil dari setiap metode kemudian dicocokan dengan *k-fold cross validation*.



Gambar 1.
Metode Penelitian

A. *Algoritma Decision Tree*

Decision Tree adalah metode yang termasuk dalam kategori metode yang relatif sederhana dan mudah di interpretasi, dimana sebuah struktur data yang terdiri dari simpul (node) dan rusuk (edge). Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar (root/node), simpul percabangan/internal (branch/internal node) dan simpul daun (leaf node) [5]. Entropy adalah ukuran tingkat keacakan atau ketidakpastian dalam kumpulan data. Dalam hal klasifikasi, Ini mengukur keacakan berdasarkan distribusi label kelas dalam kumpulan data, juga sering dikatakan ukuran kemurnian (purity) persamaan 1.

$$E(S) = \sum_{i=1}^{c} - Pi \log_2 Pi \dots (1)$$

Keterangan:

S : himpunan kasus
c : jumlah partisi S
pi : proporsi dari Si terhadap S

B. *Algoritma K-Nearest Neighbors*

Salah satu algoritma paling popular dalam *machine learning* adalah algoritma K-nearest Neighbors (k-nn) dengan proses mudah dan sederhana [8]. Berdasarkan nilai dari variabel target yang terasosiasi dengan nilai variabel prediktor k-NN salah satu dari algoritma *supervised learning* dengan proses belajar. Dan dalam algoritma k-nn data harus memiliki label ketika data baru diberikan selanjutnya dibandingkan dengan data yang telah ada, diambil dari data yang mirip melihat dari data tersebut. Berikut langkah-langkah algoritma k-NN:

1) Menentukan parameter K
2) Menghitung jarak diantara data uji dengan data latih, jika data berbentuk numerik harus menggunakan *Euclidean distance*
3) Kemudian jarak diurutkan secara descending
4) Menentukan jarak terdekat pada parameter K
5) Jumlah kelas terbanyak di klasifikasikan

Algoritma K-Nearest Neighbors (KNN) adalah mencari range (jarak) yang paling dekat dengan k tetangga (*Neighbors*) terdekat dalam data training dengan data yang akan diolah [9]. Teknik mengelompokkan data baru dengan cara menghitung jarak data baru ke beberapa data/tetangga (*Neighbors*) paling dekat. Algoritma K-Nearest Neighbors (KNN) merupakan *instead-based learning*. Dimana data traning disimpan sehingga klasifikasi untuk tumpukan dataset baru (*record*) yang belum diklasifikasi dapat ditemukan dengan membandingkan kemiripan yang paling banyak dalam data traning [10][11]**.** Permasalahan yang diketahui dimiliki dalam metode K-Nearest Neighbors (KNN) yaitu menemukan nilai terdekat K dari tetangga pada query dataset yang digunakan [12][13]. Untuk menghitung *distance* (jarak) dalam K-Nearest Neighbors (KNN) digunakan fungsi *Euclidien Distance* sebagaimana persamaan 2.

$$euc = \sqrt{\sum_{i}^{n} = ((X_2)i - (X_1)i)^2} \ldots (2)$$

Keterangan :

X_2 = Data latih

X_1 = Data uji

i = variable data

n = dimensi data

C. Algoritma Naïve Bayes

Naïve bayes merupakan salah satu metode yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi nilai atau data dari dataset yang diberikan [14]. *Naïve bayes* merupakan langkah- langkah dalam menyelesaikan masalah, klasifikasi yang sangat tepat, akurat, simple, dan efisiensi (proses penalaran dilakukan memanfaatkan input yang ada dengan cara yang relatif cepat). Langkah-langkah naïve bayes bertujuan untuk melakukan klasifikasi data terhadap kelas tertentu dan dapat menangani data baik yang bersifat diskrit maupun *continue* adalah kelebihan lain dari naïve bayes [15][16], seperti pada persamaan 3.

$$P(x1,,,,,,,,xk|C) = P(x1|C)x \ \ldots \ \ldots \ x\ P\ (xk|C) \ldots (3)$$

Adalah nilai probabilitas yang diberikan pada persamaan 2. Ketika diberikan k atribut yang saling bebas adalah dalam proses mencari kelas terbaik. Jika atribut ke-I bersifat diskrit atau kategori, maka$P\ (x_i|C)$ dapat diestimasi sebagai frekuensi relative sampel yang memiliki nilai $x_i$sebagai atribut ke-i dalam kelas C. maka $P\ (x_i|C)$ dapat dicari menggunakan *densitas gauss* (persamaan 4).

$$f(x) = \frac{1}{\sqrt{2}\pi\sigma} e - \frac{(x-\mu)^2}{2\sigma^2} \ldots (4)$$

Keterangan :

σ^2 = standar deviasi

μ = mean

D. Dataset

Pada Penelitian ini Dataset yang digunakan adalah dataset yang bersifat publik dan mengarah pada penelitian Khozeimeh [1] pengobatan kutil dengan metode *cryotherapy*, kemudian data-data di uji pada metode klasifikasi yang diterapkan pada aplikasi orange. Sehingga dilanjutkan lagi dengan validasi data.

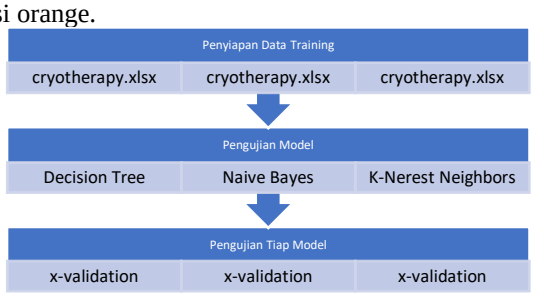
E. Akurasi

Pengujian akurasi merupakan suatu ukuran seberapa dekat (valid) hasil pengukuran terhadap angka sebenarnya. Akurasi dapat diperoleh dari persentase kebenaran, yaitu perbandingan antara jumlah data benar dengan keseluruhan dataset seperti pada persamaan 5.

$$akurasi = \frac{total\ data\ benar}{total\ data} \times 100\% \ldots (5)$$

III. HASIL DAN PEMBAHASAN

Pada penelitian pengobatan kutil dengan metode *cryotherapy* dibagi menjadi tiga model pengujian adalah pengujian dengan model decision tree, k-nearest Neighbors (k-nn) dan model naïve bayes, kemudian hasil di klasifikasi dari masing masing data yang di ambil dan dilakukan dengan proses validasi data. Kemudian hitung standar deviasi dari tiap-tiap metode. sistem ini menggunakan machine learning aplikasi orange.



Gambar 2.
Kerangka kerja implementasi sistem menggunakan *orange*

A. Pengujian Nilai K dengan K-Nearest Neighbors

Pengujian dengan K-Nearest Neighbors (K-NN) dilakukan dengan cara menginisialisasi nilai K, pada pengujian ini dilakukan sebanyak 15 kali mengubah isi nilai K dalam setiap skenario pengujian K dilakukan sejumlah 8 record pengujian data traning. Hasil dari setiap pengujian dengan nilai k terbaik akan digunakan dalam pengujian gabungan (komparasi) metode *Decision Tree*, *Naïve Bayes* dan *K-Nearest Neighbors* (K-NN). Nilai K yang diimplementasikan pada pengujian komparasi nanti hanya menggunakan angka terkecil dengan akurasi tertinggi dikarenakan untuk mempermudah perhitungan komparasi selanjutnya. Data traning yang akan diolah pada pengujian sebanyak 60 data dan data testing sebanyak 30 data. Hasil pengujian berdasarkan nilai atribut K adalah sebagai berikut.

K1 : 90,00%
K3 : 68,33%
K5 : 70,00%
K7 : 65,00%
K9 : 63,00%
K11 : 65,00%
K13 : 51,67%
K15 : 63,33%

Dikarenakan K1 telah mendapatkan tingkat akurasi yang sangat tinggi dengan nilai K lain yaitu 90,00%, maka peneliti mengambil K1 sebagai bahan penelitian dengan 60 data *training* dan 30 data *testing* dari 90 jumlah *record dataset*. Pada pengujian berdasarkan nilai atribut K data yang digunakan sebanyak 90 record dataset yang terbagi menjadi 60 data traning dan 30 data testing. Hasil pengujian terendah terjadi ketika nilai K yang bernilai 9 yaitu 63,00%. Hasil pengujian yang tertinggi dan terbaik pada nilai atribut K yang bernilai 1 yaitu 90,00%. Dapat ditarik kesimpulan bahwa akurasi K- Nearest Neighbors akan dipengaruhi terhadap jumlah nilai atribut K. Semakin banyak nilai atribut K, maka semakin rendah tingkat akurasinya, hal ini dikarenakan oleh atribut yang yang digunakan memiliki kesamaan dalam jumlah banyak sehingga semakin banyak tetangga atau nilai K yang diambil semakin banyak dari data dari kelas lain ikut dijadikan pertimbangan keputusan. Pada pengujian yang dilakukan oleh metode K-Nearest Neighbors (K-NN) akurasi yang tertinggi diperoleh pada saat nilai atribut K bernilai 1 yang nantinya akan digunakan dalam pengujian komparasi metode Decision Tree, K-Nearest Neighbors dan Naïve Bayes.

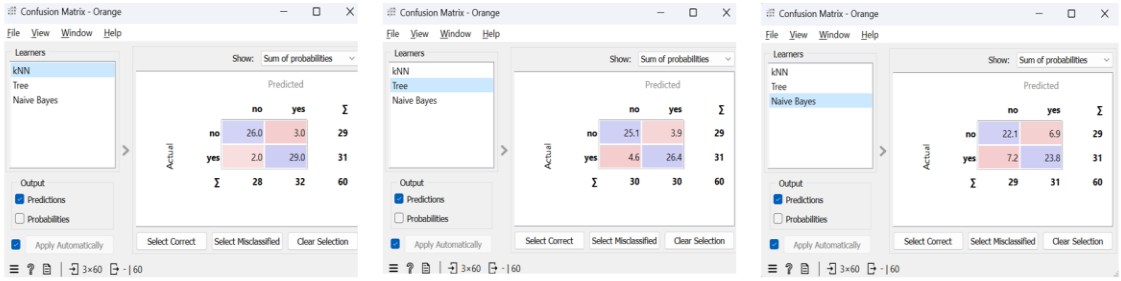
B. Pengujian Performance Tiap Model

Di tahap ini peneliti mulai membandingkan tingkat akurasi dari tiap model yaitu Decision Tree, K-Nearst Neighbors dan Naïve Bayes dengan dataset yang sama yaitu 60 *training* set dan 30 *testing* set, disini memperoleh untuk model Decision Tree dengan nilai performance 86,67%, Naïve Bayes dengan nilai performance 88,83%, dan K-NN yang mendapatkan hasil performance 90,00% dengan jumlah dataset yang sama. Tiga model klasifikasi yang memiliki ciri khas yang berbeda masing-masing model sehingga dalam proses klasifikasi memiliki langkah-langkah yang berbeda pula atau sering disebut dengan algoritma yang akan diterapkan dalam melakukan klasifikasi data pengobatan penyakit kutil dengan metode *cryotherapy*. Berikut hasil Test & Score nya (Gambar 3).

Model	Train	Test	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
kNN	0.083	0.033	0.902	0.900	0.900	0.902	0.900	0.801	0.905	3.454
Naive Bayes	0.122	0.011	0.951	0.883	0.884	0.888	0.883	0.771	0.891	0.292
Tree	0.084	0.002	0.891	0.867	0.867	0.869	0.867	0.734	0.871	2.496

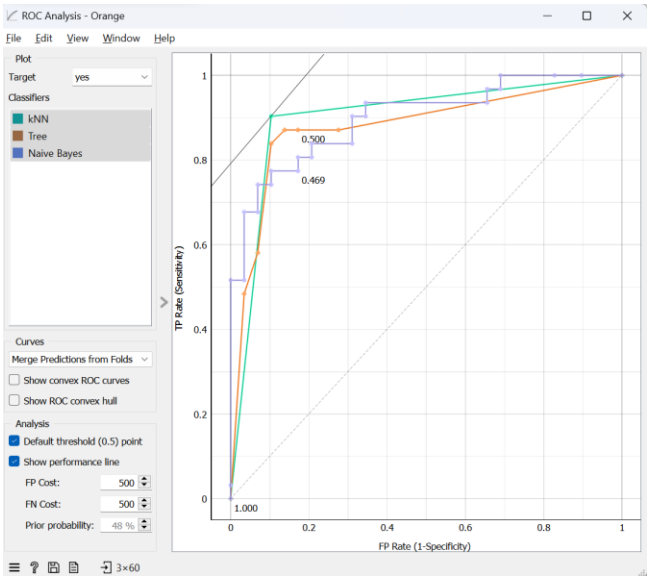
Gambar 3.
Hasil *Test & Score*

Dengan hasil yang ditunjukan pada Gambar 3 dapat dilihat bahwa *accuracy* yang diterapkan K-Nearest Neighbors (k-NN) sebesar 90,00% lebih tinggi dari Naïve Bayes dan Decision Tree. Berikut hasil Confusion Matrix nya (Gambar 4).



Gambar 4.
Confusion Matrix

Dan Berikut hasil ROC Analysis nya (Gambar 5).



Gambar 5.
ROC Analysis

Berdasarkan grafik ROC terlihat AUC (*Area Under Cover*) yang dihasilkan oleh model K-NN diatas 0,500 dengan nilai akuarsi yang terbilang medium atau rata-rata (Gambar 5). Hal tersebut dipengaruhi oleh jika semakin banyaknya data *traning* yang digunakan maka semakin baik dan semakin lengkap juga model klasifikasinya yang dibentuk berdasarkan fakta dari data traning yang diolah. Sehingga ketika saat melakukan prediksi pada klasifikasi data testing atau data yang baru, maka akurasi yang didapatkan akan semakin baik atau semakin tinggi.

IV. KESIMPULAN

Berdasarkan pengujian dari tiga metode machine learning yang telah diterapkan pada penelitian ini, memperoleh rata-rata akurasi sistem dengan penyembuhan metode crtyotherapy ketika menggunakan k-nearest Neighbors (k-nn) sebesar 90,00%, ketika menggunakan metode naïve bayes dihasilkan rata-rata akurasi sebesar 86,67% dan decision tree dihasilkan akurasi 88,33%. Namun jika diperhatikan ROC nya AUC K-NN lebih tinggi dari Decision Tree dan lebih rendah dari Naïve Bayes, dengan demikian setelah diterapkan menggunakan model k-nearest Neighbors (k-nn) nilai akurasi dalam melakukan klasifikasi lebih baik dibandingkan dengan kedua metode lainnya yang telah diuji. Selain itu, jarak akurasi setiap eksperimen dengan rata-rata akurasi lebih dekat saat menggunakan k-nearest Neighbors (k-nn) dibandingkan dengan naïve bayes dan decision tree, hal ini sesuai dengan nilai standar deviasi yang dihasilkan dari masing- masing metode.

DAFTAR PUSTAKA

[1] F. Khozeimeh et al., “Intralesional immunotherapy compared to cryotherapy in the treatment of warts,” Int. J. Dermatol., vol. 56, no. 4, pp. 474–478, 2017.
[2] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, “An expert system for selecting wart treatment method,” Comput. Biol. Med., vol. 81, pp. 167–175, 2017.
[3] H. Amalia and E. Evicienna, “Komparasi Metode Data Mining Untuk Penentuan Proses Persalinan Ibu Melahirkan,” J. Sist. Inf., vol. 13, no. 2, p. 103, 2017.
[4] N. Saputra, T. B. Adji, and A. E. Permanasari, “Analisis Sentimen Data Presiden Jokowi dengan Preprocessing Normalisasi dan Stemming menggunakan Metode Naïve Bayes dan SVM,” J. Din. Inform., vol. 5, no. 1, pp. 1–12, 2015.
[5] P. E. Utgoff, N. C. Berkman, and J. A. Clouse. Decision tree induction based on efficient tree restructuring. Machine Learning, 29:5–44, 1997.
[6] P. Piro, R. Nock, F. Nielsen, and M. Barlaud, “Leveraging k-NN for generic classification boosting,” Neurocomputing, vol. 80, pp. 3–9, 2012.
[7] D. Kurnianingtyas, B. A. Rahardian, D. P. Mahardika, A. K. A., and D. A. K., “Sistem Pendukung Keputusan Diagnosis Penyakit Sapi Potong Menggunakan K- Nearest Neighbors (K-NN),” J. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, p. 122, 2017.
[8] P. Harrington, Machine Learning in Action, vol. 37, no. 3. 2012.
[9] M. M. Jain and P. V. Richariya, “An Improved Techniques Based on Naïve Bayesian for Attack Detection,” Int. J. Emerg. Technol. Adv. Eng., vol. 2, no. 1, pp. 324–331, 2012.
[10] Yeni Kustiyahningsih and N. Syafa'ah, “Sistem Pendukung Keputusan Untuk Menentukan Jurusan Pada Siswa Sma Menggunakan Metode Knn Dan Smart,” JSII, vol. 1, no. 1, pp. 19–28, 2014.
[11] F. Gorunescu, “Data mining: Concepts, models and techniques,” Intell. Syst. Ref. Libr., vol. 12, 2011.
[12] Y. C. Liaw, M. L. Leou, and C. M. Wu, “Fast exact k nearest Neighbors search using an orthogonal search tree,” Pattern Recognit., vol. 43, no. 6, pp. 2351–2358, 2010.
[13] Y. C. Liaw, C. M. Wu, and M. L. Leou, “Fast k-nearest Neighbors search using modified principal axis search tree,” Digit. Signal Process. A Rev. J., vol. 20, no. 5, pp. 1494–1501, 2010.
[14] Bustami, “Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Data Nasabah,” TECHSI J. Penelit. Tek. Inform., vol. 4, pp. 127–146, 2010.
[15] N. D. Prayoga, N. Hidayat, and R. K. Dewi, “Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 8, pp. 2666–2671, 2018.
[16] J. Wu and Z. Cai, “Attribute weighting via differential evolution algorithm for attribute Weighted Naïve Bayes (WNB),” J. Comput. Inf. Syst., vol. 7, no. 5, pp. 1672– 1679, 2011.