PENERAPAN *DATA MINING* TERHADAP DATA COVID-19 MENGGUNAKAN ALGORITMA KLASIFIKASI

Rizka Dahlia[1], Nanik Wuryani[2], Sri Hadianti[3], Windu Gata[4], Arina Selawati[5]

[1234]Fakultas Ilmu Komputer, Sekolah Tinggi Ilmu Komputer Nusa Mandiri Jakarta
5, Jl. Kramat Raya No. 18 RT. 05 RW 7, Kwitang, Kec. Senen
Kota Jakarta Pusat DKI Jakarta 10450

[5]Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
2, Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen
Kota Jakarta Pusat, DKI Jakarta 10450
e-mail : 14002335@nusamandiri.ac.id[1], 14002336@nusamandiri.ac.id[2], 1sri.shv@nusamandiri.ac.id[3], windu@nusamandiri.ac.id[4], arina.asq@bsi.ac.id

ABSTRACT

Coronavirus 2019 or more commonly referred to as COVID-19 is a type of virus that attacks the respiratory system. Until now the number of spread and the number of deaths caused by this virus continues to increase. As of April 21, 2020, based on data from the WHO, the total number of cases infected with this virus reached 2,397,217 with 162 deaths from all over the world. For South Korea itself, as of March 21, 2020, the total number of infected cases was 10,683 with a total of 237 deaths. In this study, researchers conducted data processing on the spread of COVID-19 in South Korea with Rapidminer using a classification algorithm, namely Naïve Bayes, C4.5, and K-Nearest Neighbor by performing the stages of selection, preprocessing, transfotmating, data mining and interpretation or evaluating the quality of the best accuracy of 80.79% with AUC of 0.881 achieved by the Naïve Bayes algorithm. The distribution of the data found that the influential attribute of the isolated class factor from the patient contained in the sex attribute where more women experienced isolation.

Keywords— *COVID-19, data mining, classification, C4.5, Naïve Bayes, K-NN*

ABSTRAK

Coronavirus 2019 atau yang lebih sering disebut dengan *COVID*-19 merupakan salah satu virus jenis yang menyerang sistem pernapasan. Hingga saat ini angka penyebaran dan angka kematian yang diakibatkan virus ini terus bertambah. Per tanggal 21 April 2020, berdasarkan data dari WHO total kasus yang terinfeksi virus ini mencapai 2.397.217 dengan kasus kematian mencapai 162 kasus dari seluruh dunia. Untuk Korea Selatan sendiri, per tanggal 21 Maret 2020 total kasus yang terinfeksi mencapai 10.683 dengan total kematian sebanyak 237. Pada penelitian kali ini peneliti melakukan pengolahan data penyebaran COVID-19 di Korea Selatan dengan *Rapidminer* menggunakan algoritma klasifikasi yaitu *Naïve Bayes*, C4.5, dan *K-Nearest Neighbor* dengan melakukan tahapan *selection, preprocessing, transfotmating, data mining* dan i*nterpretation atau evaluation* menghasilkan akurasi terbaik 80,79% dengan AUC 0.881 yang diraih oleh algoritma *Naïve Bayes.* Distribusi data yang didapatkan bahwa atribut yang berpengaruh dari faktor class *isolated* dari pasien terdapat pada atribut *sex* dimana *female* lebih banyak yang mengalami *isolated*.

Kata Kunci— *COVID*-19, *data mining*, klasifikasi, C4.5, *Naïve Bayes*, K-NN

I. PENDAHULUAN

Penyakit *Coronavirus 2019* atau lebih dikenal dengan istilah *COVID*-19 merupakan suatu wabah yang awalnya terdeteksi di Kota Wuhan, Cina pada Desember 2019. Sebelum disebut sebagai *COVID*-19, WHO atau *World Health Organization* memberikan nama sementara virus baru ini sebagai *Coronavirus Novel 2019* (2019-nCoV). Dan pada 21 April 2020 WHO secara resmi menyebut virus 2019-nCoV menjadi *COVID*-19 [1][2].

COVID-19 bermula dari *betacoronavirus* (SARS-CoV-2) yang menyerang bagian saluran pernapasan bagian bawah yang berubah menjadi pneumonia di tubuh manusia. Virus *COVID*-19 merupakan *coronavirus* jenis baru. *COVID*-19 dianggap sebagai kerabat dari *Severe Acute Respiratory Syndrome* (SARS) dan *Middle East Respiratory Syndrome Coronavirus* (MERS) [1].

Berdasarkan data yang didapat dari WHO, terdapat 179 negara yang sudah terpapar virus *COVID*-19. Hal menandakan bahwa virus ini memiliki tingkat paparan yang sangat tinggi dan cepat. Cara penyebarannya juga sangat sederhana. Penyebarannya dapat berupa bersin, batuk, atau berinteraksi dengan orang yang sudah terinfeksi. Dan virus ini lebih rentan terhadap orang tua dan mereka yang memang sudah memiliki riwayat penyakit serius.

Per tanggal 21 April 2020, berdasarkan data dari WHO total kasus yang terinfeksi virus ini mencapai 2.397.217 dengan kasus kematian mencapai 162 kasus dari seluruh dunia. Untuk Korea Selatan sendiri, per tanggal 21 Maret 2020 total kasus yang terinfeksi mencapai 10.683 dengan total kematian sebanyak 237. Data ini masih melakukan pembaruan hingga saat ini dan jumlah yang terinfeksi, sembuh serta meninggal akan bertambah. Ada beberapa faktor yang mempengaruhi cepatnya penyebaran virus ini yaitu umur tua, banyaknya orang bepergian ke negara yang sudah terinfeksi, melakukan kontak dengan orang yang terinfeksi, dan sebagainya [3]. Faktor-faktor tersebut dapat menjadi data dan dapat diolah dengan *data mining.*

Penelitian terdahulu yang menggunakan *dataset Data Science for COVID-19 (DS4C)* yang diambil dari kaggle yang juga digunakan pada penelitian ini pernah dilakukan oleh Al-Najjar dan Al-Rousan membahas mengenai prediksi kesembuhan dan kematian pasien *Covid*-19 di Korea Selatan dengan algoritma yang digunakan yaitu *Artificial Neural Network* (ANN) dengan hasil yang didapatkan pada

penelitian ini adalah usulan untuk memperhatikan penyebab infeksi untuk peningkatan kesembuhan pasien *Covid*-19 dan pengendalian regional atau daerah untuk meminimalisir kematian[4].

Penelitian dengan *dataset* yang sama dilakukan juga oleh Muhammad dkk. yang membahas tentang penggunaan beberapa model untuk mendapatkan akurasi tertinggi. Model yang digunakan antara lain *Decision Tree, Support Vector Machine, Naïve Bayes, Logistic Regression, Random Forest,* dan *K-Newarest Neighbor*. Dengan akurasi paling tinggi yang didapatkan dari beberapa model yang digunakan adalah *Decision Tree* memiliki akurasi yang tertinggi dengan akurasi 99.85% [5]. Kekurangan dari penelitian ini adalah terpaku terhadap akurasi tanpa melihat matriks yang mempengaruhi baik atau buruknya sebuah model yang dibuat.

Melihat dari penelitian sebelumnya, penelitian yang akan dilakukan kali ini adalah melakukan klasifikasi dengan atribut yang digunakan yaitu *sex, age, city, infaction case* serta *state* dengan menggunakan model klasifikasi *Naïve Bayes*, C4.5, dan *K-Nearest Neighbor* dengan tujuan menghasilkan klasifikasi pasien Covid-19 di Korea Selatan dengan menguji atribut yang berhubungan dengan penentuan status pasien Covid-19.

1. *Naïve Bayes*

Naïve Bayes merupakan salah satu algoritma *data mining* yang terdapat pada klasifikasi yang di ambil dari nama seorang ahli matematika yang bernama *Thomas Bayes* dan merupakan seorang menteri *Prebysterian* Inggris. Algoritma *Naïve Bayes* menggunakan teknik percabangan matematika dengan mencari peluang terbesar dari kemungkinan dalam klasifikasi berdasarkan frekuensi tiap klasifikasi terhadap data *training* yang sering di sebut dengan teori probabilistik [6]. Adapun rumus perhitungan dari *Naïve Bayes* adalah sebagai berikut [7]:

$$P(X|Y) = \frac{P(Y|X)xP(X)}{P(Y)} \qquad (1)$$

Keterangan :

Y = data dengan kelas yang belum diketahui
X = hipotesis data Y merupakan suatu kelas spesifik
P(X|Y) = probabilitas hipotesis X berdasarkan kondisi Y
P(X) = Probabilitas hipotesis X
P(Y|X) = Probabilitas Y berdasarkan kondisi pada hipotesis X
P(Y) = Probabilitas Y

2. C4.5

C4.5 merupakan salah satu algoritma pada metode klasifikasi dengan menghasilkan sebuah pohon keputusan. Pohon keputusan berisikan aturan yang direpresentasikan dengan bahasa yang mudah dimengerti. Dalam membuat

sebuah pohon keputusan dibutuhkan atribut yang dipilih sebagai akar kemudian akan membentuk cabang setiap nilai. Proses itu akan diulang terus menerus hingga setiap cabang memiliki kelas yang sama[8]. Proses tersebut memiliki perhitungan yaitu sebagai berikut:

$$Gain(S, A) = Entropy(s) \sum n * Entropy(s) \quad (2)$$

Keterangan:
S = Himpunan Kasus
A = atribut
n = jumlah partisi dalam atribut
|Si| = jumlah kasus pada partisi ke-i
|S| = jumlah kasus dalam S

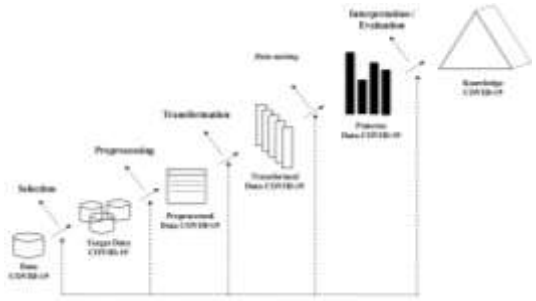
3. k-NN

k-NN atau *K-Nearest Neighbor* merupakan salah satu algoritma yang di pakai dalam melakukan klasifikasi . Algoritma k-NN melakukan klasifikasi dengan cara mencari *class* terdekat dengan jumlah data berupa *k* dengan class yang lain. *Class* terdekat dengan jumlah *k* terbesar dipilih sebagai *class* untuk diprediksi dengan *class* yang baru [9]. Untuk mencari dekat atau jauhnya sebuah *class* dicari dengan persamaan sebagai berikut.

$$d = \sqrt{(a_1 - b_1)^2} + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2 \quad (3)$$

II. METODE PENELITIAN

Metode atau tahapan penelitian sangat diperlukan ketika melakukan penelitian. Tahap penelitian dilakukan secara sistematis guna membantu penelitian agar terarah dalam melakukan penelitian. Adapun tahapan penelitian yang dilakukan pada penelitian ini sebagai berikut:



Gambar 1. Tahapan Penelitian *Data mining* COVID-19 Algoritma Klasifikasi

a. Selection

Sebelum melakukan proses *selection* lakukan terlebih dahulu pengumpulan data. Pengumpulan data dilakukan dengan mencari *dataset* pada *repository* atau jika memiliki data sendiri yang terdapat pada perusahaan jika memang dibutuhkan dalam penelitian. Kemudian pilih data yang sesuai dengan pembahasan yang ingin dibahas.

b. Preprocessing

Tahap preprocessing merupakan tahap di mana data yang sudah diseleksi kemudian dihilangkan berbagai macam data yang tidak memiliki isi atau

kandungan. Proses ini dilakukan agar mendapatkan kualitas data yang baik.

c. Transformation

Pada tahap ini data yang telah memalui tahap *preprocessing* akan dilanjutkan pemilihan atribut yang tepat untuk melakukan penelitian. Atribut ini dipilih untuk mempermudah dalam melakukan proses *data mining*.

d. *Data mining*

Proses *data mining* merupakan proses dimana algoritma yang sudah ditentukan untuk diterapkan pada *RapidMiner* dengan menggunakan data-data yang sudah diolah pada proses sebelumnya serta menerapkan model yang diinginkan. Pada Proses *data mining* menggunakan *RapidMiner* akan menghasilkan sebuah keputusan atau pola berdasarkan algoritma yang ingin digunakan.

e. Interpretation / Evaluation

Pola atau keputusan yang didapat pada proses *data mining* disajikan dalam bentuk yang mudah dimengerti. Hal ini dilakukan untuk mempermudah seseorang dalam membaca dan melakukan koreksi hasil yang telah didapat.

III. HASIL DAN PEMBAHASAN

Dalam penelitian yang menggunakan Data COVID-19 didapatkan di *repository* kaggle.com dengan nama *Science for COVID-19 (DS4C).* Terdapat beberapa data yang disediakan. Namun *dataset* yang diambil untuk melakukan penelitian ini terfokus kepada *dataset patien info*. *Dataset* ini berisikan informasi mengenai pasien yang terpapar COVID-19 di Korea Selatan. Berikut adalah bentuk data dari *dataset patien info*.

Tabel 1. Contoh Data *Patien info*

patient_id	global_num	sex	birth_year	age	country	province	city	disease	infection_case		infection_order	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date	deceased_date	state
1000000001	2	male	1964	50s	Korea	Seoul	Gangseo-gu		overseas inflow		1		75	1/22/2020	1/23/2020	2/5/2020		released
1000000002	5	male	1987	30s	Korea	Seoul	Jungnang-gu		overseas inflow		1		31		1/30/2020	3/2/2020		released
1000000003	6	male	1964	50s	Korea	Seoul	Jongno-gu		contact with patient		2	2002000001	17		1/30/2020	2/19/2020		released
1000000004	7	male	1991	20s	Korea	Seoul	Mapo-gu		overseas inflow		1		9	1/26/2020	1/30/2020	2/15/2020		released
1000000005	9	female	1992	20s	Korea	Seoul	Seongbuk-gu		contact with patient		2	1000000002	2		1/31/2020	2/24/2020		released
1000000006	10	female	1966	50s	Korea	Seoul	Jongno-gu		contact with patient		3	1000000003	43		1/31/2020	2/19/2020		released
1000000007	11	male	1995	20s	Korea	Seoul	Jongno-gu		contact with patient		3	1000000003	0		1/31/2020	2/10/2020		released
1000000008	13	male	1992	20s	Korea	Seoul	etc		overseas inflow		1		0		2/2/2020	2/24/2020		released
1000000009	19	male	1983	30s	Korea	Seoul	Songpa-gu		overseas inflow		2		68		2/5/2020	2/21/2020		released
1000000010	21	female	1960	60s	Korea	Seoul	Seongbuk-gu		contact with patient		3	1000000003	6		2/5/2020	2/29/2020		released
1000000011	23	female	1962	50s	China	Seoul	Seodaemun-gu		overseas inflow		1		23		2/6/2020	2/29/2020		released
1000000012	24	male	1992	20s	Korea	Seoul	etc		overseas inflow		1		0		2/7/2020	2/27/2020		released

Data tersebut memiliki beberapa atribut yang tidak diperlukan dalam proses pengolahan. Maka dari itu proses selanjutnya adalah memilih atribut yang dibutuhkan. Atribut yang digunakan dalam penelitian ini adalah *sex, age, city, infection case*, dan *state*. Atribut *state* berfungsi sebagai label keputusan yang akan dihasilkan. Terdapat dua *class* pada atribut *state* yang digunakan yaitu *released* dan *isolated*.

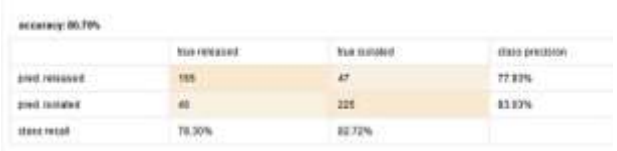
Tabel 2. Contoh Data *Patien info* Setelah Pemilihan Atribut

sex	age	city	infection_case	state
male	50s	Gangseo-gu	overseas inflow	released
male	30s	Jungnang-gu	overseas inflow	released
male	50s	Jongno-gu	contact with patient	released
male	20s	Mapo-gu	overseas inflow	released
female	20s	Seongbuk-gu	contact with patient	released
female	50s	Jongno-gu	contact with patient	released
male	20s	Jongno-gu	contact with patient	released
male	20s	etc	overseas inflow	released
male	30s	Songpa-gu	overseas inflow	released
female	60s	Seongbuk-gu	contact with patient	released
female	50s	Seodaemun-gu	overseas inflow	released
male	20s	etc	overseas inflow	released
male	80s	Jongno-gu	contact with patient	released
female	60s	Jongno-gu	contact with patient	released
male	70s	Seongdong-gu	Seongdong-gu APT	isolated
male	70s	Jongno-gu	contact with patient	released
male	70s	Jongno-gu	contact with patient	released
male	20s	etc	etc	isolated
female	70s	Jongno-gu	contact with patient	released

Setelah pemilihan atribut, kemudian menghapus data kosong. Proses menghilangkan data kosong ini dilakukan dengan melakukan *sorting* data kosong atau *blank*. Data yang didapatkan sebanyak 2420 data setelah melalui proses *sorting.* Dan data tersebut yang diterapkan di *Rapidminer* untuk proses pembuatan model *data mining* menggunakan algoritma C4.5, *Naïve Bayes*, dan k-NN.

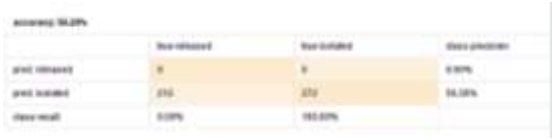
Gambar 2. Model *Rapidminer* Algoritma Klasifikasi

Setelah data dimasukkan ke dalam *RapidMiner*, jalankan model *RapidMiner*. Setelah *RapidMiner* dijalankan, maka akan tampil hasil sebagai berikut.



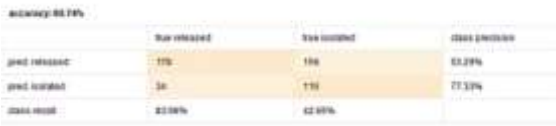
Gambar 3. Hasil Akurasi *Naïve Bayes*

Dari proses *Rapidminer* menggunakan algoritma *Naïve Bayes* menghasilkan akurasi sebesar 80,79% yang ditunjukkan pada gambar diatas.



Gambar 4. Hasil Akurasi C4.5

Untuk algoritma C4.5 mendapatkan hasil akurasi yang lebih kecil dari *Naïve Bayes* dengan hasil 54,20%.



Gambar 5. Hasil Akurasi k-NN

Sedangkan hasil yang didapatkan pada algoritma k-NN mendapatkan akurasi 60,74 %. Perbandingan akurasi dari ketiga algoritma dapat dilihat pada tabel berikut.

Tabel 3. Perbandingan Akurasi Algoritma C4.5, *Naïve Bayes*, dan k-NN

Algoritma	Akurasi	AUC
Naïve Bayes	80.79%	0.881
C4.5	56.20%	0.743
k-NN	60.74%	0.567

Dari tabel di atas dapat dilihat untuk nilai akurasi yang dicapai dari ketiga algoritma. *Naïve Bayes* memiliki akurasi 80,79% dengan AUC 0.881, algoritma C4.5 mendapatkan akurasi 56,20% dengan AUC 0.743 dan k-NN mendapatkan akurasi 60,74% dengan AUC 0.567. Jika ketiga algoritma tersebut dibandingkan maka nilai akurasi dan AUC terbaik adalah algoritma *Naïve Bayes.* Berikut adalah penyebaran data dari algoritma *Naïve Bayes*.

Tabel 4. Penyebaran Data pada Algotima *Naïve Bayes*

Attribute	Parameter	released	isolated
sex	value=female	0.555	0.560
sex	value=male	0.445	0.440
infection_case	value=contact with patient	0.346	0.388
infection_case	value=overseas inflow	0.074	0.305
age	value=20s	0.251	0.233
infection_case	value=etc	0.279	0.175
age	value=50s	0.177	0.168
age	value=30s	0.163	0.144
age	value=60s	0.096	0.139
age	value=40s	0.187	0.132
city	value=Seongnam-si	0.000	0.086
age	value=70s	0.037	0.069
age	value=10s	0.038	0.048
city	value=Gyeongsan-si	0.033	0.046
age	value=80s	0.026	0.043

IV. SIMPULAN

Hasil dari klasifikasi pada penelitian ini dengan atribut *sex, age, city, infection case,* dan *state* menggunakan algoritma C4.5, *Naïve Bayes*, dan k-NN pada *Rapidminer* menghasilkan banyak pasien masuk ke class *isolated.* Dari ketiga algoritma, didapatkan akurasi tertinggi dari algoritma *Naïve Bayes* dengan hasil 80,79% dengan hasil *isolated* dan memiliki AUC 0.881 yang dikategorikan sebagai *good clasification* atau klasifikasi yang baik. Adapun hasil distribusi data yang didapatkan bahwa atribut yang berpengaruh dari faktor class *isolated* dari

pasien terdapat pada atribut *sex* dimana *female* lebih banyak yang mengalami *isolated* dari algoritma *Naïve Bayes.*

DAFTAR PUSTAKA

[1] C. Sohrabi *et al.*, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," *International Journal of Surgery*. 2020, doi: 10.1016/j.ijsu.2020.02.034.

[2] P. Sun, X. Lu, C. Xu, W. Sun, and B. Pan, "Understanding of COVID-19 based on current evidence," *Journal of Medical Virology*. 2020, doi: 10.1002/jmv.25722.

[3] S. A. Rasmussen, J. C. Smulian, J. A. Lednicky, T. S. Wen, and D. J. Jamieson, "Coronavirus Disease 2019 (COVID-19) and Pregnancy: What obstetricians need to know.," *Am. J. Obstet. Gynecol.*, 2020, doi: 10.1016/j.ajog.2020.02.017.

[4] H. Al-Najjar and N. Al-Rousan, "A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 24, no. 6, pp. 3400–3403, 2020, doi: 10.26355/eurrev_202003_20709.

[5] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive *Data mining* Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," *SN Comput. Sci.*, vol. 1, no. 4, pp. 1–7, 2020, doi: 10.1007/s42979-020-00216-w.

[6] R. A. Saputra, "Komparasi Algoritma Klasifikasi *Data mining* Untuk Memprediksi Penyakit Tuberculosis (Tb): Studi Kasus Puskesmas Karawang," *Semin. Nas. Inov. dan Tren*, 2014.

[7] W. D. Septiani, "Komparasi Metode Klasifikasi *Data mining* Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.

[8] F. Parung, "Penerapan algoritma decision tree c4.5 dalam penerimaan guru pada smk sirajul falah parung," vol. 11, no. 2, pp. 192–198, 2018.

[9] I. G. Harsemadi, M. Sudarma, and N. Pramaita, "Implementasi Algoritma K-Nearest Neighbor pada Perangkat Lunak Pengelompokan Musik untuk Menentukan Suasana Hati," *Maj. Ilm. Teknol. Elektro*, vol. 16, no. 1, pp. 14–20, 2017, doi: 10.24843/mite.1601.03.

[10] Sabita, H., & Herwanto, R. (2020). Pantauan Prediktif Covid-19 Dengan Menggunakan Metode SIR dan Model Statistik Di Indonesia. *TEKNIKA*, *14*(2), 145-150.