**Klasifikasi Lama Waktu Gejala Vaksin Covid-19 dengan Metode *K-Nearest Neighbor*, *Random Forest* dan *Support Vector Machine***

Studi Kasus Pegawai Rumah Sakit Di Rumah Sakit Rujukan Nasional Indonesia

*(Classification of the Duration of Covid-19 Vaccine Symptoms with the K-Nearest Neighbor, Random Forest and Support Vector Machine Methods)*

Riofebri Prasetia1, Arie Wahyu Wijayanto2

*1Politeknik Statistika Sekolah Tinggi Ilmu Statistik*

E-mail: [221911192@stis.ac.id](mailto:abcd@ggg.com)

**ABSTRAK**

Pada akhir tahun 2019 virus penyakit covi-19 muncul di Wuhan, China. Virus ini memiliki kemampuan menyebar yang sangat cepat sehingga tidak lama kemudian dilaporkan sebanyak lebih dari 100 negara terkena wabah ini termasuk Indonesia. Satu di antara pencegahan penyebaran covid-19 yakni pemberian vaksin pada masyarakat. Akan tetapi terdapat efek samping pada penggunaannya. Karena belum jelasnya efek samping tersebut, untuk itu perlu model status lama waktu gejala setelah vaksin untuk ukuran dalam melakukan tindak pencegahan timbulnya potensi penyakit lain karena efek samping tersebut. Saat ini belum ada penelitian yang membahas model tersebut. Kemungkinan model ini masih dibuat dalam bentuk manual. Untuk itu perlu teknik *machine learning* dalam melakukan klasifikasi model agar pemodelan bisa dilakukan dengan akurat dan efisien dibanding dengan model yang dilakukan secara manual. Adapun model yang direkomendasikan yaitu *K-Nearest Neighbor* (KNN), *Random forest* (RF), dan *Support Vector Machine* (SVM). Hasilnya model *Support Vector Machine* memiliki akurasi jauh lebih tinggi sebesar 82% dibanding model *Random Forest* dan *K-Nearest Neighbor* dengan masing-masing akurasi 73% dan 69%. Sehingga model *Support Vector Machine* adalah yang direkomendasikan untuk digunakan.

**Kata kunci**: waktu gejala, covid-19, model

***ABSTRACT***

*At the end of 2019, the COVID-19 virus emerged in Wuhan, China. This virus has the ability to spread very quickly, so it was not long before it was reported that more than 100 countries were affected by this outbreak, including Indonesia. One of the ways to prevent the spread of COVID-19 is the provision of vaccines to the public. However, there are side effects to its use. Because the side effects are not yet clear, it is necessary to model the duration of symptoms after the vaccine for measures to prevent the emergence of other potential diseases due to these side effects. There are currently no studies that discuss this model. The possibility of this model is still made in manual form. For this reason, machine learning techniques are needed in classifying models so that modeling can be done accurately and efficiently compared to models that are done manually. The recommended models are K-Nearest Neighbor (KNN), Random forest (RF), and Support Vector Machine (SVM). As a result, the Support machine learning model has a much higher accuracy of 82% than the Random forest and K-Nearest Neighbor models with 73% and 69% accuracy, respectively. So The Support Vector Machine model is the recommended one to use.*

***Keywords****: symptom time, covid-19, model*

**PENDAHULUAN**

Pada akhir tahun 2019, Dunia dikejutkan dengan kemunculan penyakit Covid-19 yang terjadi di Wuhan, Cina. Menurut *Shereen et al. (2020)*, Covid-19 merupakan suatu wabah infeksi virus dan patogen yang disebabkan oleh Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2). Pada tanggal 12 Januari 2020, The National Health Commission of China merilis bahwa hal tersebut merupakan bagian dari virus pneumonia. Penyebaran virus ini sangat cepat hingga tidak lama kemudian dilaporkan lebih dari 100 negara di dunia terkena wabah termasuk Indonesia. Banyak kebijakan yang telah diterapkan untuk mengatasi penyebaran wabah, satu di antaranya yaitu pemberian vaksinasi ke masyarakat dengan harapan bisa membangun kekebalan komunitas (Farisa, 2020).

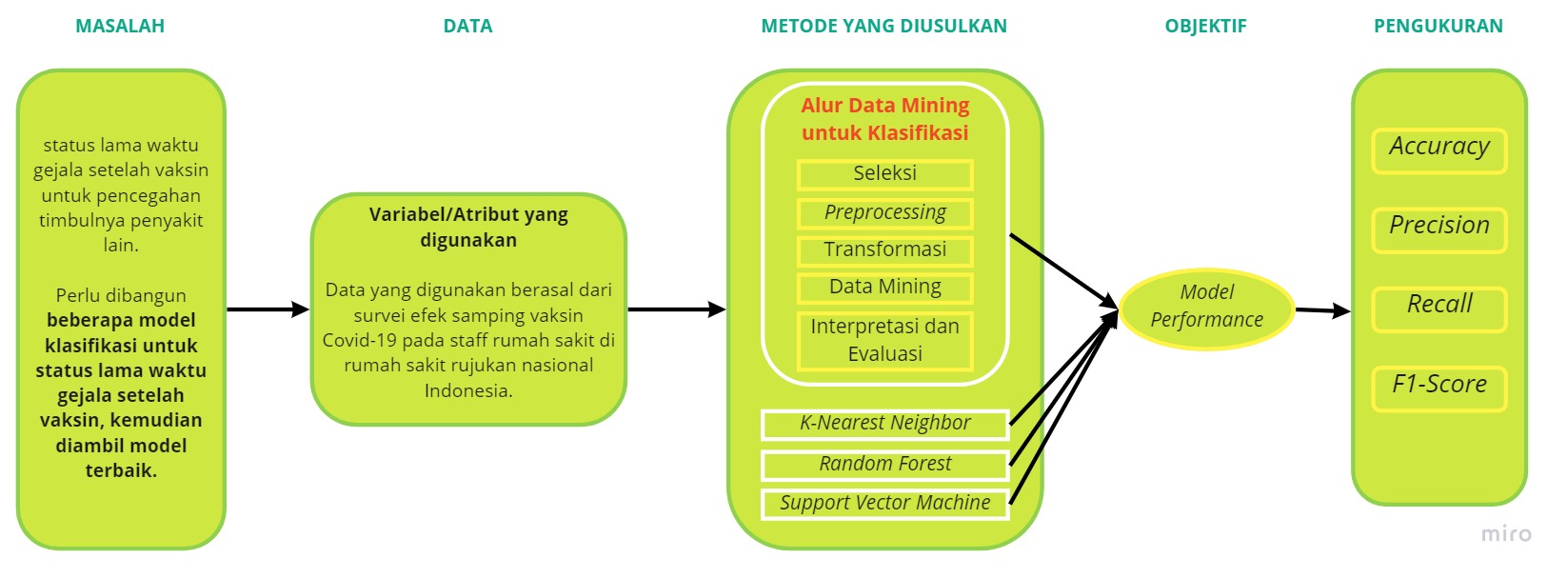
Hingga saat ini pemberian vaksin mencapai lebih dari 200 juta suntikan dimana sebanyak 123% suntikan dosis pertama, 77% suntikan dosis kedua dan 1% untuk suntikan dosis ke tiga (Rokom, 2021). Dalam penelitian Kadali et al. (2021), Salah satu jenis vaksin yakni BNT162b2 mRNA covid-19 memiliki beberapa gejala yang terbagi ke dalam kelas gejala umum dan gejala khusus. Gejala umum seperti sakit kepala, demam, berkeringat. Sedangkan gejala khusus seperti diare, maag dan gejala khusus lainnya. Dalam penanganan kasus covid-19 perlu adanya keputusan yang cepat, tepat dan efisien tidak hanya kepada penderita covid-19 tapi juga terhadap orang yang sudah melakukan vaksinasi (Handayani et al, 2020). Oleh karena itu, perlu adanya suatu model hasil analisis dari survey gejala penerima vaksin setelah vaksinasi. Model ini digunakan untuk tindak pencegahan timbulnya potensi penyakit lain setelah vaksinasi secara efisien. Karena waktu gejala setelah vaksin merupakan suatu hal yang mudah untuk diukur. Dalam hal ini, model klasifikasi lama waktu gejala setelah vaksin merupakan model yang cocok untuk menangani masalah tersebut saat ini.

Dalam Çiǧşar & Ünal (2019), terdapat teknik untuk melakukan analisis data dalam jumlah besar yang dinamakan “*Data mining*”. *Data mining* bertujuan untuk mengolah dan mengekstraksi data besar menjadi informasi yang berguna. Beberapa teknik utama Data mining ialah seperti klasifikasi dan prediksi, pengelompokan, deteksi outlier, aturan asosiasi (Y. Zhao, 2015). Klasifikasi bisa digunakan untuk mengelompokkan kelas yang sesuai pada kelompok lama waktu gejala setelah vaksinasi berdasarkan hasil survey responden. Beberapa metode klasifikasi yang biasa digunakan untuk prediksi yaitu *k-nearest neighbor, bayesian method, support vector machine dan artificial neural networks* (Gayathri & Satapathy, 2020)dalam (Shaban et al., 2020)*.*

Beberapa metode klasifikasi dilakukan penelitian di bidang medis. *K-nearest Neighbor* merupakan metode klasifikasi yang cepat dengan menggunakan pendekatan berdasarkan jumlah k tetangga terdekat. Dalam H. Zhao et al. (2021), *K-Nearest Neighbor* digunakan dalam pengembangan metode *ensemble k-nearest neighbor*. Metode ini digunakan untuk mendeteksi tingkat keparahan penyakit Parkinson dengan data yang *imbalance* atau tidak seimbang. Data diambil dari rekaman medis frekuensi. Model klasifikasi ini menghasilkan akurasi 95% dan standar deviasi 0,44% yang artinya hasil prediksi memiliki tingkat kesalahan yang kecil. Kemudian metode *random forest* juga digunakan dalam penelitian yang berkaitan dengan medis. Dalam Iwendi et al. (2020), *random forest* di modifikasi dengan algoritma *adaboost*. Model ini digunakan untuk memprediksi tingkat keparahan kasus dan probabilitas hasil kesehatan pasien covid-19. Dengan menggunakan data kesehatan dan geografi di dapat model dengan akurasi 94% dan *F1-score* 86%. Hasil dari model ini menunjukkan bahwa mayoritas pasien ialah pada usia 20 dan 70 tahun. Selanjutnya dalam Devikanniga et al (2020), *support vector machine* dengan *crows search algorithm* atau *CSA-SVM.* Metode ini digunakan untuk diagnosis efisien penyakit hati. Dalam penelitian ini secara keseluruhan memperoleh model klasifikasi dengan tingkat akurasi 99,49% berdasarkan dari data penyakit hati di India. Model ini juga termasuk memiliki akurasi yang sangat tinggi bila dibandingkan dengan model pendekatan lainnya.

Berdasarkan yang sudah di paparkan, peneliti tertarik untuk mengetahui model klasifikasi “lama waktu gejala setelah vaksinasi” dengan pendekatan data mining. Hal ini dikarenakan, belum adanya model klasifikasi yang berkaitan dengan “lama waktu gejala vaksin covid-19” dan pengembangannya. Sehingga dimungkinkan pembuatan model yang ada, masih dilakukan secara manual. Selain itu, juga dilakukan perbandingan metode klasifikasi antara metode *K-Nearest Neighbor*, *Random Forest* dan *Support Vector Machine* dalam pembuatan model ini.

**METODE**



Gambar 1. Kerangka berpikir

1. Sumber data dan deskripsi atribut

Data yang digunakan berasal dari survei efek samping vaksin Covid-19 pada staf rumah sakit di rumah sakit rujukan nasional Indonesia. Data berjumlah 840 responden yang mengisi survey dan terdapat sebanyak 22 pertanyaan atau atribut.

Table 1. Tabel Variabel dan Posisi Variabel

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Nama Atribut/Variabel** | **Nilai** | **Tipe Data** | **Variabel** |
| 1 | Jenis Kelamin | * Laki-laki * Perempuan | * String | * Independen |
| 2 | Usia | * < 20 tahun * 20-25 tahun * 26-30 tahun * 31-35 tahun * 36-40 tahun * 41-45 tahun * 46-50 tahun * 51-55 tahun * 56-60 tahun * > 60 tahun | * String | * Independen |
| 3 | Profesi | * Bidan * Perawat * Dokter medis * Dokter Spesialis * Staff non medis | * String | * Independen |
| 4 | Pendidikan | * SMP * SMA * Sarjana/S1 * Master/S2 | * String | * Independen |
| 5 | Tempat tinggal | * Pusat Kota * Pinggir Kota | * String | * Independen |
| 6 | Pembengkakan | * Ya * Tidak | * String | * Independen |
| 7 | Kemerahan | * Ya * Tidak | * String | * Independen |
| 8 | Gatal | * Ya * Tidak | * String | * Independen |
| 9 | Demam | * Ya * Tidak | * String | * Independen |
| 10 | Sakit Kepala | * Ya * Tidak | * String | * Independen |
| 11 | Nyeri Otot | * Ya * Tidak | * String | * Independen |
| 12 | Kelelahan | * Ya * Tidak | * String | * Independen |
| 13 | Batuk | * Ya * Tidak | * String | * Independen |
| 14 | Diare | * Ya * Tidak | * String | * Independen |
| 15 | Mual dan Muntah | * Ya * Tidak | * String | * Independen |
| 16 | Sesak Napas | * Ya * Tidak | * String | * Independen |
| 17 | Nyeri Sendi | * Ya * Tidak | * String | * Independen |
| 18 | Pingsan | * Ya * Tidak | * String | * Independen |
| 19 | Reaksi Anafilaksis | * Ya * Tidak | * String | * Independen |
| 20 | Perasaan geli | * Ya * Tidak | * String | * Independen |
| 21 | Pembengkakan kelenjar getah bening | * Ya * Tidak | * String | * Independen |
| 22 | Waktu Gejala | * Tidak ada * < 24 jam * 24-72 jam * > 72 jam | * String | * Dependen |

1. *K-Nearest Neighbor (KNN)*

*K-Nearest Neighbor* merupakan metode pengklasifikasian dengan tetangga terdekat berdasarkan pembelajaran dengan analogi. “Nearest” atau Kedekatan didefinisikan dengan jarak metrik, seperti jarak *Euclidian* (Han & Kamber, 2006)*. Instance* di klasifikasi berdasarkan jarak dengan k tetangga terdekat dimana k merupakan parameter bebas. Berdasarkan Parvin et al. (2010) dalam Ayyad et al. (2019) KNN memiliki kelebihan sebagai berikut:

* Klasifikasi dapat dengan mudah diterapkan;
* Ini kuat untuk *data train* yang *noisy* atau punya gangguan;
* Sesuai untuk data latih yang besar.

Namun juga memiliki kelemahan sebagai berikut:

* Pemilihan nilai K, yang berpengaruh pada kinerja KNN.
* Persyaratan perhitungan jarak untuk k tetangga tersebut.
* Degradasi akurasi dalam kasus kumpulan data multidimensi

**Kelas A**

**Kelas B**

**X1**

**X2**

objek baru

untuk

diklasifikasi

**k = 3**

**k = 6**

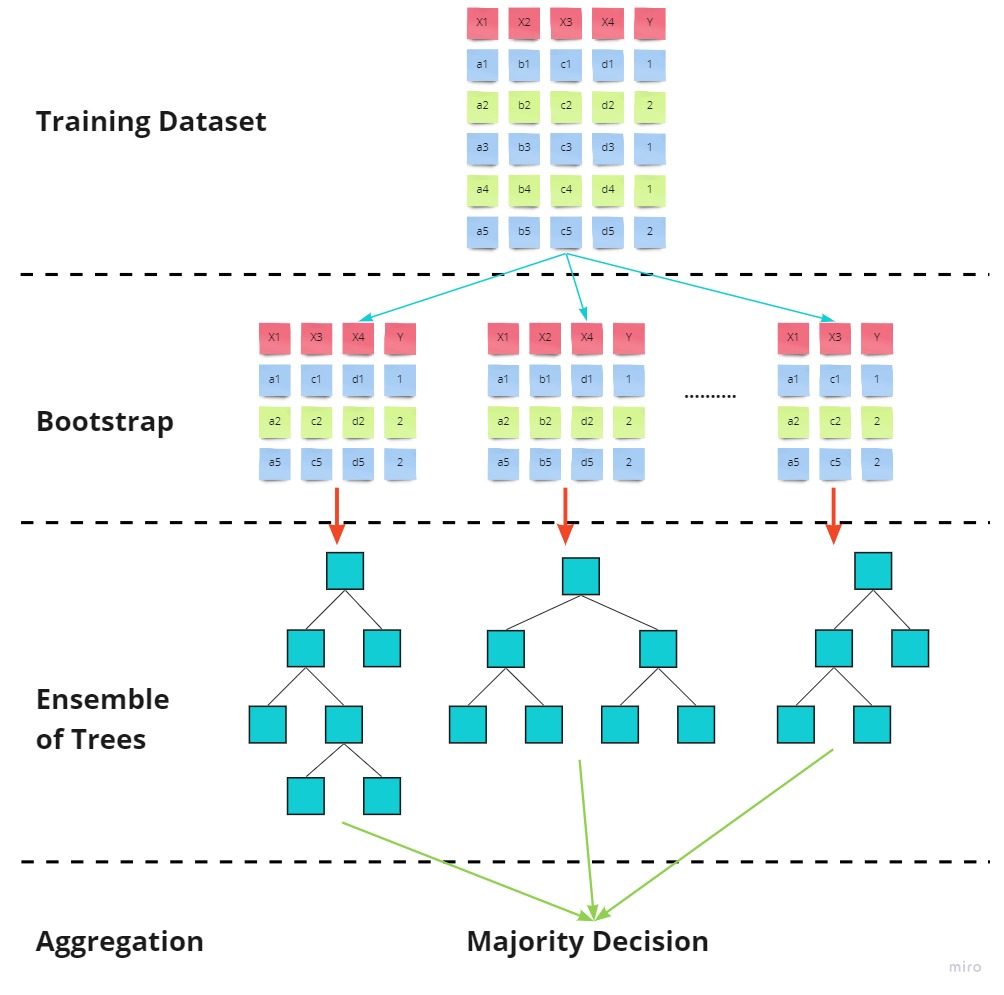
Gambar 2. Ilustrasi penggunaan parameter k pada Metode K-Nearest Neighbor (Angreni et al., 2018)

1. *Random Forest (RF)*

*Random Forest* merupakan satu diantara pembelajaran ensemble. Berdasarkan Speiser et al. (2019), metode ini menggunakan model sederhana yang menggunakan pemisahan biner pada variabel prediksi untuk menentukan prediksi hasil. Sistem kerja *Random Forest* sendiri yaitu mengambil subset *feature* secara acak sehingga secara sistematis menghindari korelasi dan meningkatkan kinerja model (Singh, 2019). Menurut Xia (2020) dalam *Random Forest - an Overview | ScienceDirect Topics* (2020), kelebihan dari RF meliputi

* Tidak rentan terhadap overfitting;
* Sangat mudah digunakan karena hanya dua parameter yang diperlukan;
* Daya pembeda yang sangat tinggi sehingga memungkinkan akurasi klasifikasi yang tinggi;
* Tidak ada asumsi distribusi tentang variabel prediktor atau respons;
* Dapat menangani situasi di mana jumlah variabel prediktor jauh melebihi jumlah pengamatan.

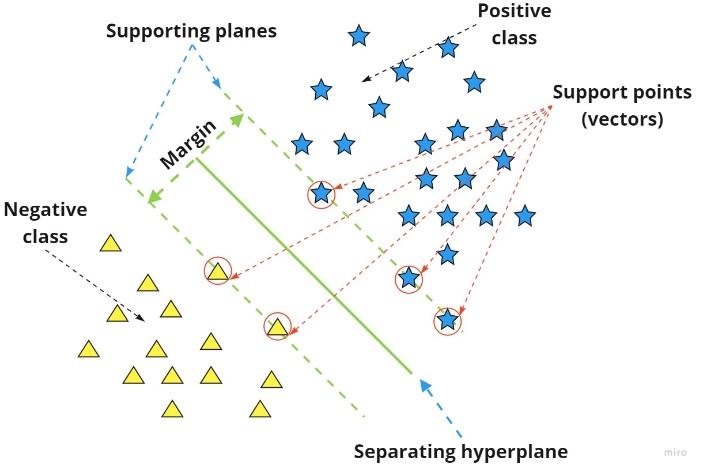
Namun kelemahan dari RF ini yaitu memiliki waktu training yang sangat lama.



Gambar 3. Ilustrasi Metode Random Forest Sumber (Misra & Li, 2020) dalam (Random Forest - an Overview | ScienceDirect Topics, 2020)

1. *Support Vector Machine (SVM)*

Berdasarkan (Han & Kamber, 2006), *Support Vector Machine* merupakan metode untuk klasifikasi data linier dan non linier menggunakan pemetaan non linier untuk mengubah *training dataset* menjadi dimensi yang lebih tinggi. Pada dimensi baru dicari *hyperplane* pemisah optimal linier. *Hyperplane* ini menggunakan vektor dukungan dan margin yang didefinisikan vektor dukungan. Dalam Han & Kamber (2006), kemampuan SVM dalam melakukan training berjalan sangat lambat dan rentan terhadap overfitting dibanding metode lain. Walaupun demikian kelebihan dari metode ini memiliki akurasi yang tinggi karena kemampuannya dalam memodelkan batas-batas keputusan non linier yang kompleks.



Gambar 4. Ilustrasi Metode Support Vector Machine (Ragab et al. (2021) dalam (Support Vector Machine - an Overview | ScienceDirect Topi, 2018)

1. Alur Data Mining

Data Mining

Preprocessing

Seleksi

Transformasi

Interpretasi dan Evaluasi

Gambar 5. Diagram alur Data Mining.

* Seleksi

Mengubah dataset menjadi format yang sesuai untuk analisis. Pada tahap ini akan diseleksi atribut. Atribut yang di buang ialah atribut yang secara teori tidak diperlukan. Selanjutnya dilakukan reduksi dimensi dan sub setting data jika dibutuhkan.

* Preprocessing

Merupakan tahap pembersihan data dimana informasi yang tidak dibutuhkan dibuang. Data di setting ulang untuk memastikan format konsisten.

* Transformasi

Pada tahap ini data di transformasi menjadi bentuk yang bisa di analisis lebih lanjut. Contoh transformasi yakni normalisasi sehingga data bernilai minimum 0 dan maksimum 1.

* Data Mining

Pada tahap ini dilakukan pengujian model yang mana label yang diklasifikasikan ialah “Waktu gejala”. Akan dilakukan pengujian model dimana proporsi yang ditentukan ialah 67% untung *training set* dan 33% untuk *test set.* Setelah dilakukan pengujian model, selanjutnya dilakukan evaluasi dan validasi hasil. Untuk evaluasi hasil menggunakan confusion matrix dan hasil klasifikasi berupa *precision, recall/sensitivity*, *F1-score* dimana dengan ini bisa dihitung *accuracy.* Kemudian pada validasi hasil digunakan *cross validation.*

1. *Confusion Matrix*

*Confusion matrix* merupakan gambaran pengelompokan kelas yang terdiri dari kelas sebenarnya/*retrieved* dengan kelas yang diprediksi/*relevant*. Dengan ini bisa digambarkan jumlah data/*instance* yang termasuk kelompok kelas berikut:

Table 2. Configuration Matrix (Han & Kamber, 2006)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predicted Class  */Relevant* | True/Actual Class/*Retrieved* | | | |
| P/A | Positive | Negative |  |
| True | True Positif  (TP) | False Positive  (FP) | P’ |
| False | False Negatif  (FN) | True Negative  (TN) | N’ |
|  | P | N | All |

1. *Precision*

Presisi adalah persentase yang relevan/*relevant* dengan tanggapan “benar” atau *True Positive*. Secara formal didefinisikan sebagai:

1. *Recall/sensitivity*

*Recall* merupakan persentase yang relevan/*relevant* dengan tanggapan “benar” yang pada kenyataannya diambil/*retrieved.*  Secara formal dapat didefinisikan sebagai:

1. *F1-score*

F1-score merupakan perbandingan rata-rata harmonik dari *recall* dan *precision.* Secara umum dapat didefinisikan sebagai berikut:

1. *Accuracy*

Ketepatan prediksi baik yang bernilai *positive* maupun bernilai *negative*, diukur dari persentase *accuracy.* Secara umum dapat didefinisikan sebagai berikut:

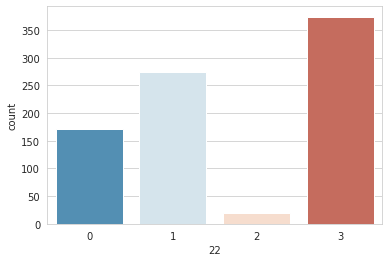
1. *Cross validation*

*Cross validation* merupakan satu di antara metode *resampling* data sampel untuk menilai generalisasi kemampuan model *predictive* dan mencegah overfitting. Menurut Berrar (2018), *cross validation* Biasanya digunakan untuk memberikan perkiraan kinerja model dengan menyetel parameter model yang diterapkan beberapa kali pada nilai parameter yang berbeda, dan parameter yang dicoba bisa meminimalkan kesalahan yang kemudian digunakan untuk membangun model akhir. Dengan demikian *cross validation* mengatasi masalah overfitting.

* Interpretasi dan evaluasi

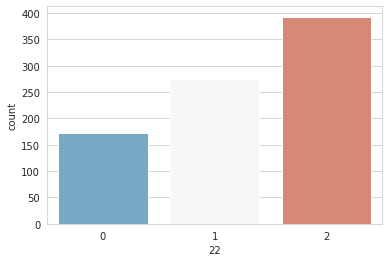
Setelah didapat hasil terbaik, maka hasil model akan di interpretasi dan evaluasi. Adapun interpretasi dan evaluasi ini difokuskan pada *accuracy, precision, sensitivity* serta akurasi model. Serta bagaimana model yang terbaik dari berbagai percobaan yang dilakukan sesuai dengan kondisi data yang ada.

**HASIL DAN PEMBAHASAN**



Gambar 6. Diagram batang frekuensi kolom pertanyaan waktu gejala

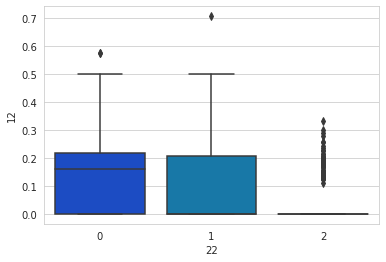
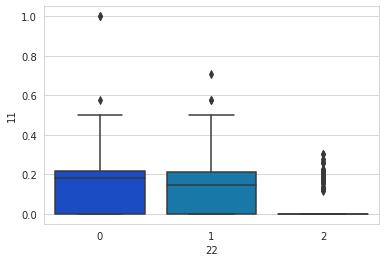
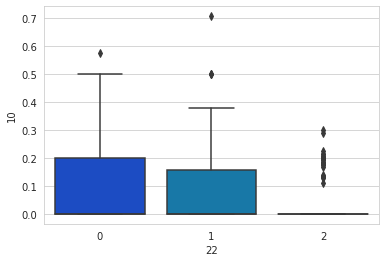
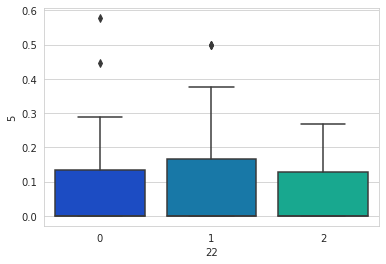
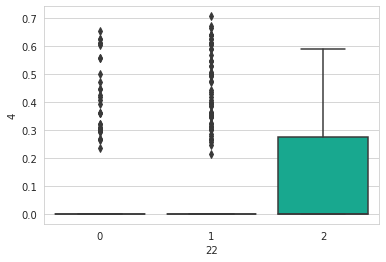
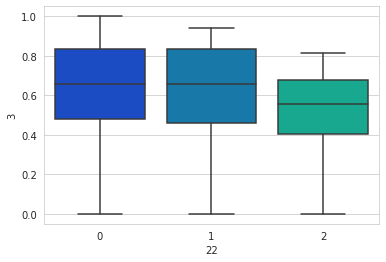
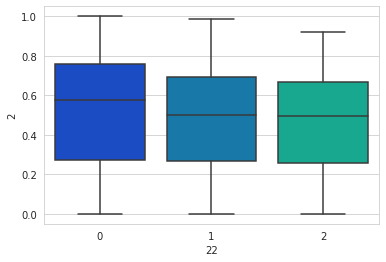
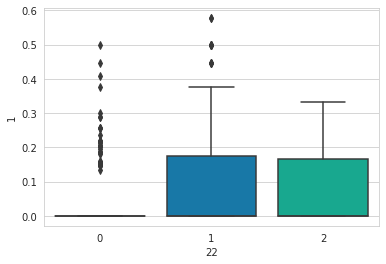
Pada atribut “22” yaitu “waktu gejala” terjadi extreme *imbalance* pada opsi jawaban 2 yakni 24 – 72 jam. Sehingga dalam hal ini dilakukan agregasi baris sehingga dijadikan 3 kelompok kelas yakni waktu gejala: “tidak ada”, “< 24 jam” dan “> 24 jam”. Sehingga menjadi:



Gambar 7. Diagram batang frekuensi kolom pertanyaan waktu gejala setelah agregasi baris

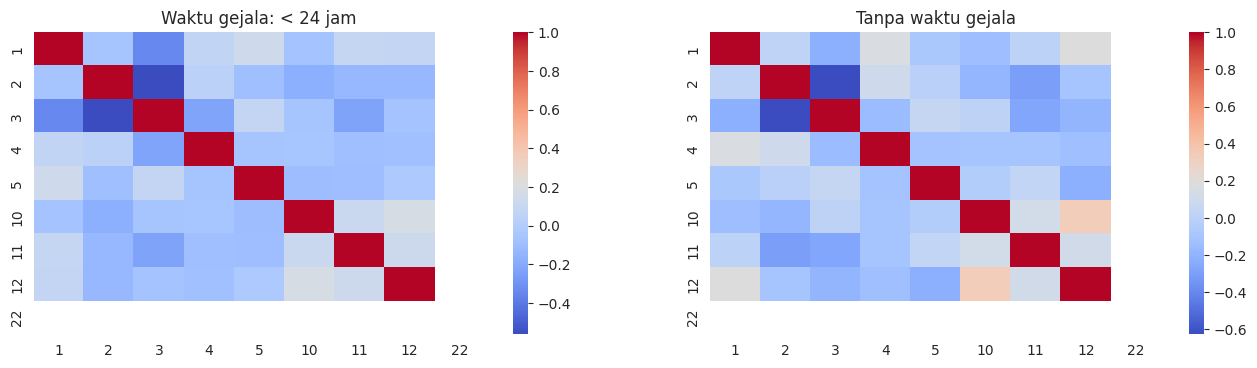
Jika diperhatikan terjadi imbalance dimana jumlah responden yang menjawab lama waktu gejala: “tidak ada” lebih kecil kurang dari 50% jumlah responden yang menjawab “> 24 jam”. Selanjutnya di lakukan drop pada atribut no 8 yakni pertanyaan tentang “gatal” karena jawaban semua responden adalah “tidak” atau hanya memiliki 1 jenis nilai, sehingga atribut ini tidak bisa membantu dalam klasifikasi.

Selanjutnya di lihat bagaimana sebaran masing-masing kelas berdasarkan masing-masing atribut. Atribut yang memiliki sebaran antar kelas yang mirip akan di *drop* atau dihilangkan. Berikut merupakan atribut yang terpilih:

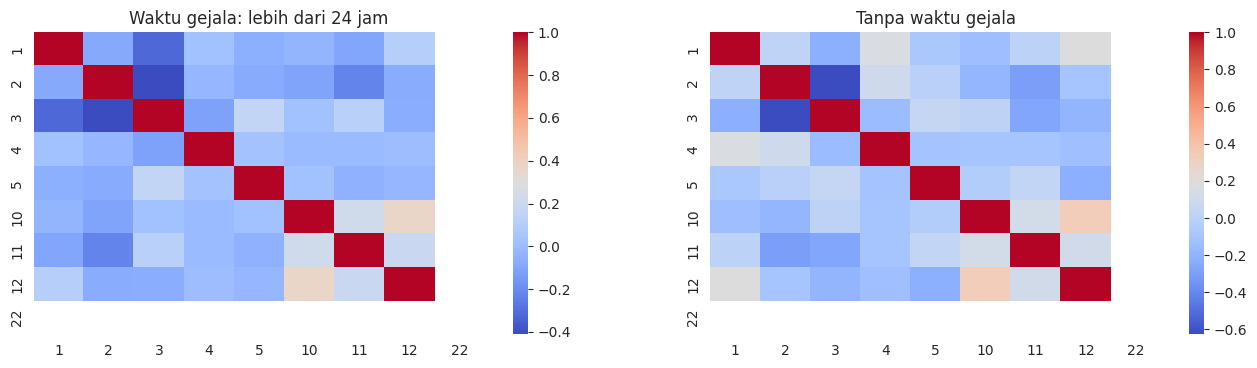


Gambar 8. Boxplot kelas berdasarkan atribut

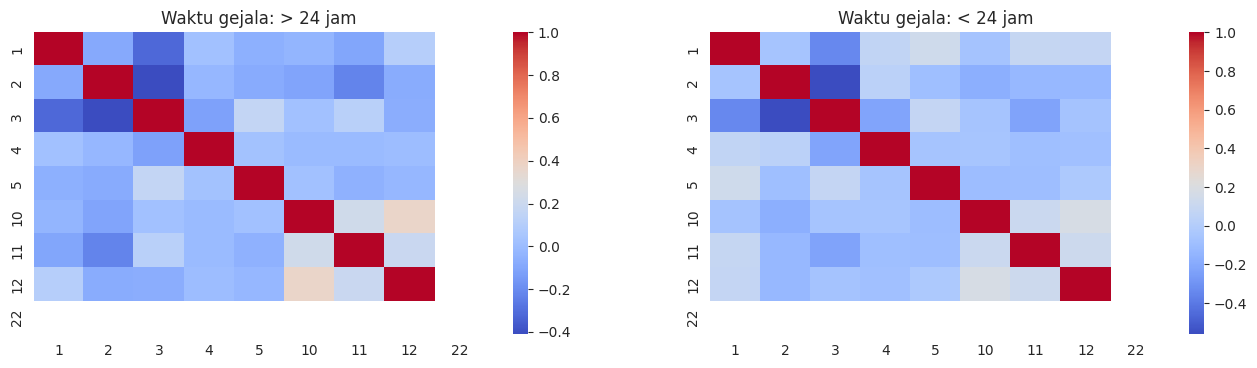
Berdasarkan boxplot, terdapat *outliers*. Akan tetapi tidak perlu suatu tindakan untuk mengatasi *outliers* tersebut. Kemudian bisa dilihat bagaimana pengaruh atributdalam mengklasifikasikan kelas “Waktu gejala” serta perbandingannya pada nilai klasifikasi yang lain seperti pada gambar:



Gambar 9. Korelasi perbedaan antar atribut terhadap label antara kelas “waktu gejala < 24 jam” (kelas kode “1”) dan kelas “Tanpa waktu gejala” (kode kelas “0”).



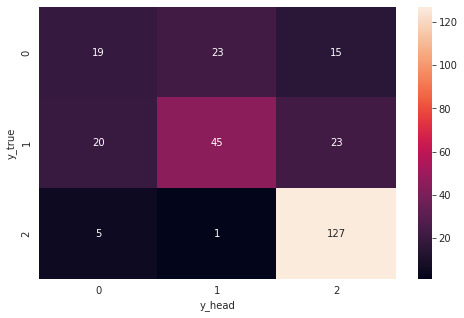
Gambar 10. Korelasi perbedaan antar atribut terhadap label antara kelas “waktu gejala > 24 jam” (kelas kode “2”) dan kelas “Tanpa waktu gejala” (kode kelas “0”).



Gambar 11. Korelasi perbedaan antar atribut terhadap label antara kelas “waktu gejala < 24 jam” (kelas kode “1)” dan kelas “waktu gejala > 24 jam” (kode kelas “2”).

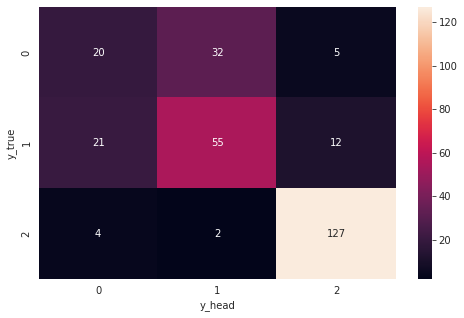
Hingga saat ini masih belum terdapat antara atribut penjelas yang memiliki korelasi tinggi (“r” lebih dari 0,7). Sehinggatidak terdapat redundansi pada data. Sehingga atribut tersebut dapat digunakan untuk membantu dalam klasifikasi.

Kemudian didapat hasil dari “*cross validation”* dimana ketika dilakukan uji dengan data set yang sama, menghasilkan klasifikasi model dengan *confusion matrix*:



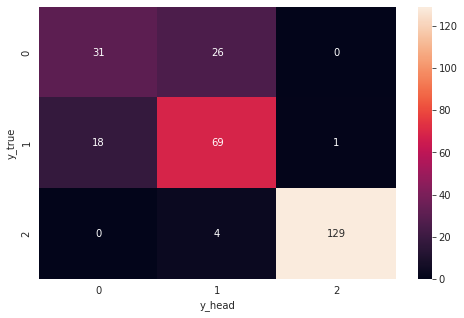
Gambar 12. Confusion Matrix pada model K-Nearest Neighbor

Dalam *Confusion matrix* model *K-Nearest Neighbor* bisa kita lihat bahwa terdapat error dimana terdapat 15 orang yang diprediksi tanpa waktu gejala, padahal waktu gejala yang sebenarnya ialah waktu gejala lebih dari 24 jam setelah vaksin. Kemudian terdapat error dimana 5 orang yang diprediksi punya waktu gejala lebih dari 24 jam yang mana kelas sebenarnya yaitu tanpa waktu gejala.



Gambar 13. Confusion Matrix pada model Random Forest

Dalam *Confusion matrix* model *Random Forest* bisa kita lihat bahwa terdapat error dimana 5 orang yang diprediksi tanpa waktu gejala, padahal waktu gejala yang sebenarnya ialah lebih dari 24 jam setelah vaksin. Kemudian terdapat error dimana 4 orang yang diprediksi punya waktu gejala lebih dari 24 jam yang mana waktu gejala sebenarnya ke-4 orang tersebut ialah tanpa waktu gejala.



Gambar 14. Confusion Matrix pada model Support Vector Machine

Dalam model *Support Vector Machine* tidak seperti 2 model sebelumnya yang mempunyai error yang berbahaya. Dalam model ini hanya terdapat error yang bisa diperhatikan atau tidak terlalu berbahaya yaitu terdapat 26 orang yang diprediksi tanpa waktu gejala, padahal waktu gejala yang sebenarnya ialah kurang dari 24 jam setelah vaksin. Kemudian terdapat error dimana 18 orang yang diprediksi punya waktu gejala kurang dari 24 jam yang mana waktu gejala sebenarnya ke-18 orang tersebut ialah tanpa waktu gejala.

Berdasarkan dengan *precision, recall, f1-score dan accuracy* dari masing-masing model bisa kita lihat:

Table 3. Evaluasi Model K-Nearest Neighbor

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Precision* (%) | *Recall* (%) | F1-score (%) |
| 0 | 43 | 33 | 38 |
| 1 | 65 | 51 | 57 |
| 2 | 77 | 95 | 85 |
| *Accuracy (%)* | | | 69 |

Table 4. Evaluasi Model Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Precision* (%) | *Recall* (%) | F1-score (%) |
| 0 | 44 | 35 | 39 |
| 1 | 62 | 62 | 62 |
| 2 | 88 | 95 | 92 |
| *Accuracy (%)* | | | 73 |

Table 5. Evaluasi Model Support Vector Machine

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Precision* (%) | *Recall* (%) | F1-score (%) |
| 0 | 63 | 54 | 58 |
| 1 | 70 | 78 | 74 |
| 2 | 99 | 97 | 98 |
| *Accuracy (%)* | | | 82 |

Berdasarkan tabel evaluasi ketiga model. Model *SVM* memiliki *precision* dan *recall* lebih tinggi dibanding model *RF* kemudian disusul model KNN. Saat memprediksi waktu gejala dengan value “2” atau gejala lebih dari 24 jam, menghasilkan f1\_score yang tinggi terutama pada model *SVM,* kemudian disusul model RF dengan urutan kedua dan model KNN dengan urutan ketiga. Kemudian jika dibandingkan f1\_score dari ketiga model, saat melakukan prediksi pada responden yang memiliki waktu gejala kurang dari 24 jam (bernilai “1”) maka model *SVM* memiliki nilai f1\_score yang tinggi, lebih baik dibanding model *RF* di urutan ke-dua dan *KNN* di urutan ke-tiga. Kemudian saat dilakukan prediksi pada responden yang tidak memiliki lama waktu gejala, model *Support Vector Machine* lebih baik dibanding kedua model lainnya karena memiliki f-1 score yang tinggi. Kemudian berdasarkan *accuracy*, model *SVM* memiliki accuracy yang tinggi dibanding model RF dan KNN yang hampir sama.

**KESIMPULAN**

Berdasarkan percobaan ke-tiga model uji yaitu model *K-Nearest Neighbor*, *Random Forest* dan *Support Vector Machine* pada kasus ini, model *SVM* lebih baik dibanding *RF* (di peringkat 2) dan *KNN* (di peringkat 3). Hal ini juga untuk mengurangi resiko agar pemilihan keputusan tidak membawa dampak yang membahayakan. Dalam hal ini terdapat perbedaan yang signifikan dari hasil ke-tiga model yang mana model SVM memiliki *accuracy* jauh lebih tinggi dibanding model RF dan KNN yang mana tingkat *accuracy* nya hampir sama. Saran untuk ke depan sebaiknya perlu dilakukan perlakuan data yang berbeda saat processing atau seleksi untuk setiap model hingga didapat perbedaan yang lebih signifikan lagi dan hasil yang lebih optimal.

**DAFTAR PUSTAKA**

Angreni, I. A. A., Adisasmita, S. A., & Ramli, M. I. (2018). *PENGARUH NILAI K PADA METODE K-NEAREST NEIGHBOR (KNN) TERHADAP TINGKAT AKURASI IDENTIFIKASI KERUSAKAN JALAN*. *7*(2), 63–70.

Ayyad, S. M., Saleh, A. I., & Labib, L. M. (2019). Gene expression cancer classification using modified K-Nearest Neighbors technique. *BioSystems*, *176*(July 2018), 41–51. https://doi.org/10.1016/j.biosystems.2018.12.009

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1*–*3*(April), 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Çiǧşar, B., & Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming*, *2019*. https://doi.org/10.1155/2019/8706505

Devikanniga, D., Ramu, A., & Haldorai, A. (2020). Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. *EAI Endorsed Transactions on Energy Web*, *7*(29), 1–10. https://doi.org/10.4108/EAI.13-7-2018.164177

Farisa, F. C. (2020). Alasan Jokowi Putuskan Vaksin Covid-19 Digratiskan untuk Masyarakat... *Nasional.Kompas.Com*. https://nasional.kompas.com/read/2020/12/17/09084331/alasan-jokowi-putuskan-vaksin-covid-19-digratiskan-untuk-masyarakat?page=all

Gayathri, G. V., & Satapathy, S. C. (2020). A Survey on Techniques for Prediction of Asthma. *Smart Intelligent Computing and Applications*, *159*, 751–758.

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. In A. Stephan (Ed.), *Morgan Kaufmann* (Second, Vol. 12). Diane Cerra. https://doi.org/10.1007/978-3-642-19721-5

Handayani, R. T., Kuntari, S., Darmayanti, A. T., Widiyanto, A., & Atmojo, J. T. (2020). Factors Causing Stress in Health and Community When the Covid-19 Pandemic. *Jurnal Keperawatan Jiwa*, *8*(3), 353. https://doi.org/10.26714/jkj.8.3.2020.353-360

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health*, *8*(July), 1–9. https://doi.org/10.3389/fpubh.2020.00357

Kadali, R. A. K., Janagama, R., Peruru, S., & Malayala, S. V. (2021). Side effects of BNT162b2 mRNA COVID-19 vaccine: A randomized, cross-sectional study with detailed self-reported symptoms from healthcare workers. *International Journal of Infectious Diseases*, *106*(2001), 376–381. https://doi.org/10.1016/j.ijid.2021.04.047

Misra, S., & Li, H. (2020). Chapter 9 - Noninvasive fracture characterization based on the classification of sonic wave travel times. In S. Misra, H. Li, & J. He (Eds.), *Machine Learning for Subsurface Characterization* (pp. 243–287). Gulf Professional. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Parvin, H., Alizadeh, H., & Minati, B. (2010). A Modification on K-Nearest Neighbor Classifier. *Global Journal of Computer Science and Technology*, *10*(14), 37–41.

Ragab, A., Koujok, M. El, Ghezzaz, H., & Amazouz, M. (2021). Chapter 10 - Fault diagnosis in industrial processes based on predictive and descriptive machine learning methods. In J. Ren, W. Shen, Y. Man, & L. Dong (Eds.), *Applications of Artificial Intelligence in Process Systems Engineering* (pp. 207–254). Elsevier. https://doi.org/10.1016/B978-0-12-821092-5.00002-4

*Random Forest - an overview | ScienceDirect Topics*. (2020). Elsevier B.V. https://www.sciencedirect.com/topics/engineering/random-forest

Rokom. (2021, November 5). *Vaksinasi COVID-19 di Indonesia Capai 200 Juta Suntikan*. https://sehatnegeriku.kemkes.go.id/baca/umum/20211105/1038788/vaksinasi-covid-19-di-indonesia-capai-200-juta-suntikan/

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowledge-Based Systems*, *205*, 106270. https://doi.org/10.1016/j.knosys.2020.106270

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, *24*, 91–98. https://doi.org/10.1016/j.jare.2020.03.005

Singh, H. (2019). *Understanding Random Forests*. https://medium.com/@harshdeepsingh\_35448/understanding-random-forests-aa0ccecdbbbb

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028

*Support Vector Machine - an overview | ScienceDirect Topi*. (2018). https://www.sciencedirect.com/topics/chemical-engineering/support-vector-machine/pdf

Xia, Y. (2020). Chapter Eleven - Correlation and association analyses in microbiome study integrating multiomics in health and disease,. In *Progress in Molecular Biology and Translational Science* (Vol. 171, pp. 309–491). Elsevier Inc. https://doi.org/10.1016/bs.pmbts.2020.04.003.

Zhao, H., Wang, R., Lei, Y., Liao, W.-H., Cao, H., & Cao, J. (2021). Severity Level Diagnosis of Parkinson’s Disease by Ensemble K-Nearest Neighbor Under Imbalanced Data. *Expert Systems with Applications*, *189*(September 2021), 116113. https://doi.org/10.1016/j.eswa.2021.116113

Zhao, Y. (2015). R and Data Mining: Examples and Case Studies Messages from the Author. In *Elsevier* (Issue December 2012).