KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)
Terakreditasi Nomor 204/E/KPT/2022 | Vol. 6, No. 2, April (2025), pp. 537-546

# Penerapan Model Machine Learning (KNN, Random Forest, dan Naive Bayes) untuk Menganalisis Pengaruh Kualitas Udara terhadap COVID-19: Studi Normalisasi pada Data COVID-19

*Irene Paskalita Ponamon[1], Alz Danny Wowor[2]*
*[1,2]Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana, Salatiga, Indonesia*
*E-mail: [1]672021244@student.uksw.edu, [2]alzdanny.wowor@uksw.edu*

***Abstract***

*The study was intended to analyze the effects of air quality on covid-19 in jagakconclusion, south Jakarta, using the machine learning classification algorithm: random forest, naive bayes, and k-herd prediction (KNN). The data used consisted of about 6,713 records that included case files of covid-19 and data of air quality such as pm10, so2, co, o3, and no2 obtained from the official air and source monitoring stations. The research process involves the stage of data preparation, normalization, exploration, and training and model evaluation. The results showed that random forest algorithm with 100 trees reached its highest level of accuracy, around 97.6% of the data had been modernized, and was consistently the best performance compared to naive bayes and KNN. Furthermore, analysis suggests that the normal of data significantly increases model performance. The conclusion from this study suggests that air quality affects the spread and severity of covid-19 in the region, and that the random forest model is the best option for prediction and analysis of the environmental impact on the covid-19 case. The results of this study are expected to be a reference to more effective development of health and environmental policies.*

***Keywords*** *: Air quality, Covid-19, Machine learning, Normalisasi, Random Forest.*

***Abstrak***

*Penelitian ini bertujuan untuk menganalisis pengaruh kualitas udara terhadap COVID-19 di Kecamatan Jagakarsa, Jakarta Selatan, dengan menggunakan algoritma klasifikasi machine learning yaitu: Random Forest, Naïve Bayes, dan K-Nearest Neighbors (KNN). Data yang digunakan terdiri dari sekitar 6.713 records yang mencakup data kasus COVID-19 dan data kualitas udara seperti PM10, SO2, CO, O3, dan NO2 yang diperoleh dari stasiun pemantauan udara dan sumber resmi terkait. Proses penelitian meliputi tahap persiapan data, normalisasi, eksplorasi, serta pelatihan dan evaluasi model. Hasil menunjukkan bahwa algoritma Random Forest dengan 100 pohon mencapai tingkat akurasi tertinggi, sekitar 97.6% pada data yang telah dinormalisasi, dan secara konsisten menunjukkan performa terbaik dibandingkan Naïve Bayes dan KNN. Selain itu, analisis menunjukkan bahwa normalisasi data secara signifikan meningkatkan performa model. Kesimpulan dari studi ini menegaskan bahwa kualitas udara berpengaruh terhadap penyebaran dan tingkat keparahan COVID-19 di wilayah tersebut, dan bahwa model Random Forest merupakan pilihan terbaik untuk prediksi dan analisis dampak lingkungan terhadap kasus COVID-19. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan kebijakan kesehatan dan lingkungan yang lebih efektif.*

***Kata kunci:*** *Kualitas Udara, COVID-19, Machine Learning, Normalization, Random Forest.*

ISSN: 2720-992X

## 1. Pendahuluan

COVID-19, atau penyakit coronavirus 2019, pertama kali ditemukan pada akhir tahun 2019, tepatnya pada bulan Desember di Kota Wuhan, Provinsi Huebei, China. Sejak itu, penyakit ini kemudian menyebar ke hampir seluruh dunia. COVID-19 adalah penyakit menular yang disebabkan oleh Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2)[1]. Sumber penularan virus COVID-19 belum diketahui, tetapi virus ini termasuk dalam kategori zoonosis, yang berarti itu ditularkan antara hewan dan manusia[2]. Banyak hewan liar membawa patogen dan penyebab penyakit menular. Sumber utama dari severe acute respiratory syndrome (SARS) dan Middle East respiratory syndrome (MERS) adalah corona virus yang ditemukan pada kelelawar, tikus bambu, unta, dan musang.(PDPI, 2020)[3]. Jenis corona virus baru yang menular ke manusia disebut Severe Acute Respiratory Syndrome Coronavirus 2 (SARSCoV-2). Virus ini juga dikenal sebagai virus Corona. Bayi, anak-anak, orang dewasa, orang tua, ibu hamil, dan ibu menyusui semua dapat terinfeksi virus ini[4]. Virus ini juga memiliki tingkat penularan yang tinggi dan dapat dengan mudah menyebar dari satu individu ke individu lainnya melalui berbagai bentuk kontak dengan penderita[5]. Meskipun gejalanya hampir sama dengan flu namun virus corona berkembang lebih cepat, menyebabkan infeksi yang lebih parah dan bisa mengakibatkan gagal organ[6].

Terdapat banyak faktor yang bisa menyebabkan penyebaran virus covid-19 ini sendiri, salah satunya adalah kualitas udara. Manusia membutuhkan udara bersih untuk proses pernapasan mereka. Kualitas udara buruk akan berdampak pada saluran pernapasan setiap orang di suatu tempat. Penduduk yang berada di daerah dengan kualitas udara buruk tentunya juga mempunyai risiko buruk yang lebih tinggi bila terkena COVID-19[7]. Becchetti et al. (2020) menjelaskan bahwa polusi udara dapat menjadi penggerak COVID-19 karena dua alasan. Pertama, orang yang tinggal di daerah dengan kualitas udara yang buruk lebih cenderung memiliki masalah paru-paru, yang membuat mereka lebih rentan terhadap penyakit pernapasan seperti COVID-19. Faktor kedua adalah bahwa zat partikulat dapat berfungsi sebagai pembawa virus yang tetap berada di udara[8].

Kualitas udara di wilayah Jakarta disebut-sebut juga sebagai salah satu yang paling buruk di dunia, dan bahkan pernah berada di tingkat teratas. Kecamatan Jagakarsa adalah salah satu kecamatan yang berada di wilayah Kota Administrasi Jakarta Selatan. Berdasarkan Surat Keputusan Gubernur DKI Jakarta dengan nomor 1251 Tahun 1986, 435 Tahun 1966, dan 1986 Tahun 2000, luas wilayah Kecamatan Jagakarsa ditetapkan sebesar 25,01 km², yang terbagi ke dalam 54 RW dan 545 RT. Jumlah penduduk Jagakarsa pada tahun 2020 di Jakarta Selatan adalah 222.004 jiwa. Untuk kelurahan jagakarsa ini sendiri jumlah kasus positif covid di seluruh Kecamatan Jagakarsa mencapai 3.555 kasus.

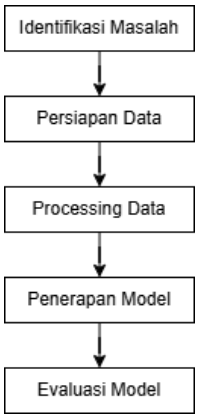
Menurut penelitian yang dilakukan sebelumnya, Monica Dias Febriyanti melakukan penelitian dengan judul "Identifikasi Pengaruh Kualitas Udara Terhadap Kondisi Pasien Covid-19 dengan Algoritma Naive Bayes" hasil penelitian menunjukkan bahwa Naive Bayes memperoleh hasil akurasi 82,73% dalam mengklasifikasikan seberapa besar pengaruh kualitas udara terhadap pasien COVID-19 di Jakarta[9]. Pada penelitian yang dilakukan oleh Suci Anggraini yang membahas tentang "Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan *Machine Learning*" di temukan bahwa algoritma Neural Network dan Naive Bayes memiliki nilai *accuracy* tertinggi[10]. Penelitian yang juga dilakukan oleh Fauzan Azimah dengan judul "Klasifikasi Deteksi Gejala Awal Covid-19 Dengan Metode Logistic Regression, Random Forest Classifier, dan Support Vector Machine" menemukan bahwa algoritma Logistic Regression memiliki performa terbaik dalam mendeteksi gejala awal Covid-19[11]. Penelitian lain yang dilakukan oleh Agung Supoyo yang membahas tentang "Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid di DIY" ditemukan hasil yang menunjukkan

bahwa *performance* Random Forest memiliki *accuracy* terbaik jika dibandingkan dengan KNN dan Deep Learning Neural Network[12].

Normalisasi data pada penelitian ini diterapkan pada dataset COVID-19 karena dataset yang diperoleh masi belum cukup baik sehingga dilakukan normalisasi agar kualitas data lebih bagus. Penelitian yang dilakukan oleh penulis sendiri menggunakan modelling Random Forest untuk menyelesaikan masalah. Random Forest digunakan untuk penelitian ini karena mampu menangani data kompleks dan multivariat yang dapat dilihat dalam dataset yang terdiri dari banyak variabel misalnya variabel kualitas udara meliputi (PM10, CO, NO2, SO2, O3) dan data COVID-19 yang fluktuatif.

## 2. Metodologi Penelitian

Penelitian akan dilakukan melalui beberapa tahap yaitu identifikasi masalah, persiapan data, pre-processing data, perancangan model, dan evaluasi model, yang diilustrasikan dalam Gambar 1.




**Gambar 1.** Tahapan Penelitian

Tahap pertama adalah identifikasi masalah, dengan tujuan untuk menganalisis dan memahami hubungan antara kualitas udara dengan angka penyebaran COVID-19. Penelitian ini bertujuan untuk mengidentifikasi bagaimana faktor-faktor kualitas udara, seperti pm10, so2, co, o3, dan no2, dapat memengaruhi angka covid 19 dilihat dari data jumlah orang yang meninggal, jumlah pasien rawat inap, jumlah orang yang positif terkena virus, suspek atau orang yang dicurigai terkena covid, serta jumlah orang yang sembuh dari virus ini.

Tahap kedua adalah persiapan data, dataset kualitas udara yang digunakan dalam penelitian ini didapat dari Stasiun Pemantauan Kualitas Udara (SPKU) Jagakarsa yang terletak di Jakarta yang berupa file *csv*. Pada *dataset* ini, memuat informasi kualitas udara pada sebuah stasiun jagakarsa berupa hasil pengukuran Polusi Udara (PM10), *Sulfur Dioksida* (SO2), *Karbon Monoksida* (CO), *Ozon* (O3), dan *Natrium Dioksida* (NO2) pada bulan juli hingga bulan november 2020. Dataset ini memuat total sebanyak 6.713 *records* data.

Tahap ketiga adalah *processing* data, pada tahap ini dilakukan pengecekan *Missing Value* dengan tujuan untuk menilai seberapa banyak data yang memiliki kesalahan atau ketidaksempurnaan yang dapat memengaruhi hasil analisis. Kemudian langkah selanjutnya yang dilakukan adalah normalisasi data. Metode normalisasi data sendiri dilakukan dengan tujuan untuk membuat beberapa variabel memiliki rentang nilai yang sama, tidak terlalu besar maupun terlalu kecil. Ini membuat analisis stastistik lebih mudah. Pada penelitian ini peneliti menggunakan normalisasi Min-Max dengan tujuan untuk memperbaiki kualitas data agar lebih baik. Secara lebih spesifik, analisis korelasi dilakukan untuk melihat hubungan linier antara variabel-variabel yang tersedia dalam data, sehingga langkah ini membantu untuk menemukan pola dan tren, dan memahami

bagaimana perubahan dalam satu variabel berhubungan dengan perubahan dalam valeur[13].

Tahap keempat adalah perancangan model, pada tahap ini peneliti membuat pemodelan menggunakan algoritma *machine learning* untuk menganalisis data dalam jumlah yang besar dan kompleks sperti data kualitas udara dan covid-19. Modelling ya di pakai adalah *Random Forest*, *Naive Bayes*, dan *K-Nearest Neighbor*.

Tahap kelima adalah evaluasi model, pada tahap ini penulis membandingkan setiap modelling dengan tujuan untuk melihat model mana yang efektif digunakan di wilayah Jagakarsa Jakarta dengan melihat tingkat *accuracy* dari setiap pemodelan dan memilih akurasi tertinggi yang kemudian digunakan untuk menyelesaikan masalah.

## 3. Hasil dan Pembahasan

### 3.1. *Processing* Data

Langkah awal dalam pengolahan data ini adalah melakukan pengecekan *Missing Value*. *Missing Value* adalah nilai yang hilang dari data yang mungkin terjadi karena proses pengambilan data yang tidak sempurna . Untuk mengatasi nilai yang hilang, kita dapat menghapus data yang hilang, melakukan estimasi parameter seperti menggunakan algoritma Estimasi-Maximisasi, atau imputasi, yaitu menghitung nilai pengganti data yang hilang yang tersebar[14].

Data yang di dapat peneliti dalam penelitian ini memiliki rentang nilai yang jauh sehingga penulis melakukan normalisasi data dengan tujuan untuk membentuk data dalam posisi nilai dengan rentang yang sama. Dengan demikian, normalisasi deviasi dari outlier memastikan distribusi data tetap normal[15]. Skala yang dihasilkan berada di antara 0 hingga 1.

$$x' = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

Rumus normalisasi Min-Max ditunjukkan pada Persamaan (1). Di mana $xi$ adalah data yang dinormalisasi, $x'$ menunjukkan hasil normalisasi, $min(x)$ merupakan data terkecil suatu fitur, dan $max(x)$ adalah data terbesar suatu fitur[16]. Sebelum dilakukan normalisasi *accuracy* yang dihasilkan dari setiap pemodelan sangatlah rendah namun setelah dilakukan normalisasi hasil dari *accuracy* dari setiap pemodelan mengalami kenaikan.

**Tabel 1.** Hasil Accuracy Modelling Sebelum Normalisasi

| Modelling | 70%:30% | 75%:25% | 80%:20% | 85%:15% | 90%:10% |
|---|---|---|---|---|---|
| **Random Forest** | 0.000000 | 0.000000 | 0.035714 | 0.047619 | 0.071429 |
| **Naïve Bayes** | 0.023810 | 0.028571 | 0.035714 | 0.000000 | 0.000000 |
| **KNN** | 0.023810 | 0.028571 | 0.071429 | 0.047619 | 0.071429 |

Pada 1 hasil *accuracy* tertinggi dari *modelling* Random Forest adalah 0.071429 pada rasio 90%:10% hasil *accuracy* terendah-nya 0.000000 pada rasio 70%:30% dan rasio 75%:25%, *accuracy* tertinggi dari *modelling* Naïve Bayes adalah 0.035714 pada rasio 80%:20% hasil *accuracy* terendah-nya 0 pada rasio 85%:15% dan rasio 90%:10%. Untuk *modelling* KNN hasil *accuracy* tertinggi adalah 0.071429 pada rasio 80%:20% dan rasio 90%:10%, hasil *accuracy* terenda-nya sendiri adalah 0.023810 pada rasio 70%:30%.

**Tabel 2.** Hasil Accuracy Modelling Setelah Normalisasi

| Modelling | 70%:30% | 75%:25% | 80%:20% | 85%:15% | 90%:10% |
|---|---|---|---|---|---|
| **Random Forest** | 0.976190 | 0.914286 | 0.964286 | 0.904762 | 0.857143 |
| **Naïve Bayes** | 0.904762 | 0.885714 | 0.857143 | 0.809524 | 0.714286 |
| **KNN** | 0.690476 | 0.685714 | 0.678571 | 0.666667 | 0.714286 |

Pada Tabel 2sendiri hasil *accuracy* teringgi dari *modelling* Random Forest adalah 0.976190 pada rasio 70%:30% hasil *accuracy* terendah-nya 0.857143 pada rasio 90%:10%, hasil *accuracy* tertinggi untuk *modelling* Naïve Bayes adalah 0.904762 pada rasio 70%:30% hasil *accuracy* terendah-nya 0.714286 pada rasio 90%:10%, dan hasil *accuracy* tertinggi untuk *modelling* KNN adalah 0.714286 pada rasio 90%:10% dan hasil *accuracy* terendah-nya adalah 0.666667 pada rasio 85%:15% . Setelah dilakukan normalisasi hasil *accuracy* dari setiap *modelling* pada semua kategori mengalami peningkatan.

### 3.2. Penerapan Modelling

#### a. Implementasi Algoritma Random Forest

Pada bagian ini langkah yang dilakukan adalah split data menjadi data latih dan data uji kemudian membuat model Random Forest dengan 100 pohon Keputusan (*estimators)* melatih model dengan data latih, membuat prediksi menggunakan data uji, dan menghitung akurasi dari model. Random Forest membuat banyak pohon secara acak dengan tujuan agar hasil prediksi lebih stabil dan kemudian menggabungkan hasilnya untuk membuat keputusan akhir.

**Tabel 3.** Hasil Accuracy Random Forest

| No | **Atribut** | 70%:30% | 75%:25% | 80%:20% | 85%:15% | 90%:10% |
|---|---|---|---|---|---|---|
| 1 | **Meninggal** | 0.998381 | 0.971232 | 0.700475 | 0.914461 | 0.857143 |
| 2 | **Rawat Inap** | 0.500000 | 0.542857 | 0.535714 | 0.428571 | 0.571429 |
| 3 | **Positif** | 0.976190 | 0.914286 | 0.964286 | 0.904762 | 0.857143 |
| 4 | **Suspek** | 0.976190 | 0.942857 | 0.690476 | 0.904762 | 0.857143 |
| 5 | **Sembuh** | 0.952381 | 0.942857 | 0.928571 | 0.904762 | 0.857143 |

Merupakan hasil dari perhitungan algoritma *Random Forest* dengan menggunakan split data 70%:30%, 75%:25%, 80%:20%, 85%:15%, dan 90%:10% di mana atribut rawat inap mendapatkan nilai *accuracy* terendah dibandingkan dengan atribut-atribut lainnya, sedangkan meninggal memiliki *accuracy* tertinggi jika di bandingkan dengan atribut-atribut lainnya. Untuk atribut positif,suspek, dan sembuh juga menjadi factor kuat dari pengaruh kualitas udara pada pasien *covid-19* yang dapat dilihat dari tabel di atas.

#### b. Implementasi Algoritma Naïve Bayes

Menyajikan hasil kerja dari Python untuk perhitungan data dari kualitas udara dan data dari pasien *covid-19* menggunkaan algoritma *Naïve Bayes*

**Tabel 4.** Hasil Accuracy Naive Bayes

| No | **Atribut** | 70%:30% | 75%:25% | 80%:20% | 85%:15% | 90%:10% |
|---|---|---|---|---|---|---|
| 1 | **Meninggal** | 0.904762 | 0.871667 | 0.847183 | 0.795523 | 0.704285 |
| 2 | **Rawat Inap** | 0.452381 | 0.428571 | 0.535714 | 0.523810 | 0.571429 |
| 3 | **Positif** | 0.904762 | 0.885714 | 0.857143 | 0.809524 | 0.714286 |
| 4 | **Suspek** | 0.904762 | 0.885714 | 0.857143 | 0.809524 | 0.714286 |
| 5 | **Sembuh** | 0.904762 | 0.885714 | 0.857143 | 0.809524 | 0.714286 |

Pada Tabel 4 merupakan hasil dari perhitungan algoritma Naïve Bayes dengan menggunakan split data 70%:30%, 75%:25%, 80%:20%, 85%:15%, dan 90%:10%. Pada tabel di atas atribut yang memiliki *accuracy* paling rendah adalah rawat inap sedangkan untuk atribut lain seperti "Meninggal", "Positif", "Suspek", dan "Sembuh" menunjukkan performa model yang baik dan konsisten. Jika dilihat pada tabel di atas, semakin sedikit data latih (misalnya 90%:10%), akurasi model cenderung menurun.

**c. Implementasi Algoritma KNN**

Pada KNN peneliti membuat *split* data sama seperti algoritma sebelum-sebelumnya yaitu data 70%:30%, 75%:25%, 80%:20%, 85%:15%, dan 90%:10%.

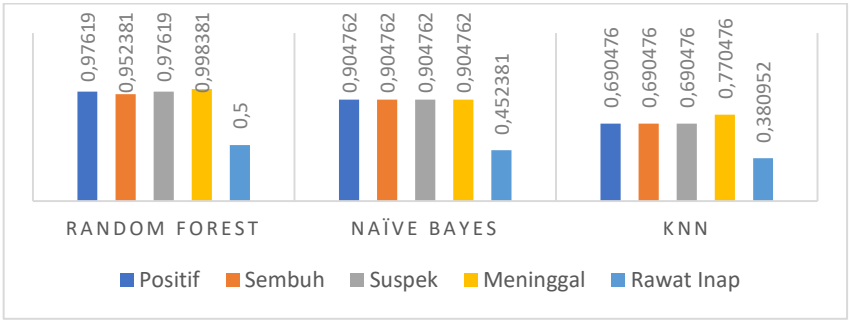
**Tabel 5.** Tabel Accuracy KNN

| No | Atribut | 70%:30% | 75%:25% | 80%:20% | 85%:15% | 90%:10% |
|---|---|---|---|---|---|---|
| 1 | Meninggal | 0.770476 | 0.681444 | 0.720492 | 0.546617 | 0.694125 |
| 2 | Rawat Inap | 0.380952 | 0.371429 | 0.392857 | 0.380952 | 0.285714 |
| 3 | Positif | 0.690476 | 0.685714 | 0.678571 | 0.666667 | 0.714286 |
| 4 | Suspek | 0.690476 | 0.690476 | 0.690476 | 0.666667 | 0.714286 |
| 5 | Sembuh | 0.690476 | 0.685714 | 0.678571 | 0.666667 | 0.714286 |

Dalam Tabel 5 akurasi tertinggi untuk kategori meninggal 0.770476 pada rasio 70%:30% dan akurasi terendah 0.546617 pada rasio 85%:15%, akurasi tertinggi kategori rawat inap 0.392857 pada rasio 80%:20% akurasi terendah 0.285714 pada rasio 90%:10%, akurasi tertinggi kategori positif 0.714286 pada rasio 90%:10% akurasi terendah 0.666667 pada rasio 85%:15%, akurasi tertinggi kategori suspek 0.714286 pada rasio 90%:10% akurasi terendah 0.666667 pada rasio 85%:15%, dan akurasi tertinggi untuk kategori sembuh adalah 0.714286 pada rasio 90%:10% akurasi terendah 0.666667 pada rasio 85%:15%.

**3.3. Komparasi Algoritma**

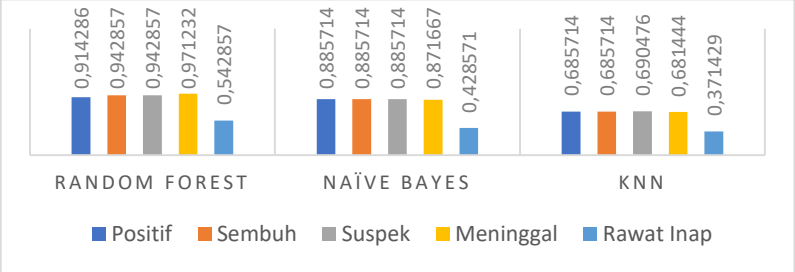
Split data adalah membedakan kumpulan data menjadi dua atau, biasanya menjadi: Data latih, juga dikenal sebagai data pelatihan, digunakan untuk membangun atau "melatih" model. Data uji, di sisi lain, digunakan untuk menguji atau menilai kinerja model yang telah dilatih.




**Gambar 2.** Grafik Algoritma Perbandingan Data 70% : 30%

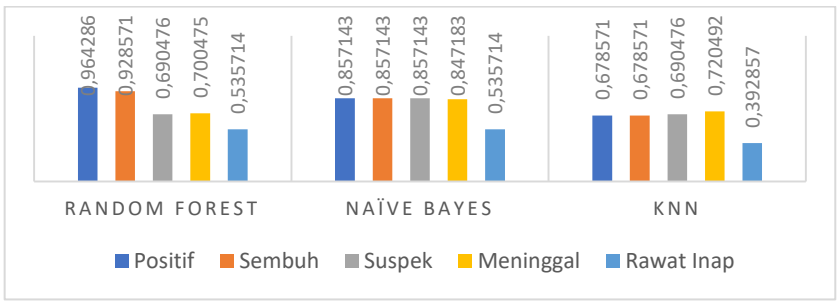
Gambar 2 menunjukkan hasil *accuracy* kategori positif menggunakan Random Forest memiliki selisih 0,071428 dengan hasil *accuracy* kategori positif pada Naïve Bayes. Sedangkan hasil *accuracy* kategori positif menggunakan Naïve Bayes memiliki selisih 0,214286 dengan kategori positif menggunakan KNN. Untuk kategori sembuh menggunakan Random Forest sendiri hasil *accuracy*-nya memiliki selisih 0,047619 dengan kategori sembuh pada Naïve Bayes. Hasil *accuracy* kategori sembuh menggunakan Naïve Bayes memiliki selisih 0,214286 dengan kategori sembuh menggunakan KNN. Hasil *accuracy* dari kategori suspek memiliki selisih 0,071428 dengan kategori suspek meggunakan Naïve Bayes. Untuk hasil *accuracy* kategori suspek menggunakan Naïve Bayes memiliki selisih 0,214286 dengan kategori suspek menggunakan KNN. Hasil *accuracy* kategori meninggal menggunakan Random Forest memiliki selisih 0,093619 dengan hasil *accuracy* kategori meninggal pada Naïve Bayes. Hasil *accuracy* kategori meninggal menggunakan Naïve Bayes memiliki selisih 0,134286 dengan hasil *accuracy* kategori meninggal pada KNN. Sedangkan untuk hasil *accuracy*

kategori Rawat Inap menggunakan Random Forest memiliki selisih 0,047619 dengan kategori rawat inap pada Naïve Bayes. Dan hasil *accuracy* kategori Rawat Inap menggunakan Naïve Bayes memiliki selisih 0,071429 jika di bandingkan dengan hasil *accuracy* kategori Rawat Inap menggunakan KNN. Meskipun selisih dari setiap kategori dengan menggunakan 3 pemodelan memiliki nilai yang berbeda namun modelling yang memiliki hasil *accuracy* terbaik untuk setiap kategori adalah Random Forest.




**Gambar 3.** Grafik Algoritma Perbandingan Data 75% : 25%

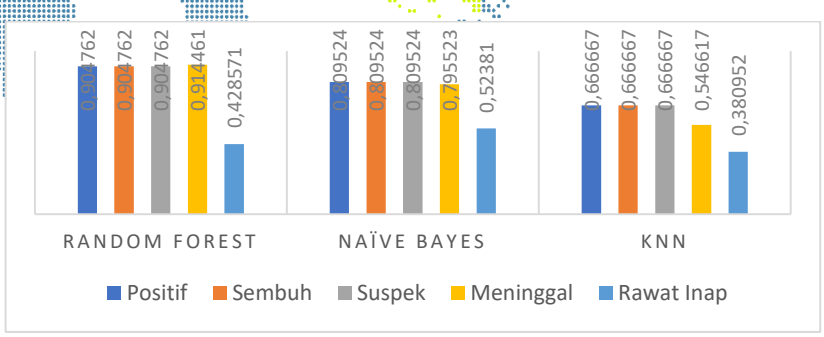
Gambar 3 Grafik menunjukkan hasil accuracy kategori positif pada Random Foreset memiliki selisih 0,028572 dengan positif pada Naïve Bayes. Untuk hasil accuracy positif pada Naïve Bayes sendiri memiliki selisih 0,2 dengan positif pada KNN. Hasil accuracy kategori sembuh dan suspek menggunakan Random Forest memiliki selisih 0,057143 jika di bandingkan dengan sembuh dan suspek pada Naïve Bayes. Untuk hasil accuracy sembuh pada Naïve Bayes memiliki selisih 0,2 dengan sembuh pada KNN dan hasil accuracy suspek pada Naïve Bayes memiliki selisih 0,195238 dengan suspek pada KNN. Hasil accuracy meninggal pada Random Forest memiliki selisih 0,099565 dengan meninggal pada Naïve Bayes. Dan untuk kategori meninggal pada Naïve Bayes memiliki selisih 0,190223 dengan kategori meninggal pada KNN. Hasil accuracy rawat inap pada Random Forest memiliki selisih 0,114286 dengan rawat inap pada Naïve Bayes. Dan hasil accuracy rawat inap pada Naïve Bayes memiliki selisih 0,057142 dengan rawat inap pada KNN.




**Gambar 4.** Grafik Algoritma Perbandingan Data 80%:20%

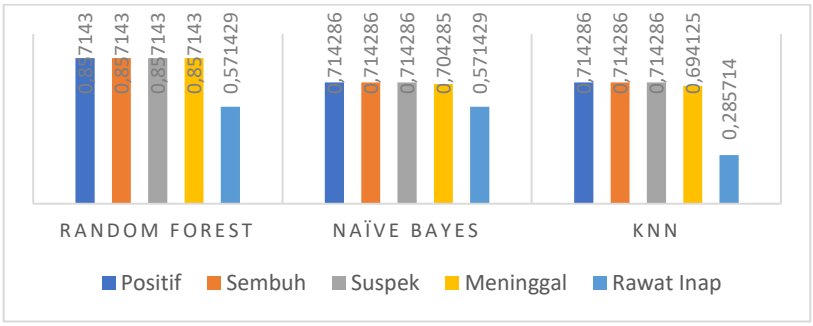
Gambar 4 menampilkan hasil *accuracy* positif pada Random Forest memiliki selisih 0,107143 dengan positif pada Naïve Bayes. Untuk hasil *accuracy* positif pada Naïve Bayes memiliki selisih 0,178572 dengan positif pada KNN. Sembuh pada Random Forest sendiri memiliki selisih *accuracy* 0,071428 dengan sembuh pada Naïve Bayes. *Accuracy* sembuh pada Naïve Bayes memiliki selisih 0,178572 dengan sembuh pada KNN. Hasil *accuracy* suspek pada Naïve bayes memiliki selisih 0,166667 dengan suspek pada Random Forest dan suspek pada KNN. Untuk meninggal pada Naïve Bayes memiliki selisih 0,146708 dengan meninggal pada Random Forest. Hasil *accuracy* meninggal pada Naïve Bayes memiliki selisih 0,126691 dengan meninggal pada KNN. Hasil *accuracy*

rawat inap pada Random Forest dan Naïve bayes memiliki selisih 0,142857 dengan rawat inap pada KNN.




**Gambar 5.** Grafik Algoritma Perbandingan Data 85%:15%

Gambar 5 menampilkan hasil *accuracy* dari kategori positif, sembuh, dan suspek pada pemodelan Random Forest memiliki selisih 0,095238 dengan *accuracy* kategori positif, sembuh, dan suspek pada Naïve Bayes. Sementara hasil *accuracy* pada kategori positif, sembuh, dan suspek pada pemodelan Naïve Bayes memiliki selisih 0,142857 dengan hasil *accuracy* kategori positif, sembuh, dan suspek menggunkan pemodelan KNN. Untuk kategori meninggal pada Random Forest sendiri hasil *accuracy*-nya memiliki selisih 0,118938 dengan hasil *accuracy* meninggal pada Naïve Bayes. Sementara hasil *accuracy* meninggal menggunakan pemodelan Naïve Bayes memiliki selisih 0,248906 dengan meninggal pada pemodelan KNN. Hasil *accuracy* rawat inap menggunakan Naïve Bayes memiliki selisih 0,095239 dengan Random Forest. Dan hasil *accuracy* rawat inap menggunakan pemodellan Naïve Bayes memiliki selisih 0,142858 dengan rawat inap menggunakan pemodellan KNN.




**Gambar 6.** Grafik Algoritma Perbandingan Data 90%:10%

Gambar 6 menunjukkan hasil *accuracy* kategori positif, sembuh, dan suspek pada pemodelan Random Forest memiliki selisih 0,142857 dengan hasil *accuracy* kategori positif, sembuh dan suspek pada pemodelan Naïve Bayes. Sementara hasil *accuracy* kategori positif, sembuh, dan suspek menggunakan pemodellan Naïve Bayes memiliki selisih 0% dengan hasil *accuracy* kategori positif, sembuh, dan suspek dengan pemodellan KNN. Dapat dilihat kategori positif, sembuh dan suspek menggunkan pemodelan Naïve Bayes dan KNN memiliki nilai yang identik sama hal ini bisa saja disebabkan oleh kondisi, jika data dibagi 90:10, hanya sepuluh persen dari total data digunakan sebagai data uji. Jika jumlah data uji kecil dan distribusinya merata sehingga mengakibatka setiap kelas memiliki jumlah *accuracy* yang sama (positif, sembuh, dan suspek), maka kedua model dapat memberikan hasil prediksi *accuracy* yang sama untuk kelas tertentu. Hasil *accuracy* kategori meninggal pada Random Forest memiliki selisih 0,152858 dengan kategori meninggal pada Naïve Bayes. Untuk hasil *accuracy* kategori meninggal pada Naïve Bayes memiliki selisih 0,01016 dengan kategori meninggal pada

KNN. Untuk hasil *accuracy* rawat inap pada pemodellan Random Forest dan Naïve Bayes memiliki selisih 0,285715 dengan hasil *accuracy* rawat inap menggunkan pemodellan KNN.

### 3.4. Evaluasi Model

Berdasarkan lima grafik akurasi dari berbagai skenario pembagian data pada komparasi algoritma dengan split data: 70-30, 75-25, 80-20, 85-15, 90-10 dan tiga algoritma klasifikasi (Random Forest, Naïve Bayes, dan KNN). Maka dapat dapat dilakukan evaluasi kepada setiap algoritma klasifikasi pemodelan. Algoritma pertama yaitu Random Foreset memiliki kelebihan yang sangat akurat dalam hampir semua kategori, termasuk Positif, Sembuh, Suspek, dan Meninggal. sangat baik dalam pembagian data besar untuk training (70:30 dan 75:25), dengan nilai akurasi hampir atau di atas 0.95. Namun memiliki kekurangan pada kinerja Rawat Inap yang jauh lebih buruk, dengan akurasi hanya 0,5–0.57. Ini menunjukkan kelemahan dalam menangani kelas tersebut, hal ini dapat terjadi karena distribusi data yang tidak seimbang. Untuk Algoritma kedua yaitu Naïve Bayes memiliki kelebihan hampir di semua kategori memiliki akurasi yang cukup baik, dengan rata-rata 0.80–0.90 untuk sebagian besar kelas. Sederhana dan efektif, cocok untuk data yang tidak terlalu kompleks dan bersih. Kekurangan algoritma ini sendiri adalah memiliki nilai akurasi rawat inap yang rendah (sekitar 0.42–0.57), seperti Random Forest hal ini terjadi karena asumsi independensi antar variabel yang tidak dapat menangani hubungan antar fitur dengan baik. Dan algoritma klasifikasi yang terakhir adalah KNN memiliki kelebihan yaitu performanya baik di beberapa kelas seperti pada kelas meninggal, terutama di split data 70:30 (akurasi 0.77). Klasifikasi ini cenderung memberikan hasil yang seimbang jika jumlah data tidak terlalu besar. Namun memiliki kekurangan yaitu hasil *accuracy*-nya lebih rendah jika di bandingkan dengan dua model lainnya. Kinerja nya buruk di kelas rawat inap, khususnya pada split data 90:10 (akurasi hanya 0.28). Serta sensitif terhadap outlier dan sangat bergantung pada nilai k dan skala data yang dipilih.

## 4. Kesimpulan

Dari perhitungan hasil *accuracy* implementasi dataset covid-19 dan kualitas udara pada sebuah stasiun pemantauan kualitas udara Jagakarsa yang terletak di Jakarta, algoritma klasifikasi yang paling sesuai digunakan di daerah ini adalah Random Foreset karena klasifikasi ini sendiri memiliki performa yang cukup baik dalam memprediksi pengaruh kualitas udara terhadap covid-19 dilihat dari *accuracy* nya yang sangat tinggi pada hampir semua kategori. Melalui hasil *accuracy* yang dihasilkan oleh algoritma klasifikasi yang sudah di terapkan, dapat dilihat juga bahwa kualitas udara sangat mempengaruhi angka covid-19 di Jagakarsa baik itu untuk jumlah pasien positif, jumlah pasien yang sembuh, jumlah pasien yang meninggal, jumlah pasien rawat inap maupun jumlah orang yang dicurigai terkena virus (suspek).

## Daftar Pustaka

[1] M. Mudzakkir, N. Risnasari, M. F. E. Nugraha, And S. A. Mawadha, "Upaya Pencegahan Penularan Covid-19 Pada Masyarakat Kab. Kediri," *Kontribusi J. Penelit. Dan Pengabdi. Kpd. Masy.*, Vol. 2, No. 1, Pp. 56–65, 2021, Doi: 10.53624/Kontribusi.V2i1.85.

[2] A. Sari And I. Budiono, "Faktor Yang Berhubungan Dengan Perilaku Pencegahan Tbc Paru," *Indones. J. Public Heal. Nutr.*, Vol. 1, No. 1, Pp. 50–61, 2021, [Online]. Available: Http://Journal.Unnes.Ac.Id/Sju/Index.Php/Ijphn

[3] A. Syauqi, "Jalan Panjang Covid19," *J. Keuang. Dan Perbank. Syariah*, Vol. 1, No. 1, Pp. 1–19, 2020, Doi: 10.24260/Jkubs.V1i1.115.

[4] L. Aplikasi And B. Andorid, "M Engenal C Ovid -19 Dan C Egah P Enyebarannya

D Engan ‘ P Eduli L Indungi ’ A Plikasi B Erbasis A Ndorid,” No. April, 2020.

[5] B. Fish, “No 主観的健康感を中心とした在宅高齢者における健康関連指標に関する共分散構造分析title,” Vol. 2507, No. February, Pp. 1–9, 2020.

[6] L. Amalia, I. Irwan, And F. Hiola, “Analisis Gejala Klinis Dan Peningkatan Kekebalan Tubuh Untuk Mencegah Penyakit Covid-19,” *Jambura J. Heal. Sci. Res.*, Vol. 2, No. 2, Pp. 71–76, 2020, Doi: 10.35971/Jjhsr.V2i2.6134.

[7] M. D. Febriyanti, A. D. Wowor, And M. A. I. Pakereng, “Identifikasi Pengaruh Kualitas Udara Terhadap Kondisi Pasien Covid-19 Dengan 1 Algoritma Naive Bayes,” *Jiko (Jurnal Inform. Dan Komputer)*, Vol. 8, No. 2, P. 222, 2024, Doi: 10.26798/Jiko.V8i2.867.

[8] R. Zulkarnain And K. D. Ramadani, “Air Quality And The Potency Of Covid-19 Transmission In Java,” *Semin. Nas. Off. Stat. 2020*, No. 2, Pp. 23–33, 2020, [Online]. Available: Https://Prosiding.Stis.Ac.Id/Index.Php/Semnasoffstat/Article/Download/398/88/

[9] S. Adi And A. Wintarti, “Komparasi Metode Support Vector Machine (Svm), K-Nearest Neighbors (Knn), Dan Random Forest (Rf) Untuk Prediksi Penyakit Gagal Jantung,” *Mathunesa J. Ilm. Mat.*, Vol. 10, No. 2, Pp. 258–268, 2022, Doi: 10.26740/Mathunesa.V10n2.P258-268.

[10] S. Anggraini, M. Akbar, A. Wijaya, H. Syaputra, And M. Sobri, “Klasifikasi Gejala Penyakit Coronavirus Disease 19 (Covid-19) Menggunakan Machine Learning,” *J. Softw. Eng. Ampera*, Vol. 2, No. 1, Pp. 57–68, 2021, Doi: 10.51519/Journalsea.V2i1.105.

[11] F. Azimah And K. Rizky Nova Wardani, “Sistem Pendeteksi Gejala Awal Covid-19 Dengan Penggunaan Metode Al Project Cycle,” *J. Locus Penelit. Dan Pengabdi.*, Vol. 1, No. 6, Pp. 405–418, 2022, Doi: 10.36418/Locus.V1i6.135.

[12] A. Supoyo And P. T. Prasetyaningrum, “Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di Diy,” *Bianglala Inform.*, Vol. 10, No. 1, Pp. 21–29, 2022, Doi: 10.31294/Bi.V10i1.11890.

[13] N. Syahira And D. B. Arianto, “Prediksi Tingkat Kualitas Udara Dengan Pendekatan Algoritma K-Nearest Neighbor,” *J. Ilm. Inform. Komput.*, Vol. 29, No. 1, Pp. 45–59, 2024, Doi: 10.35760/Ik.2024.V29i1.10069.

[14] H. Maulana *Et Al.*, “Clustering Rfm (Recency, Frequency, Monetary) Publisher Gim Menggunakan Algoritma K-Means,” *Semin. Nas. Inform. Bela Negara*, Vol. 3, Pp. 2747–0563, 2023.

[15] Ahmad Harmain, P. Paiman, H. Kurniawan, K. Kusrini, And Dina Maulina, “Normalisasi Data Untuk Efisiensi K-Means Pada Pengelompokan Wilayah Berpotensi Kebakaran Hutan Dan Lahan Berdasarkan Sebaran Titik Panas,” *Tek. Teknol. Inf. Dan Multimed.*, Vol. 2, No. 2, Pp. 83–89, 2022, Doi: 10.46764/Teknimedia.V2i2.49.

[16] P. P. Allorerung, A. Erna, And M. Bagussahrir, “Analisis Performa Normalisasi Data Untuk Klasifikasi K-Nearest Neighbor Pada Dataset Penyakit,” Vol. 9, No. 3, Pp. 178–191, 2024.