JIKO (Jurnal Informatika dan Komputer)
Vol. 4, No. 2, Agustus 2021, hlm. 98-104
DOI: 10.33387/jiko

Akreditasi KEMENRISTEKDIKTI, No.36/E/KPT/2019
p-ISSN: 2614-8897
e-ISSN: 2656-1948

ANALISIS DAN KOMPARASI ALGORITMA KLASIFIKASI DALAM INDEKS PENCEMARAN UDARA DI DKI JAKARTA

Syekh S A Umri[1], Muhammad S Firdaus[2], Aji Primajaya[3]

[1,2,3] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Kec. Telukjambe Tim., Kabupaten Karawang, Jawa Barat 41361
Email: [1]syekh.syihabuddin17023@student.unsika.ac.id, [2]muhammad.syam17155@student.unsika.ac.id,
[3]aji.primajaya@staff.unsika.ac.id

(Naskah masuk: 3 Maret 2021, diterima untuk diterbitkan: 1 April 2021)

Abstrak

Udara mengambil peran penting dalam menjaga kehidupan makhluk hidup di bumi. DKI Jakarta merupakan salah satu kota dengan peringkat tertinggi dalam kualitas udara yang terburuk di dunia. Hal ini diakibatkan karena faktor utama selain aktivitas umum manusia yang menyumbang udara tidak baik ini berasal dari pabrik industri dan operasi pembangkit listrik berbahan bakar fosil pada daerah lintas batas administratif Jakarta. Fokus penelitian ini adalah melakukan analisis dan komparasi dari berbagai algoritme klasifikasi yakni *Neural Network*, *Support Vector Machine*, *K-Nearest Neighbors*, *Naive Bayes*, dan *Decission Tree* dengan menggunakan *T-Test* sebagai metode uji *parametrik* untuk menghasilkan perbandingan metode yang lebih baik dalam menentukan level Indeks Standar Pencemar Udara berdasarkan lima pencemar utama yakni arbon monoksida (CO), sulfur dioksida (SO2), nitrogen dioksida (NO2), ozon permukaan (O3), dan partikel debu (PM10) menghasilkan performa model terbaik yakni *Decission Tree* dengan nilai akurasi sebesar 99.80%, nilai kappa yang hampir sempurna yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Meskipun begitu, *Neural Network Backpropagation*, *K-Nearest Neighbors*, *Support Vector Machine*, dan *Naive Bayes* juga masih dapat digunakan sebagai model klasifikasi yang baik karena mendapatkan nilai akurasi yang tinggi di atas 90% dan nilai kappa di atas 0.8 dan nilai RMSE di bawah 0.3.

Kata kunci: *kualitas udara, ISPU, jakarta, data mining, klasifikasi*

ANALYSIS AND COMPARISON OF CLASSIFICATION ALGORITHM IN AIR POLLUTION INDEX IN DKI JAKARTA

Abstract

Air plays an important role in maintaining the life of living things on earth. DKI Jakarta is one of the cities with the highest ranking for the worst air quality in the world. This is because the main factors other than general human activities that contribute to this bad air come from industrial factories and the operation of fossil fuel-fired power plants in areas across the administrative boundaries of Jakarta. The focus of this research is to analyze and compare various classification algorithms, namely Neural Networks, Support Vector Machines, K-Nearest Neighbors, Naive Bayes, and Decission Tree using the T-Test as a parametric test method to produce a better comparison of methods in determining levels. The Air Pollutant Standard Index based on five main pollutants namely carbon monoxide (CO), sulfur dioxide (SO2), nitrogen dioxide (NO2), surface ozone (O3), and dust particles (PM10) produces the best model performance, namely the Decission Tree with an accuracy value of 99.80%, the kappa value is almost perfect, namely 0.996, the smallest RMSE value is below 0.1, namely 0.039, and the time needed is only 0.8 seconds. Even so, Neural Network Backpropagation, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes can still be used as good classification models because they get high accuracy values above 90% and kappa values above 0.8 and RMSE values below 0.3.

Keywords: *air quality, ISPU, jakarta, data mining, classification*

1. PENDAHULUAN

Udara mengambil peran penting dalam menjaga kehidupan makhluk hidup di bumi. Proses metabolisme yang terjadi dalam tubuh makhluk hidup tidak dapat berlangsung tanpa adanya oksigen yang berasal dari udara. Namun, selain oksigen, kandungan dalam udara terdapat juga zat lain seperti karbon monoksida, karbon dioksida, sulfur dioksida,

nitrogen oksida, ozon dan zat lain sebagainya. Beberapa zat - zat tersebut masih dapat ditoleransi oleh tubuh dalam konsentrasi yang masih di bawah batas wajar, namun jika melampaui batas wajar dapat menyebabkan masalah serius pada kesehatan manusia seperti penyakit jantung, infeksi pernapasan, stroke, asma hingga penyakit paru [1]. Salah satu faktor penyebab peningkatan konsentrasi zat - zat tersebut yang ada di dalam udara adalah aktivitas manusia.

Berdasarkan pemantauan kualitas udara yang dilakukan oleh US Air Quality Indeks (AQI) pada kuarter tiga tahun 2019, DKI Jakarta pernah menduduki peringkat pertama dalam kualitas udara terburuk di dunia dengan angka 179 kategori tidak sehat melalui parameter PM2,5 dengan tingkat konsentrasi hingga 110 µg/m³ [2], selain itu dalam penelitian lain yang dilakukan oleh [1] di DKI Jakarta mengungkapkan bahwa selama 2017 hingga 2019 Jakarta hanya mengalami kualitas udara yang terus memburuk kualitasnya yang ditandai semakin banyaknya jumlah hari tidak sehat yang terus meningkat setiap tahunnya. Pada pemantauan tahun 2020, meskipun terdapat penurutan aktivitas manusia yang disebabkan oleh COVID-19, namun kualitas udara dari 3 bulan yakni bulan Maret hingga Mei masih dalam tingkat sedang hingga tidak sehat. Hal ini diakibatkan karena faktor utama selain aktivitas umum manusia yang menyumbang udara tidak baik ini berasal dari pabrik industri dan operasi pembangkit listrik berbahan bakar fosil pada daerah lintas batas administratif Jakarta.

Pemerintah Indonesia berdasarkan Keputusan Badan Pengendalian Dampak Lingkungan (Bapedal) Nomor KEP-107/Kabapedal/11/1997 menetapkan Indeks Standar Pencemar Udara (ISPU) untuk menentukan kualitas udara di suatu daerah dan bagaimana dampaknya terhadap kesehatan setelah menghirup udara tersebut selama beberapa jam hingga hari. Semakin tinggi tingkatan level ISPU semakin tidak baik udara tersebut terhirup oleh tubuh. ISPU memiliki lima level yakni Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat dan Berbahaya [3].

Penentuan tingkat level ISPU dapat dipermudah dengan menggunakan proses klasifikasi dari *data mining*. *Data mining* merupakan salah satu cara dalam menggali informasi baru dengan menemukan aturan atau pola tertentu dari berbagai data yang berjumlah besar [4]. Beberapa penelitian dalam mengklasifikasikan level ISPU di DKI Jakarta sebelumnya pernah dilakukan oleh [5] pada tahun 2019, yang mengklasifikasikan pencemaran udara di wilayah DKI Jakarta menggunakan algoritme *Support Vector Machine* (SVM) dengan hasil akurasi tertinggi yaitu 96,03%.

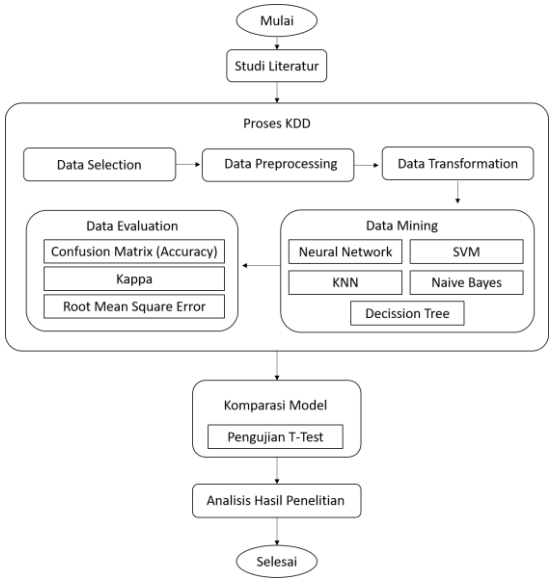
Penelitian serupa lainnya telah dilakukan oleh [6] pada tahun 2019, yang meneliti tentang perbandingan metode *Naive Bayes* dan *K-Nearest Neighbor* pada klasifikasi kualitas udara di DKI Jakarta. Hasil dari penelitian perbandingan algoritme tersebut akurasi tertinggi diperoleh oleh hasil proses data dengan algoritme *Naive Bayes*, dengan tingkat akurasi sebesar 97,3396%.

Sayangnya pada penelitian sebelumnya hanya melakukan evaluasi pada berdasarkan akurasi saja tanpa melihat nilai evaluasi lainnya yang dapat dijadikan salah satu pertimbangan dalam melakukan komparasi untuk mendapatkan model algoritme terbaik. Oleh karena itu pada penelitian ini dilakukan analisis dan komparasi dari berbagai algoritme klasifikasi dan membandingkan tidak hanya dengan nilai evaluasi berupa akurasi namun juga dengan nilai evaluasi kappa dan *root mean square error* sehingga mendapatkan model dari algoritme yang terbaik. Algoritme yang dipakai yakni *Neural Network Backpropagation, Support Vector Machine, K-Nearest Neighbors, Naive Bayes,* dan *Decission Tree*. Algoritme tersebut dipakai karena mampu memproses data dalam bentuk *polynomial* yang dimiliki oleh data ISPU. Kemudian dilakukan pengujian dengan menggunakan *T-Test* sebagai uji *parametrik* dari nilai evaluasi *confusion matrix/accuracy*, *kappa*, dan *root mean square error*. Hasil uji yang diperoleh tersebut dilakukan komparasi sehingga menghasilkan metode model algoritme yang lebih baik dalam menentukan level ISPU berdasarkan lima pencemar utama yakni karbon monoksida (CO), sulfur dioksida (SO2), nitrogen dioksida (NO2), Ozon permukaan (O3), dan partikel debu (PM10).

2. METODE PENELITIAN

2.1. Skema Alur Penelitian

Penelitian dalam mengklasifikasikan level ISPU yang dilakukan untuk menentukan kualitas udara di DKI Jakarta menggunakan skema alur penelitian sebagai berikut :



Gambar 1. Skema Penelitian

Proses penelitian ini diawali dengan mencari informasi dengan studi literatur. Kemudian dilanjutkan dengan proses dari metodologi *data mining* yakni *Knowledge Discovery In Database*. *Knowledge Discovery In Database* (KDD) merupakan metode untuk memperoleh pengetahuan dari tabel – tabel yang saling terhubung dari basis data yang ada sehingga dapat digunakan sebagai basis pengetahuan untuk dijadikan sebagai pengambilan keputusan [7].

Proses KDD ini dimulai dari *data selection*, *data preparation*, *data transformation*, *data mining* yang menggunakan lima algoritme klasifikasi, dan data evaluation dengan menggunakan nilai evaluasi *confusion matrix/accuracy*, *kappa*, dan *root mean square error*. Dari hasil evaluasi tersebut di komparasi menggunakan pengujian para metrik *T-Test* yang selanjutnya dianalisis untuk membentuk pengetahuan baru.

Penelitian ini dilakukan dengan menggunakan komputer pribadi dengan *processor* Intel Core i7 9750H @ 2.60GHz dengan *boost clock* hingga 4.5 GHz, GPU Nvidia GTX 1660 TI 6GB, RAM 16 GB DDR4 2400MHz, dan sistem operasi Microsoft Windows 10 Pro 64 bit Versi 20H2. Sedangkan untuk lingkungan pengembangan aplikasi yang digunakan adalah RapidMiner Studio versi 9.8.

2.2. Objek Penelitian

Obyek yang diteliti pada penelitian ini adalah data pencemar udara dari Indeks Pencemar Standar Udara di DKI Jakarta. Kota DKI Jakarta dipilih karena Berdasarkan pemantauan kualitas udara yang dilakukan oleh US Air Quality Indeks (AQI) pada kuarter tiga tahun 2019, DKI Jakarta pernah menduduki peringkat pertama dalam kualitas udara terburuk di dunia [2], selain itu dalam penelitian lain yang dilakukan oleh [1] di DKI Jakarta mengungkapkan bahwa selama 2017 hingga 2019 Jakarta hanya mengalami kualitas udara yang terus memburuk kualitasnya yang ditandai semakin banyaknya jumlah hari tidak sehat yang terus meningkat setiap tahunnya.

Dari data tersebut diteliti untuk membuat perbandingan antar algoritme klasifikasi untuk mencari algoritme terbaik dalam memprediksi kualitas udara yang baik, sedang, tidak sehat, sangat tidak sehat dan berbahaya menggunakan *data mining*.

2.3. Studi Literatur

Tahap awal dalam merumuskan penelitian ini dilakukan dengan cara mengumpulkan informasi yang menunjang penelitian yakni berkaitan dengan kualitas udara, ISPU dan berbagai algoritme klasifikasi yang diperoleh dari buku, artikel, jurnal, dan dokumen lainnya.

2.4. *Data Selection*

Dataset Indeks Pencemar Standar Udara DKI Jakarta dikumpulkan dan diseleksi dari laman situs terbuka pemerintah DKI Jakarta http://www.data.jakarta.go.id/ berupa CSV yang diambil per SPKU dan per daerah yakni Jakarta Pusat (DKI1), Jakarta Utara (DKI2), Jakarta Selatan (DKI3), Jakarta Timur (DKI4), dan Jakarta Barat (DKI5) dari tahun 2017 hingga bulan Juni 2020.

2.5. *Data Preparation*

Data yang sudah dikumpulkan dan diseleksi akan memasuki tahap *data preparation*. Pada tahap ini data akan di reduksi dengan menghapus data yang memiliki nilai kosong, data duplikat, data yang terisi namun tidak memiliki data secara lengkap, dan penghapusan atribut yang tidak digunakan. Kemudian karena klasifikasi masuk dalam *supervised learning* maka dibutuhkan tahap *labeling* pada atribut yang akan diklasifikasikan.

2.6. *Data Transformation*

Pada tahap ini dari dataset yang masih terpisah antar berkas CSV diintegrasikan dan disesuaikan atribut yang ada sehingga menjadi kesatuan dataset untuk mempermudah dalam proses selanjutnya.

2.7. *Data Mining*

Tahap inti dari penelitian ini, dengan membuat model klasifikasi menggunakan algoritme *Neural Network Backpropagation*, *Support Vector Machine*, *K-Nearest Neighbors*, *Naive Bayes*, dan *Decission Tree*.

Neural Network adalah algoritme *data mining* yang mengambil konsep dari sistem saraf manusia [8].Salah satu metode dari *Neural Network* adalah *Backpropagation* yang memiliki satu atau lebih lapisan yang tersembunyi dan proses propagasi balik untuk perbaikan kesalahan yang ada [9]. *Support Vector Machine* merupakan algoritme yang memisahkan dua kelompok data dari dua kelas berbeda dengan memaksimalkan batas fungsi pemisah yang disebut *hyperplane* [10]. *K-Nearest Neighbor* (KNN) merupakan algoritme untuk mengklasifikasi suatu objek berdasarkan k buah data latih yang jaraknya paling dekat dengan objek tersebut [11]. *Naive Bayes* merupakan algoritme klasifikasi data yang menggunakan metode probabilitas dan statistik dan menggunakan aturan yakni *bayes rule* [12]. Decission Tree merupakan algoritme yang memiliki struktur yang digunakan untuk membagi himpunan data besar menjadi beberapa himpunan *record* yang lebih kecil dengan menggunakan serangkaian aturan keputusan [13].

Kelima algoritme tersebut dipilih dikarenakan algoritme tersebut dapat memproses data dalam bentuk *polynomial* yang ada pada format label ISPU. *Backward elimination* digunakan sebagai *feature*

selection dalam mereduksi informasi yang tidak relevan dalam dataset. 10-*fold cross validation* digunakan untuk membentuk kombinasi data yang baik dan model yang efisien sehingga mencegah data menjadi *overfitting* [14]. Hal ini dilakukan dengan membagi dataset menjadi 10 *fold* berukuran sama dan membuat 10 *subset* data yang mana masing – masing *subset* terdapat 9 *fold* untuk pelatihan dan 1 *fold* untuk pengujian. Hasil validasi tersebut akan menghasilkan nilai akurasi yang didapatkan dari *confusion matrix* yang dapat dilihat dari tabel berikut ini :

Tabel 1. *Confusion Matrix*

		Aktual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negatif (TN)

Tabel *confusion matrix* tersebut akan menghasilkan nilai TP, FP, FN, dan TN yang dapat dihitung menjadi nilai akurasi dengan menggunakan persamaan 1.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

Selain nilai akurasi, sebagai metrik perbandingan klasifikasi pada penelitian menggunakan nilai kappa. Nilai kappa digunakan untuk mengukur nilai konsistensi antar penilai (*inter-rater agreement*) [15]. Nilai Kappa dapat dihitung dari tabel kontingensi (Tabel 2). Namun dengan menggunakan hasil dari *confusion matrix* nilai kappa dapat didapatkan dengan menggunakan persamaan 2. Nilai kappa yang dihasilkan dari rumus tersebut dapat diinterpretasi seperti yang ditunjukkan pada Tabel 3.

Tabel 2. 2x2 Tabel Kontingensi

Penilai 1	**Penilai 2**		**Total**
	1	**2**	
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

$$K = \frac{p_0 - p_e}{1-p_e} \quad (2)$$

Dari persamaan tersebut yang mana jika dihubungkan dengan *confusion matrix* nilai p_0 dan p_e :

$$p_0 = p_{11} + p_{12} = \frac{TP+TN}{TP+FN+FP+FN} \quad (3)$$

$$p_e = p_{.1}p_1 + p_{.2}p_2 \quad (4)$$

dengan

$$p_{.1}\,p_1 = \frac{(TP+FN)(TP+FP)}{(TP+FN+FP+TN)^2} \quad (5)$$

$$p_{.2}\,p_2 = \frac{(FP+TN)(FN+TN)}{(TP+FN+FP+TN)^2} \quad (6)$$

Tabel 3. Interpretasi Kappa

Kappa	**Agreement**
<0.00	Buruk
0.00-0.20	Sedikit buruk
0.21-0.40	Cukup
0.41-0.60	Sedang
0.61-0.80	Bagus
0.81-1.00	Hampir Sempurna

Metrik perbandingan selanjutnya yang dipakai adalah *Root Mean Square Error* (RMSE) yang merepresentasikan standari deviasi dari perbedaan atau ukuran kesalahan nilai yang diprediksi dengan nilai asli atau nilai yang diobservasi. Semakin keci nilai RMSE maka semakin baik model tersebut [16]. Untuk mendapatkan nilai RMSE dapat menggunakan persamaan 16.

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(Z_t - \hat{Z}_t)^2} \quad (16)$$

Keterangan :
Z_t : Nilai Observasi
$\hat{Z}_t$: Nilai Prediksi

2.8. Komparasi Model

Dari evaluasi yang dihasilkan berupa nilai akurasi dari masing – masing model algoritme akan di komparasi dengan menggunakan pengujian hipotesis *parametrik T-Test* apakah hasil tersebut memiliki perbedaan yang signifikan atau tidak.

3. HASIL DAN PEMBAHASAN

Hasil seleksi pengumpulan dataset Indeks Pencemar Standar Udara berupa berkas CSV dari laman situs terbuka http://www.data.jakarta.go.id/ yang diambil per SPKU dan per daerah yakni Jakarta Pusat (DKI1), Jakarta Utara (DKI2), Jakarta Selatan (DKI3), Jakarta Timur (DKI4), dan Jakarta Barat (DKI5) sejak tahun 2017 hingga Juni 2020 yang sudah dikumpulkan per tahun adalah sebagai berikut :

Tabel 4. Jumlah *Record* dalam Dataset

Tahun	**Jumlah Dataset**	**Jumlah *Record* dalam Seluruh Dataset**
2017	1	1.825
2018	24	2.190
2019	24	2.170
2020	12	1.092
Total	61	7.277

Dari 61 dataset yang terdiri dari 7.277 *record* perlu disesuaikan untuk atribut yang dipakai dan atribut yang memerlukan proses *data preparation* untuk membantu dan mengefisienkan proses klasifikasi level ISPU ditunjukkan pada tabel 5 berikut.

Tabel 5. Atribut Data ISPU DKI Jakarta

Atribut	Data Kosong	Proses	Keterangan
Tanggal	0	Tidak digunakan	Waktu pengambilan data ISPU
Lokasi SPKU	4591	Tidak digunakan	Lokasi SPKU yang melakukan pengambilan data
Stasiun	2872	Tidak digunakan	Lokasi stasiun yang melakukan pengambilan data
PM10	240	Digunakan	Partikel debu yang berukuran kurang dari 10 mikron
SO2	220	Digunakan	Sulfur dioksida
CO	232	Digunakan	Karbon monoksida
O3	240	Digunakan	Ozon
NO2	274	Digunakan	Nitrogen oksida
Max	31	Tidak Digunakan	Nilai tertinggi dari seluruh parameter atribut yang diukur di waktu yang bersamaan
Critical	100	Tidak Digunakan	Atribut parameter yang memiliki nilai paling tinggi
Categori	0	Digunakan	Kategori hasil perhitungan level ISPU

Dari Tabel 5 di atas, terlihat bahwa atribut lokasi SPKU dan stasiun memiliki data kosong yang besar karena terdapat dua tipe dataset yakni dataset per SPKU yang memiliki atribut lokasi SPKU dan dataset per daerah yang memiliki atribut stasiun, namun karena atribut ini tidak digunakan sebagai parameter klasifikasi maka hal tersebut bukan menjadi masalah. Selain itu, terdapat atribut tanggal, *max*, dan *critical* yang juga tidak digunakan. Untuk atribut PM10, SO2, CO, O3, NO2 diperlukan tahap *preparation* untuk mengatasi data yang kosong yakni penghapusan data tersebut. Sedangkan untuk atribut "categori" meskipun tidak mempunyai data kosong namun terdapat inkonsistensi data yakni penggunaan huruf kapital dan huruf kecil yang tidak konsisten dan penghapusan data yang memiliki nilai "TIDAK ADA DATA" karena data tersebut tidak memiliki makna dan tidak digunakan. Kemudian dilakukan pengecekan nilai duplikat untuk dihilangkan karena dapat mempengaruhi proses klasifikasi.

Selanjutnya dari semua dataset yang berjumlah 61 dataset di integrasikan menjadi satu dataset dengan jumlah *record* tereduksi menjadi 6.536 setelah proses *preparation*. Kemudian dilakukan proses *data mining* dari algoritme yang berbeda, yakni SVM, *Decission Tree, K-Nearest Neighbor, Naive Bayes*, dan *Neural Network Backpropagation* untuk dijadikan model klasifikasi yang diperlakukan hal yang sama yakni menggunakan *feature selection* berupa *backward elimination* untuk mereduksi informasi yang tidak relevan dan juga 10-*fold cross validation* untuk membagi dataset menjadi data latih dan data uji.

Pembuatan model menggunakan algoritme SVM dilakukan dengan menggunakan LibSVM dengan parameter tipe SVM = C-SVC, tipe *kernel*= rbf, gamma = 0.0, C = 0.0, ukuran cache = 80, epsilon = 0.001, mengaktifkan opsi *shrinking* dan *confidence for multiclass*. Untuk algoritme *Decission Tree* menggunakan parameter *criterion = grain_ratio*, *maximal depth* = 10 dengan mengaktifkan opsi apply *pruning, confidence* 0.1 dengan mengaktifkan opsi *prepruning, minimal gain* = 0.01, *minimal leaf size* 2, *minimal size for split* = 5 dan n*umber of prepuring alternatives* = 3. Untuk algoritme KNN menggunakan parameter k = 5, mengaktifkan opsi *weighted vote, measure types = MixedMeasures*, dan *mixed measure = MixedEuclideanDistance*. Untuk algoritme *Naive Bayes* mengaktifkan opsi *laplace correction*. Untuk algoritme *Neural Network Backpropagation* menggunakan parameter *hidden* layer *sigmoid*, *training cycles* = 1000, *learning rate* = 0.2, momentum 0.9, mengaktifkan opsi *shuffle* dan *normalize*, dan *error* epsilon = 1.0E-4.

Hasil pengujian dari kelima algoritme tersebut menghasilkan tabel *confusion matrix* yang dapat dilihat pada Tabel 6 hingga Tabel 10 dan dari tabel *confusion matrix* tersebut dapat direpresentasi ke dalam bentuk nilai *accuracy*, *kappa*, dan RMSE pada Tabel 11.

Tabel 6. *Confusion Matrix* SVM

	TS	TB	TTS	TSTS	CP
PS	3973	106	43	0	96.39%
PB	66	1030	0	0	93.98%
PTS	3	0	1240	3	99.52%
PSTS	0	0	1	71	98.61%
CR	98.29%	90.57%	96.57%	95.95%	

Tabel 7. *Confusion Matrix Decission Tree*

	TS	TB	TTS	TSTS	CP
PS	4033	0	1	0	99.98%
PB	9	1136	0	0	99.21%
PTS	0	0	1282	2	99.84%
PSTS	0	0	1	72	98.63%
CR	99.78%	100%	99.84%	97.30%	

Tabel 8. *Confusion Matrix* KNN

	TS	TB	TTS	TSTS	CP
PS	3960	46	24	0	98.26%
PB	72	1090	0	0	93.80%
PTS	10	0	1255	3	98.97%
PSTS	0	0	5	71	93.42%
CR	97.97%	95.95%	97.74%	95.95%	

Tabel 9. *Confusion Matrix Naive Bayes*

	TS	TB	TTS	TSTS	CP
PS	3899	170	188	1	91.57%
PB	135	966	7	1	87.11%
PTS	8	0	1086	17	97.75%
PSTS	0	0	3	55	94.83%
CR	96.46%	85.04%	84.58%	74.32%	

Tabel 10. *Confusion Matrix Neural Network*

	TS	TB	TTS	TSTS	CP
PS	3980	7	29	2	99.05%
PB	58	1129	0	0	95.11%
PTS	4	0	1255	30	97.36%
PSTS	0	0	0	42	100%
CR	98.47%	99.38%	97.74%	56.76%	

Keterangan :
TS : *True* SEDANG
TB : *True* BAIK
TTS : *True* TIDAK SEHAT
TSTS : *True* SANGAT TIDAK SEHAT
PS : Prediksi SEDANG
PB : Prediksi BAIK
PTS : Prediksi TIDAK SEHAT
PSTS : Prediksi SANGAT TIDAK SEHAT
CP : *Class Precision*
CR : *Class Recall*

Tabel 11. Hasil Evaluasi Klasifikasi

Algoritme	Accuracy	Kappa	RMSE	Waktu
SVM	96.60%	0.937	0.184	11 detik
Decission Tree	99.80%	0.996	0.039	0.8 detik
KNN	97.55%	0.955	0.135	3 detik
Naive Bayes	91.89%	0.848	0.255	0.2 detik
Neural Network	98.01%	0.964	0.132	2 menit 28 detik

Dari hasil evaluasi yang dapat dilihat pada Tabel 11 menunjukkan bahwa model dengan nilai *accuracy* dan nilai kappa tertinggi adalah dengan menggunakan algoritme *Decission Tree* sebesar 99.80% dan 0.996 yang berarti hampir sempurna dengan nilai RMSE terkecil yakni 0.039 dan waktu eksekusi hanya 0.8 detik. Meskipun begitu, *Neural Network Backpropagation, KNN, SVM*, dan *Naive Bayes* juga masih dapat digunakan sebagai model klasifikasi yang baik karena mendapatkan nilai akurasi yang tinggi di atas 90% dan nilai kappa di atas 0.8. *Neural Network Backpropagation* yang memiliki nilai akurasi sebesar 98.01%, nilai kappa 0.964, dan nilai RMSE 0.132 namun sayangnya memerlukan waktu yang relatif jauh lebih lama sebesar 2 menit 28 detik. KKN memiliki nilai akurasi sebesar 97.55%, nilai kappa sebesar 0.955, nilai RMSE sebesar 0.135 dan memerlukan waktu sebanyak 3 detik. SVM memiliki nilai akurasi sebesar 96.60%, nilai kappa sebesar 0.937, nilai RMSE sebesar 0.184 dan memerlukan waktu 11 detik. *Naive Bayes* yang mendapatkan nilai lebih rendah dibandingkan yang lain mulai dari *accuracy* hanya 91.89%, nilai kappa 0.848, dan nilai RMSE 0.255, meskipun dalam waktu merupakan yang tercepat yakni 0.2 detik.

Selanjutnya dari hasil evaluasi yang dihasilkan dilakukan pengujian hipotesis *T-Test* untuk melihat adakah tidak ada perbedaan yang signifikan di bawah rata-rata antara hasil akurasi ke lima model yang dihasilkan tersebut (H_0) atau ada perbedaan yang signifikan (H_1) . Taraf signifikansi yang dipakai pada penelitian ini adalah 0.5 dengan hasil sebagai berikut :

Tabel 12. Hasil Uji T-Test

	SVM	Decission Tree	KNN	Naive Bayes	Neural Network
SVM	-	0.000	0.000	0.000	0.000
Decission Tree	0.000	-	0.000	0.000	0.000
KNN	0.000	0.000	-	0.248	0.002
Naive Bayes	0.000	0.000	0.248	-	0.001
Neural Network	0.000	0.000	0.002	0.001	-

Berdasarkan hasil pada Tabel 12, dapat dinyatakan bahwa baik dari penggunaan algoritme *SVM, Decission Tree, KNN*, dan *Neural Network* masing – masing jika dibandingkan satu sama lain mempunyai nilai rata-rata yang signifikan (H_1). H_0 atau tidak memiliki rata-rata perbedaan yang signifikan hanya terjadi pada perbandingan KNN dengan *Naive Bayes* dan sebaliknya.

4. KESIMPULAN

Penelitian dengan menggunakan dataset Indeks Pencemar Udara pada wilayah DKI Jakarta dengan melakukan komparasi atau perbandingan antara lima algoritme yaitu SVM, *Decission Tree,* KNN, dan *Neural Network Backpropagation* dengan menggunakan validasi 10-*fold cross validation* dan *feature selection* berupa *backward elimination* menghasilkan algoritme dengan performa terbaik yakni *Decission Tree* dengan nilai akurasi sebesar 99.80%, nilai kappa yang hampir sempurna yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Meskipun begitu, *Neural Network Backpropagation, KNN, SVM*, dan *Naive Bayes* juga masih dapat digunakan sebagai model klasifikasi yang baik karena mendapatkan nilai akurasi yang tinggi di atas 90% dan nilai kappa di atas 0.8.

5. DAFTAR PUSTAKA

[1] L. Myllyvirta., I. Suarez., dan E. Uusivuori., 2020. Pencemaran Udara Lintas Batas di provinsi Jakarta, Banten, dan Jawa Barat. Tersedia [https://energyandcleanair.org/wp/wp-content/uploads/2020/08/Jakarta-Transboundary-Pollution_Final-Bahasa.pdf] diakses 1 Maret 2021.

[2] M. Linawati., 2019. “Data AirVisual: Kualitas Udara DKI Jakarta Peringkat 1 Terburuk di

Dunia - News Liputan6.com," Liputan 6, 23 September. Tersedia [https://www.liputan6.com/news/read/4069080/data-airvisual-kualitas-udara-dki-jakarta-peringkat-1-terburuk-di-dunia] diakses 1 Maret 2021.

[3] Bapelda,. 1999 "PEDOMAN TEKNIS PERHITUNGAN DAN PELAPORAN SERTA INFORMASI INDEKS STANDAR PENCEMAR UDARA,".Tersedia [http://www.cets-uii.org/BML/Udara/ISPU/ISPU%20(Indeks%20Standar%20Pencemar%20Udara).htm.] diakses 1 Maret 2021.

[4] D. Kartini, 2017. "Penerapan Data Mining dengan Algoritma Neural Network (Backpropagation) Untuk Prediksi Lama Studi Mahasiswa*," PROSIDING Seminar Nasional Sisfotek*, 3584, pp. 235–241. Tersedia [https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/44/36] diakses 2 Maret 2021.

[5] A. Hermawan, 2019. "SPKU: Sistem Prediksi Kualitas Udara (Studi Kasus: Dki Jakarta)". Yogyakarta. Tersedia [http://eprints.uty.ac.id/3552/] diakses 2 Maret 2021.

[6] M. J. Sodiq dan E. I. Sela, 2019. "Perbandingan Metode Naive Bayes Dan K-Nearest Neighbor Pada Klasifikasi Kualitas Udara Di DKI Jakarta."

[7] Y. Mardi, 2017. "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik Informatika*, vol.2(2), pp. 213–219.

[8] A. H. Baksir, et al. 2020. "Prediction of Fertility Quality Levels With Artificial Neural Network of Backpropagation," *JIKO (Jurnal Informatika dan Komputer)*, vol.3(2), pp. 107–112. doi: 10.33387/jiko.

[9] D. Jauhari., A. Himawan dan C. Dewi, 2016. "Prediksi Distribusi Air PDAM Menggunakan Metode Jaringan Syaraf Tiruan Backpropagation Di PDAM Kota Malang," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol.3(2), pp. 83. doi: 10.25126/jtiik.201632155.

[10] S. E. Anindika Sari., et al. 2020. "Klasifikasi Kabupaten Tertinggal di Kawasan Timur Indonesia dengan Support Vector Machine," *JIKO (Jurnal Informatika dan Komputer*), vol.3(3), pp. 188–195. doi: 10.33387/jiko.v3i3.2364.

[11] M. Rivki dan A. M. Bachtiar, 2017. "Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter Yang Menggunakan Bahasa Indonesia," *Jurnal Sistem Informasi*, vol.13(1), pp. 31. doi: 10.21609/jsi.v13i1.500.

[12] N. Nuraeni, 2017. "Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC," *Jurnal Teknik Komputer AMIK BSI (JTK)*, vol.3(1), pp. 9–15.

[13] P. B. N. Setio., et al. 2020. "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," PRISMA, *Prosiding Seminar Nasional Matematika*, vol.3, pp. 64–71.

[14] T. Tušar., et al. 2017. "A study of overfitting in optimization of a manufacturing quality control procedure," Applied Soft Computing Journal, vol.59, pp. 77–87. doi: 10.1016/j.asoc.2017.05.027

[15] J. R. Landis dan G. G. Koch, 1977. "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol.33(1), pp. 159. doi: 10.2307/2529310.

[16] N. Fadjrin., et al. 2020. "Pemodelan Deret Waktu Point Liga Italia Serie A Dengan Pendekatan Regresi Berdasarkan RMSE (Root Mean Square Score) Terkecil dan Skor Maksimal Tiap Pekan," *Statistika*, vol.8(1), pp. 78–87.