ANALISIS DATA MINING UNTUK KLASIFIKASI DATA KUALITAS UDARA DKI JAKARTA MENGGUNAKAN ALGORITMA *DECISION TREE* DAN *SUPPORT VECTOR MACHINE*

DATA MINNING ANALYSIS FOR CLASSIFICATION OF AIR QUALITY DATA DKI JAKARTA USING DECISION TREE ALGORTHM AND SUPPORT VECTOR MACHINER ALGORITHM

Adinda Inez Sang[1], Edi Sutoyo[2], Irfan Darmawan[3]

[1,2,3] Universitas Telkom, Bandung
[1]adindainezsang@student.telkomuniversity.ac.id, [2]edisutoyo@telkomuniveristy.co.id, [3]irfandarmawan@telkomuniversity.ac.id

Abstrak

Pertumbuhan dan perkembangan suatu kota merupakan salah satu faktor penyebab terjadinya pencemaran udara karena kualitas udara sudah tergabung dengan berbagai komponen senyawa. Menurut IQ Air 2020 prov. Untuk memantau pencemaran udara setiap harinya, Dinas Lingkungan Hidup Pemerintah Provinsi DKI Jakarta mengoperasikan Stasiun Pemantau Kualitas Udara (SPKU). Penggunaan data mining merupakan metode yang cocok untuk mengetahui informasi pencemaran udara di Provinsi DKI Jakarta. Metode data mining yang digunakan yaitu klasifikasi karena metode ini dapat mengolah data parameter ISPU menjadi informasi yang memberitahukan tingkat kualitas udara perharinya dengan menggunakan algoritma *Decision Tree* dan *Support Vector Machine* (SVM). Hasil dari penerapan *data mining* untuk klasifikasi kualitas udara di DKI Jakarta yaitu algoritma *Decision Tree* memiliki performa yang lebih baik dengan rasio terbaik 90:10 dibandingkan dengan algoritma SVM dengan rasio terbaik 60:40 dan untuk melakukan klasifikasi kualitas udara di DKI Jakarta. Pada algoritma *Decision Tree* mendapatkan nilai *Precision* sebesar 99,02%, *Recall* 99,73%, *F1-Measure* 99,37%, Akurasi 99,40% dan pada algoritma SVM mendapatkan nilai *Precision* sebesar 95,82%, *Recall* 88,89%, *F1-Measure* 92,22% dan Akurasi 94,93%.

Kata kunci : Klasifikasi, Kualitas Udara, *Decision Tree*, SVM

Abstract

The growth and development of a city are some of the factors causing air pollution because air quality has been combined with various components of compounds. According to IQ Air 2020 Prov. To monitor daily air pollution, the DKI Jakarta Provincial Government Environment Service operates an Air Quality Monitoring Station (SPKU). The use of data mining is a suitable method to find out information on air pollution in the DKI Jakarta Province. The data mining method used is classification because this method can process ISPU parameter data into information that tells the daily air quality level using the Decision Tree and Support Vector Machine (SVM) algorithms. The results of the application of data mining for air quality classification in DKI Jakarta, namely the Decision Tree algorithm has better performance with the best ratio of 90:10 compared to the SVM algorithm with the best ratio of 60:40 and for classifying air quality in DKI Jakarta. The Decision Tree algorithm gets a Precision value of 99.02%, Recall 99.73%, F1-Measure 99.37%, Accuracy 99.40%, and the SVM algorithm gets a Precision value of 95.82%, Recall 88.89%, F1-Measure 92.22%, and Accuracy 94.93%.

Keywords: Clustering, Air Quality, K-Medoids

1. Pendahuluan

Kemajuan teknologi informasi telah berpengaruh dalam segala aspek kehidupan baik di bidang ekonomi, politik, seni dan budaya bahkan didunia pendidikan [1]. Lingkungan yang baik merupakan kebutuhan paling mendasar bagi seluruh umat manusia untuk mempertahankan kehidupan, maka dari itu manusia sangat membutuhkan udara yang baik [2]. Suatu kota yang memiliki pertumbuhan dan perkembangan padat dan pesat menjadi salah satu faktor terjadinya pencemaran udara di suatu perkotaan. Pencemaran udara terjadi karena kualitas udara sudah tergabung dengan berbagai komponen senyawa. Pencemaran udara yang merugikan kesehatan masyarakat merupakan masalah yang tersebar luas di banyak negara di dunia [3]. Kota dengan kualitas udara tidak sehat salah satunya kota yang berada di benua Asia yaitu DKI Jakarta sebagai ibukota Indonesia dengan label warna orange yang menandakan tidak sehat bagi sekelompok *sensitive* (IQAir, 2020). Indeks Standar Pencemaran

Udara (ISPU) digunakan sebagai parameter untuk mengukur kualitas udara. Pada pengukuran ISPU menetapkan lima parameter pencemaran udara yang digunakan untuk pengamatan, yaitu Karbon Monoksida (CO), Sulfur Dioksida (SO_2), Nitrogen Dioksida (NO_2), Ozon Permukaan (O_3), Partikel Debu (PM_{10}). Pencemaran udara dapat disebabkan dari perkembangan kendaraan bermotor. Banyaknya masyarakat yang memiliki kendaraan pribadi mengakibatkan pencemaran udara semakin meningkat. Pertumbuhan kendaraan bermotor di DKI Jakarta dari tahun 2018 hingga tahun 2020 mengalami peningkatan yang signifikan sebesar 19% atau sebanyak 8.381.000 kendaraan dari tahun sebelumnya [5].

Berdasarkan permasalahan yang ada, maka dari itu Dinas Lingkungan Hidup Provinsi DKI Jakarta memerlukan pengolahan data yang dapat memberikan informasi atau yang biasa disebut *data mining.* Data mining untuk Dinas Lingkungan Hidup Provinsi DKI Jakarta berfungsi untuk menggali informasi lebih dalam dengan jumlah data yang banyak. Implementasi *data mining* menjadi solusi untuk mengetahui kualitas udara DKI Jakarta menggunakan teknik klasifikasi. Klasifikasi merupakan sesuatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu daru sejumlah kelas yang tersedia [6]. Dalam melakukan klasifikasi kualitas udara DKI Jakarta menggunakan Algoritma *Decision Tree* dan *Support Vector Machine (*SVM). Beberapa penelitian sebelumnya menggunakan Algoritma *Decision Tree* dan SVM yaitu penelitian tentang klasifikasi kualitas udara menggunakan algoritma *Decision tree* dimana penelitian ini mendapat hasil akurasi 85,71%, presisi 81,82%, dan spesifisitas 92,31% [7]. Selanjutnya penelitian tentang perbandingan algoritma *decision tree* dan *naïve bayes* dalam melakukan klasifikasi kualitas udara yang mendapatan hasil bahwa algoritma *decision tree* memberikan akurasi 91,9978% yang lebih tinggi dari Algoritma *Naïve Bayes* yaitu 86,663% [8]. Selanjutnya penelitian tentang perbandingan algoritma SVM, *Neural Network*, dan KNN dalam melakukan klasifikasi kualitas udara yang mendapatan hasil bahwa SVM berkinerja lebih baik algoritma Neural Network dan KNN di hal akurasi serta kompleksitas [9]. Penelitian tentang klasifikasi kualitas udara menggunakan algoritma SVM yang mendapatan hasil bahwa performansi ketepatan klasifikasi terbaik pada node sensor yaitu 99,33% [10]. Serta beberapa penelitian terkait yang memberikan hasil bahwa algoritma-algoritma tersebut memiliki performansi yang lebih baik dibandingkan algoritma lainnya [11]–[13].

Data yang digunakan memiliki kemiripan dengan data penelitian kualitas udara yang sudah dilakukan sebelumnya, yaitu data bersifat kontinu, berbentuk numerik, dan memiliki topik yang sama. Sehingga klasifikasi kualitas udara di DKI Jakarta cocok dilakukan dengan menggunakan algoritma *Decision Tree* dan SVM. Harapannya adalah dapat menghasilkan informasi bagaimana kualitas udara di DKI Jakarta berdasarkan data yang didapatkan dari Website *Jakarta Open Data*. Selain itu, tujuan penelitian ini menggunakan algoritma *Decision Tree* dan SVM adalah diharapkan nantinya *output* dari penelitian ini dapat menjadi acuan untuk Dinas Lingkungan Hidup dapat melakukan klasifikasi data kualitas udara berdasarkan atribut yang ada agar pihak Dinas Lingkungan Hidup dapat melakukan penangan lebih cepat.

2. Dasar Teori dan Metodologi

2.1 Dasar Teori

2.1.1 Polusi Udara

Masalah dunia yang dihadapi oleh disebagian besar negara di dunia adalah polusi udara [14]. Polusi udara menjadi perhatian besar yang efeknya jika tidak dikurangi dapat menyebabkan masalah kesehatan pada tubuh manusia [15]. Dengan adanya polusi udara yang semakin serius, mengembangkan sistem peringatan dini untuk prakiraan kualitas udara sangat penting untuk memantau dan mengendalikan kualitas udara [16]. Pada tingkat konsentrasi tertentu, pencemaran udara dapat berakibat langsung terhadap kesehatan manusia, dimulai dari saluran pernafasan, iritasi mata, dan alergi kulit [17]. Karena dampak kesehatan dan ekonomi dari polusi udara, sensor kualitas udara berbiaya rendah dan portabel dapat digunakan secara luas untuk mendapatkan paparan polutan udara pribadi [18].

2.1.2 *Data Mining*

Data mining adalah proses menganalisis data dari berbagai perpektif dan merangkum menjadi informasi yang berguna, informasi tersebut dapat digunakan untuk meningkatkan pendapatan dan memangkas biaya, hal ini memungkinkan pengguna untuk menganalisis data dari berbagai dimensi, mengkategorikan, dan merangkum hubungan yang diidentifikasi [19]. Tetapi dalam beberapa tahun terakhir, selain untuk mengolah data, teknik *data mining* dapat berubah sesuai kecanggihan dan kemudahan penggunaan alat untuk menganalisis data sesuai dengan peningkatan jumlah peneliti untuk menerapkan *data mining* [20]. Informasi yang paling relevan relevan sebagai hasil dari *data mining* adalah mendapatkan hubungan di antara berbagai item. Saat ini, banyak perusahaan yang menggunakan teknik *Data mining* dalam membantu aktivitas perusahaan [21]. Hal penting menurut Novianti et al (2016) yang terkait dengan *data mining* adalah:

1. *Data mining* adalah pemrosesan otomatis data yang ada.
2. Data yang akan diolah memiliki bentuk data yang sangat besar.
3. Tujuan dari data mining adalah untuk menemukan hubungan atau pola yang dapat memberikan indikasi yang berguna.

2.1.3 *Klasifikasi*

Klasifikasi adalah proses pencarian sekumpulan model, pola atau fungsi yang menggambarkan dan membedakan objek data untuk dikelompokan ke dalam kelas tertentu dari sejumlah kelas yang tersedia [6]. Akurasi klasifikasi dihitung dengan menentukan *persentase instance* yang diberi label kelas yang benar [8]. Tujuan klasifikasi untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui [23].

2.1.4 Algoritma *Decision Tree*

Decision Tree (Pohon Keputusan) adalah pohon dimana setiap cabangnya menunjukkan pilihan diantara sejumlah alternatif pilihan yang ada, dan setiap daunnya menunjukkan keputusan yang dipilih. [24]. Decision tree biasa digunakan untuk mendapatkan informasi untuk tujuan pengambilan sebuah keputusan. Decision Tree digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel attribut x dan variabel target y dalam bentuk pohon [25]. Kelebihan *Decision Tree* yaitu Daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik, Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode pohon keputusan maka sampel diuji hanya berdasarkan criteria atau kelas tertentu, Fleksibel untuk memilih fitur dari node internal yang berbeda, fitur yang terpilih akan membedakan suatu kriteriadibandingkan kriteria yang lain dalam node yang sama. Kefleksibelan metode pohon keputusan ini meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang lebih konvensional, Dalam analisis multivarian, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut [26].

2.1.5 Algoritma SVM

Support Vector Machine (SVM) yaitu sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi – fungsi linier dalam sebuah fitur yang berdimensi tinggi dan dilatih dengan menggunakan algoritma pembelajaran yang didasarkan pada teori optimasi [27].Tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan dengan *Support Vector Machine (SVM)* sangat bergantung terhadap fungsi kernel dan parameter yang digunakan [28]. Berdasarkan dari karakteristiknya, metode *Support Vector Machine (SVM)* dibagi menjadi dua, yaitu *Support Vector Machine (SVM) Linier* dan *Support Vector Machine (SVM) Non-Linier*. *Support Vector Machine (SVM) linier* merupakan data yang dipisahkan secara *linier*, yaitu memisahkan kedua *class* pada *hyperplane* dengan *soft margin*. Sedangakan *Support Vector Machine* SVM Non-Linier yaitu menerapkan fungsi dari *kernel trick* terhadap ruang yang berdimensi tinggi [29].

2.1.6 Evaluasi Performance

Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. Evaluasi dimaksudkan untuk menguji model klasifikasi data mining untuk mengetahui kinerja sistem [30]. Salah satu metode untuk mengukur evaluasi performansi adalah *confusion matrix*. *Confusion matrix* adalah alat ukur berbentuk matrix yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi terhadap kelas dengan algoritma yang dipakai [31].

Tabel 1 Klasifikasi Performa Berdasarkan Nilai Akurasi
Sumber [31]

Rentang Nilai	Klasifikasi Performa
90%-100%	Sangat Baik
80%-90%	Baik
70%-80%	Cukup
60%-70%	Buruk
<=60%	Sangat Buruk

Evaluasi performansi dapat diukur dengan *precision* untuk menunjukkan tingkat ketepatan atau ketelitian dalam pengklasifikasian, *recall* berfungsi untuk mengukur proporsi positif aktual yang benar diidentifikasi, *F1-Measure* dapat didefinisikan sebagai alternatif dari metode akurasi yang diperoleh dari hasil perhitunganantara presisi dan *recall*, dan akurasi dapat didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual [31]. Masing-masing perhitungan memiliki persamaan:

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

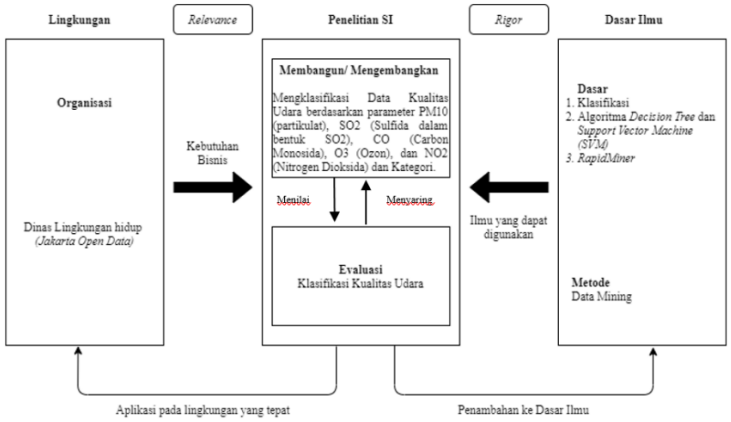
2.1.7 *RapidMiner*

RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan rediksi dalam memberikan wawasan kepada

pengguna sehingga dapat membuat keputusan yang paling baik [32]. *RapidMiner* berkembang sejak tahun 2001, sebelumnya disebut dengan nama *YALE (Yet Another Learning Environment)* [33]. Salah satu ciri utama *RapidMiner* adalah kemampuannya yang canggih untuk memprogram eksekusi alur kerja yang kompleks, semua dilakukan dalam antarmuka pengguna visual, tanpa perlu tradisional keterampilan pemrograman [34]. Software ini dikembangkan oleh Ralf Klinkenberg, Ingo Mierswa, serta Simon Fischer pada Unit Artificial Intelligence dari Technical University of Dortmund . *RapidMiner* merupakan platform analisis modern yang *meliputi data mining, mechine learning*, analisis prediktif, text mining dan analisis bisnis [35].

2.2 Metodologi Penelitian

Model konseptual digunakan untuk mengidentifikasi dan mengevaluasi masalah pada penelitian sistem informasi menggunakan konsep *design science* dan *behavior science*. *Design science* membahas penelitian melalui pembangunan dan evaluasi artefak yang dirancang untuk memenuhi kebutuhan bisnis lalu ada *behavior science* yang membahas tentang penelitian melalui pengembangan dan justification teori yang fenomena terkait dengan kebutuhan bisnis [36].



Gambar 1 Metode Konseptual

Pada gambar 1 merupakan model konseptual yang berisi tentang data-data yang dibutuhkan dalam proses penelitian yang digunakan untuk menggambarkan konsep permasalahan yang akan diteliti agar mudah dipahami. Berdasarkan Gambar 1 penelitian ini mengenai klasifikasi kualitas udara di DKI Jakarta. Lingkungan yang terdapat pada penelitian ini adalah DKI Jakarta karena penelitian ini menggunakan *dataset* yang didapat dari website *Jakarta Open Data* untuk melakukan klasifikasi Kualitas Udara di DKI Jakarta. Pada bagian Dasar Ilmu ditampilkan metode *data mining* untuk mengklasifikasi data kualitas udara berdasarkan bulan wilayah, parameter PM10, SO2, CO, O3, NO2. Untuk melakukan klasifikasi yang akurat maka algoritma yang digunakan yaitu *Algoritma Decision Tree* dan SVM.

3. Pembahasan

3.1 Pre-Processing

Dataset yang digunakan didapatkan dari *website Jakarta Open Data* pada tahun 2020. Data yang didapatkan dari website *Jakarta Open Data* dalam bentuk format *CSV (Comma Separated Values*) dalam bentuk file terpisah berdasarkan bulan dari bulan Januari hingga bulan Desember, selanjutnya data tersebut diubah menjadi format .xlsx atau excel dan terdapat 1830 baris data dengan atribut tanggal, wilayah, parameter PM10, CO2, CO, O3, NO2, Maximum, Critical. Kemudian data digabungkan menjadi satu server basis data yang sama. Data sampel yang digabungkan seperti pada Tabel 2:

Tabel 2 Penggabungan Data

Tanggal	Stasiun	PM10	SO2	CO	O3	NO2	Max	Critical	Categori
01/01/2020	DKI1 (Bunderan HI)	30	20	10	32	9	32	O3	Baik
02/01/2020	DKI1 (Bunderan HI)	27	22	12	29	8	29	O3	Baik
03/01/2020	DKI1 (Bunderan HI)	39	22	14	32	10	39	PM10	Baik
04/01/2020	DKI1 (Bunderan HI)	34	22	14	38	10	38	O3	Baik
05/01/2020	DKI1 (Bunderan HI)	35	22	12	31	9	35	PM10	Baik
06/01/2020	DKI1 (Bunderan HI)	46	23	16	32	9	46	PM10	Baik
07/01/2020	DKI1 (Bunderan HI)	37	23	26	33	11	37	PM10	Baik
08/01/2020	DKI1 (Bunderan HI)	41	26	20	30	11	41	PM10	Baik

Selanjutnya *data selection*, tujuan dari seleksi data unutk mengambil kolom yang diteliti. Pada penelitian ini mengambil sebanyak 6 kolom dengan 5 atribut parameter kualitas udara seperti PM10, SO2, CO, O3, NO2 dan kategori. Kemudian mauk ke tahap *cleaning data*, dimana pada tahap ini menghapus serta memperbaiki dan menemukan data yang rusak atau tidak akurat dengan melakukan pengaturan kembali pada data-data yang ada pada catatan, tabel, atau *database*. *Cleaning data* juga merupakan suatu proses analisis kualitas dari suatu data dengan cara mengubah, mengoreksi, atau menghapus data-data yang salah, tidak lengkap, tidak akurat, atau memiliki format yang salah dalam basis data guna menghasilkan data berkualitas tinggi. *Data cleansing* juga biasa disebut dengan *data cleaning* atau *data scrubbing*. Hasil dari *cleaning data* dari 1830 baris data menghasilan 1678 data. Sampel data yang perlu dilakukaukan *cleaning data* seperti pada tabel 3.

Tabel 3 Sampel *Cleaning Data*

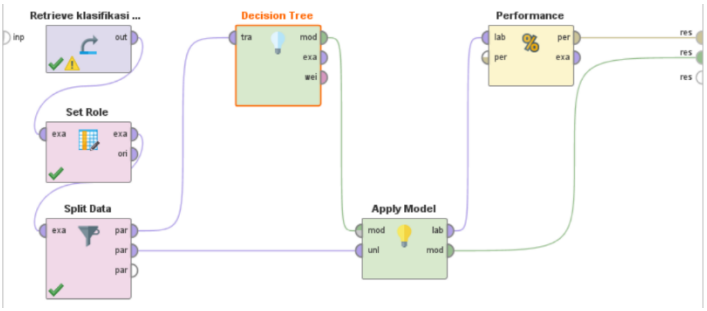
PM10	SO2	CO	O3	NO2	Max	Critical	Categori
---	35	13	101	4	101	O3	Tidak Sehat
---	33	16	---	4	33	SO2	Baik
81	33	19	---	4	81	PM10	Sedang
74	26	47	191	7	191	O3	Tidak Sehat
67	23	27	118	4	118	O3	Tidak Sehat
43	21	22	170	4	170	O3	Tidak Sehat
62	23	32	93	4	93	O3	Sedang
54	22	20	51	3	54	PM10	Sedang
30	20	11	68	3	68	O3	Sedang
34	20	22	80	4	80	O3	Sedang
58	19	34	85	4	85	O3	Sedang
39	---	---	---	---	39	PM10	Baik
32	19	---	65	---	65	O3	Sedang

Setelah mendapatkan hasil dari *cleaning data* maka didapatkan data siap olah dengan menggunakan tools *RapidMiner*. Berikut tabel *output* yang dihasilkan setelah melakukan *preprocessing data*. Dalam tabel tersebut terdapat 5 atribut parameter pengukuran kualitas udara sepeti PM10, SO2, CO, O3, NO2, dan atribut kategorinya. Data inilah yang selanjutnya akan dilakukan klasifikasi menggunakan *Tools RapidMiner*. Berikut tampilan hasil data seperti tabel 4 :

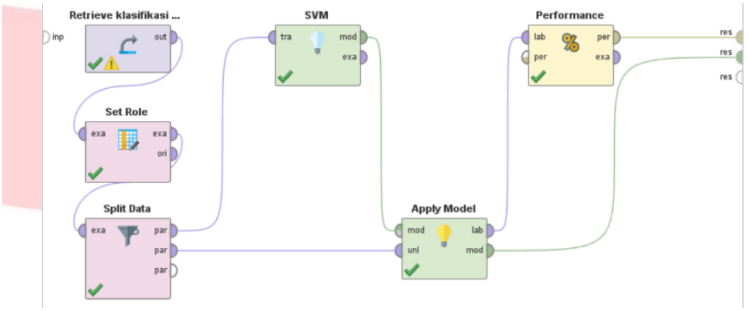
Tabel 4 Data Siap Olah

No	PM10	SO2	CO	O3	NO2	Categori
1	30	20	10	32	9	Baik
2	27	22	12	29	8	Baik
3	39	22	14	32	10	Baik
4	34	22	14	38	10	Baik
5	35	22	12	31	9	Baik
6	46	23	16	32	9	Baik
7	37	23	26	33	11	Baik
8	41	26	20	30	11	Baik
9	52	23	29	24	12	Sedang
10	24	24	18	25	8	Baik
11	34	31	25	23	8	Baik
12	27	23	9	33	4	Baik
13	33	26	12	36	8	Baik
14	34	28	13	27	7	Baik
15	29	22	13	36	8	Baik
…	…	…	…	…	…	…
1678	51	34	21	74	20	Sedang

3.2 Proses Klasifikasi

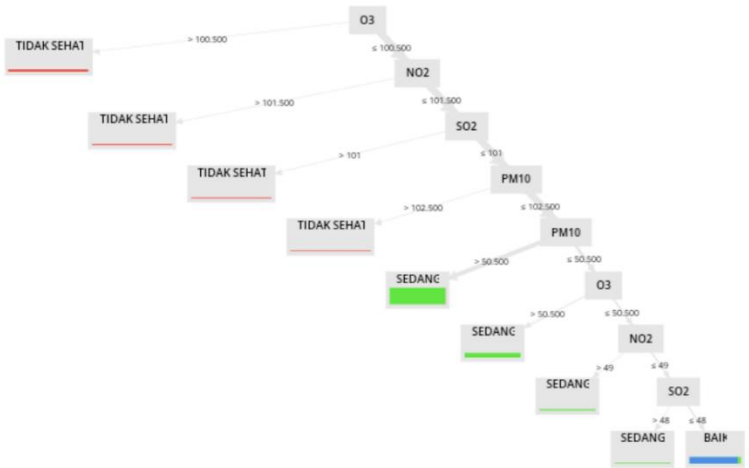


Gambar 2 Proses Klasifikasi menggunakan Algoritma *Decision Tree*



Gambar 3 Proses Klasifikasi menggunakan Algoritma SVM

Pada Gambar 2 dan gambar 3 adalah proses klasifikasi dengan menggunakan *algoritma Decision Tree* dan *algoritma SVM* yang dilakukan menggunakan *tools RapidMiner*. Data yang diolah sebanyak 1678 baris data, didapatkan setelah melakukan pengumpulan dan pembersihan data menggunakan *Microsoft Excel*. Selanjutnya untuk proses klasifikasi menggunakan *tools RapidMiner*. Proses pertama yaitu input dataset yang digunakan kemudan dihubungkan dengan *set role* untuk penandaan kolom sebagai label. Kemudian split data yang bertujuan untuk membagi menjadi 2 data yaitu data uji dan data latih. Peneliti menggunakan beberapa rasio pembagian data yang dilakukan pengujian terhadap model. Rasio pembagian data tang digunakan adalah rasio 60:40, 70;30, 80:20, dan 90:10. Rasio tersebut akan diuji satu persatu terhadap model kemudian dibandingkan dengan satu sama lain dengan hasil evalusai terbaik dan dihubungkan ke algoritma yang digunakan yaitu algoritma *Decision Tree* dan *algoritma SVM*. Selanjutnya dihubungkan ke *Apply Model* untuk menguji *data testing* dan *data training*. Selanjutnya pada operator *performance* akan melakukan evaluasi kinerja dan akan memberikan nilai dari klasifikasi.



Gambar 4 Pohon keputusan Algoritma Decision Tree

Berdasarkan Gambar 5 merupakan pohon keputusan dari algoritma *Decision Tree* untuk melakukan klasifikasi kualitas udara DKI Jakarta. Jika nilai O3 lebih besar dari 100,5 maka masuk kedalam kategori tidak sehat, tetapi jika nilai O3 lebih kecil sama dengan 100,5 dan nilai NO2 lebih besar dari 101,5 maka masuk kedalam kategori tidak sehat. Selanjutnya jika nilai NO2 lebih kecil sama dengan 101,5 tetapi nilai SO2 lebih besar dari 101 maka

masuk kedalam kategori tidak sehat. Jika nilai SO2 lebih kecil sama dengan 101 tetapi nilai PM10 lebih besar dari 102,5 maka masuk kedalam kategori tidak sehat. Jika nilai PM10 lebih kecil sama dengan 102,5 tetapi diatas 50,5 maka masuk kedalam kategori sedang. Jika nilai PM10 lebih kecil sama dengan 50,5 tetapi nilai O3 lebih besar dari 50,5 maka masuk kedalam kategori sedang, selanjutnya jika nilai O3 lebih kecil sama dengan 50,5 tetapi nilai NO2 lebih besar dari 49 maka masuk kedalam kategori sedang. Jika nilai NO2 lebih kecil sama dengan 49 tetapi nilai SO2 lebih besar 48 maka masuk kedalam kategori baik dan jika nilai SO2 lebih kecil dari 48 maka masuk ke kategori baik.

3.3 Evaluasi Performance Algoritma Decision Tree

Setelah dilakukannya proses klasifikasi pada *tools RapidMiner* menggunakan algoritma *Decision tree* mendapatkan hasil seperti tabel 5.

Tabel 5 Hasil Perbandingan nilai Akurasi algoritma *Decision Tree*

Rasio	Akurasi
60:40	98,51%
70:30	98,21%
80:20	98,51%
90:10	99,40%

Sehingga setelah dibandingkan dari hasil tersebut, maka peneliti menyimpulkan rasio untuk algoritma *Decision Tree* dengan akurasi terbaik dari perbandingan data training dan data adalah 90:10 yaitu sebanyak 99,40%.

3.1.1 *Evaluasi Performance Algoritma SVM*

Setelah dilakukannya proses klasifikasi pada *tools RapidMiner* menggunakan algoritma SVM mendapatkan hasil seperti tabel 6.

Tabel 6 Hasil Perbandingan nilai Akurasi algoritma *SVM*

Rasio	Akurasi
60:40	94,93%
70:30	93,85%
80:20	94,05%
90:10	94,01%

Sehingga setelah dibandingkan dari hasil tersebut, maka peneliti menyimpulkan rasio untuk algoritma SVM dengan akurasi terbaik dari perbandingan *data training* dan data adalah 60:40 yaitu sebanyak 94,93%.

Selanjutnya merupakan tabel perbandingan algoritma *Decision Tree* dan SVM disetiap rasio yang digunakan yaitu rasio 60:40, 70:30, 80:20, dan 90:10.

Tabel 7 Perbandingan rasio algoritma

Rasio	Akurasi	
	Decision Tree	SVM
60:40	98,51%	94,93%
70:30	98,21%	93,85%
80:20	98,51%	94,05%
90:10	99,40%	94,01%

Berdasarkan Tabel 7 nilai akurasi yang didapatkan berdasarkan nilai *weighted average* pada masing-masing *confusion matrix*. Pada rasio 60:40 algoritma *Decision Tree* memiliki nilai akurasi lebih tinggi sebesar 99,40% dibandingkan nilai akurasi alrotima SVM yaitu sebesar 94,93%, selanjutnya pada rasio 70:30 algoritma decision tree memiliki nilai akurasi lebih tinggi sebesar 98,21% dibandingkan nilai akurasi alrotima SVM yaitu sebesar 93,85%, selanjutnya pada rasio 80:20 algoritma decision tree memiliki nilai akurasi lebih tinggi sebesar 98,51% dibandingkan nilai akurasi alrotima SVM yaitu sebesar 94,05%, dan pada rasio 90:10 algoritma *decision tree* memiliki nilai akurasi lebih tinggi sebesar 99,40% dibandingkan nilai akurasi alrotima SVM yaitu sebesar 94,01%. Rasio terbaik pada algoritma *Decision Tree* dan SVM untuk melakukan klasifikasi kualitas udara DKI Jakarta yaitu pada algoritma *Decision Tree* menggunakan rasio 90:10 dengan nilai akurasi sebesar 99,40% dan pada algoritma SVM menggunakan rasio 60:40 dengan nilai akurasi sebesar 94,93%. Nilai akurasi yang didapatkan dinyatakan sangat baik berdasarkan rentang pada Klasifikasi Performa [31].

Maka dari itu peneliti menyimpulkan bahwa algoritma *Decision Tree* memiliki nilai akurasi yang lebih unggul daripada algoritma SVM untuk melakukan klasifikasi kualitas udara di DKI Jakarta, dibuktikan dengan nilai akurasi disetiap rasio yang digunakan oleh algoritma *Decision Tree* lebih tinggi daripada algoritma SVM.

Tabel 8 Hasil Evaluasi Performasi

Metode	Rasio	*Precision*	*Recall*	*Accuracy*
Decision Tree	90:10	99,02%	99,73%	99,40%
SVM	60:40	95,82%	88,89%	94,93%

Pada Tabel 8 merupakan nilai *precision, recall, F1-Measure* diambil berdasarkan nilai *weighted average* pada masing-masing *confusion matrix* berdasarkan rasio terbaik dari masing-masing algoritma, pada algoritma *Decision Tree* menggunakan rasio 90:10 dan pada algoritma SVM menggunakan rasio 60:40. Pada algoritma *Decision Tree* mendapatkan nilai *precision* sebesar 99,02%, *recall* 99,73%, *F1-Measure* 99,37% dan pada algoritma SVM mendapatkan nilai *precision* sebesar 95,82%, *recall* 88,89%, dan *F1-Measure* 92,22%. Berdasarkan hasil evaluasi performansi dan klasifikasi performa, Algoritma *Decision Tree* dan algoritma SVM memiliki *classifier* dengan klasifikasi performa sangat baik berdasarkan rentang nilai dari *confusion matrix* [31]. Algoritma *Decision Tree* memiliki performa yang lebih baik dibandingkan dengan algoritma SVM, baik dari nilai *Precision, Recall* dan *F1-Measure*.

4. Kesimpulan

Berdasarkan analisis data dan pembahasan yang telah dilakukan, dapat disimpulkan penelitian ini menggunakan Algoritma *Decision Tree* dan SVM untuk mengklasifikasi data kualitas udara DKI Jakarta yang didapatkan dari *website Jakarta Open Data* pada portal data.jakarta.go.id. Data yang digunakan adalah data harian dari bulan Januari hingga bulan Desember tahun 2020 dan terdapat 5 atribut parameter pengukuran kualitas udara PM10, SO2, CO, O3, NO2 dan atribut kategorinya. Dalam mengimplemestasi algoritma *Decision Tree* dan SVM menggunakan tools *RapidMiner.* Data yang diolah harus dilakukan proses *pre-procesing*. Hasil dari penerapan *data mining* untuk klasifikasi kualitas udara di DKI Jakarta yaitu algoritma *Decision Tree* memiliki performa yang lebih baik dibandingkan dengan algoritma SVM untuk melakukan klasifikasi kualitas udara di DKI Jakarta, baik dari nilai *Precision, Recall* dan *F1-Measure* dan akurasi. Pada penelitian ini, hasil yang didapatkan bahwa rasio terbaik untuk melakukan klasifikasi kualitas udara di DKI Jakarta dari masing-masing algoritma, pada algoritma *Decision Tree* menggunakan rasio 90:10 dan pada algoritma SVM menggunakan rasio 60:40 karena menghasilkan tingkat akurasi tertinggi dari rasio yang digunakan seperti 60:40, 70:30, 80:20, dan 90:10. Pada algoritma *Decision Tree* mendapatkan nilai *Precision* sebesar 99,02%, *Recall* 99,73%, *F1-Measure* 99,37%, Akurasi 99,40% dan pada algoritma SVM mendapatkan nilai *Precision* sebesar 95,82%, *Recall* 88,89%, *F1-Measure* 92,22% dan Akurasi 94,93%. Berdasarkan hasil evaluasi performansi dan klasifikasi performa, Algoritma *Decision Tree* dan algoritma SVM memiliki *classifier* dengan klasifikasi performa sangat baik berdasarkan rentang nilai dari *confusion matrix.*

Referensi :

[1] Y. M. Jamun, "Desain Aplikasi Pembelajaran Peta NTT Berbasis Multimedia," *J. Pendidik. dan Kebud. Missio*, vol. 8, no. 1, pp. 144–150, 2016.

[2] K. Prabowo and B. Muslim, *Penyehat Udara*, Pertama. Jakarta Selatan: Pusat Pendidikan Sumber Daya Manusia Kesehatan, 2018.

[3] H. Zheng, Y. Cheng, and H. Li, "Investigation of model ensemble for fine-grained air quality prediction," *China Commun.*, vol. 17, no. 7, pp. 207–223, 2020, doi: 10.23919/J.CC.2020.07.015.

[4] IQAir, "World Air Quality Report," *2020 World Air Qual. Rep.*, no. August, pp. 1–35, 2020, [Online]. Available: https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf.

[5] D. P. Sari, "Peningkatan Jumlah Kendaraan Bermotor di DKI Jakarta," 2020. https://jakarta.bps.go.id/indicator/17/786/1/jumlah-kendaraan-bermotor-menurut-jenis-kendaraan-unit-di-provinsi-dki-jakarta.html.

[6] I. G. Bendesa Subawa, "Teorema Bayes Ke Kelulusan," vol. 8, no. August, pp. 227–236, 2019.

[7] B. Sugiarto and R. Sustika, "Data classification for air quality on wireless sensor network monitoring system using decision tree algorithm," *Proc. - 2016 2nd Int. Conf. Sci. Technol. ICST 2016*, pp. 172–176, 2017, doi: 10.1109/ICSTC.2016.7877369.

[8] R. W. Gore and D. S. Deshpande, "An approach for classification of health risks based on air quality levels," *Proc. - 1st Int. Conf. Intell. Syst. Inf. Manag. ICISIM 2017*, vol. 2017-Janua, pp. 58–61, 2017, doi: 10.1109/ICISIM.2017.8122148.

[9] B. Liu, H. Wang, A. Binaykia, C. Fu, and B. Xiang, "Multi-level air quality classification in China using information gain and support vector machine hybrid model," *Nat. Environ. Pollut. Technol.*, vol. 18, no. 3, pp. 697–708, 2019.

[10] N. Giarsyani, A. F. Hidayatullah, and R. Rahmadi, "KLASIFIKASI KUALITAS UDARA DENGAN METODE SUPPORT VECTOR MACHINE," *Jire*, vol. 3, no. 1, pp. 48–57, 2020.

[11] E. Sutoyo, A. Almaarif, and others, "Educational Data Mining for Predicting Student Graduation Using the Na{\"\i}ve Bayes Classifier Algorithm," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 4, no.

1, pp. 95–101, 2020.
[12] E. Sutoyo and A. Musnansyah, "A Hybrid of Seasonal Autoregressive Integrated Moving Average (SARIMA) and Decision Tree for Drought Forecasting," in *Proceedings of the International Conference on Engineering and Information Technology for Sustainable Industry*, 2020, pp. 1–6.
[13] R. Rachmat and A. Ibrahim, "Decision Support System for Receiving Waste Retribution at the Housing and Sanitation Services Uses the Na{\"\i}ve Bayes Algorithm," *Bull. Comput. Sci. Electr. Eng.*, vol. 2, no. 1, 2021.
[14] Y. Alyousifi, M. Othman, R. Sokkalingam, I. Faye, and P. C. L. Silva, "Predicting daily air pollution index based on fuzzy time series markov chain model," *Symmetry (Basel).*, vol. 12, no. 2, pp. 1–18, 2020, doi: 10.3390/sym12020293.
[15] J. W. Koo, S. W. Wong, G. Selvachandran, H. V. Long, and L. H. Son, "Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models," *Air Qual. Atmos. Heal.*, vol. 13, no. 1, pp. 77–88, 2020, doi: 10.1007/s11869-019-00772-y.
[16] J. Wang, H. Li, and H. Lu, "Application of a novel early warning system based on fuzzy time series in urban air quality forecasting in China," *Appl. Soft Comput. J.*, vol. 71, pp. 783–799, 2018, doi: 10.1016/j.asoc.2018.07.030.
[17] A. Budiyono, "Pencemaran Udara : Dampak Pencemaran Udara Pada Lingkungan," *Dirgantara*, vol. 2, no. 1, pp. 21–27, 2010.
[18] N. Hossein Motlagh *et al.*, "Low-cost Air Quality Sensing Process: Validation by Indoor-Outdoor Measurements," pp. 223–228, 2020, doi: 10.1109/iciea48937.2020.9248348.
[19] T. Kasbe and R. S. Pippal, "Design of heart disease diagnosis system using fuzzy logic," *2017 Int. Conf. Energy, Commun. Data Anal. Soft Comput. ICECDS 2017*, pp. 3183–3187, 2018, doi: 10.1109/ICECDS.2017.8390044.
[20] L. C. Liñán and Á. A. J. Pérez, "Minería de datos educativos y análisis de datos sobre aprendizaje: Diferencias, parecidos y evolución en el tiempo," *RUSC Univ. Knowl. Soc. J.*, vol. 12, no. 3, pp. 98–112, 2015, doi: 10.7238/rusc.v12i3.2515.
[21] S. Bhise and P. S. Kale, "Effieient Algorithms to find Frequent Itemset Using Data Mining," *Int. Res. J. Eng. Technol.*, vol. 4, no. 6, pp. 2645–2648, 2017, [Online]. Available: https://irjet.net/archives/V4/i6/IRJET-V4I6664.pdf.
[22] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak)," *J. Coding, Sist. Komput. Untan*, vol. 04, no. 3, pp. 75–84, 2016.
[23] Ramadina, "Penerapan Fungsi Data Mining Klasifikasi Untuk Prediksi Masa Studi Mahasiswa Tepat Waktu Pada Sistem Informasi Akademik Perguruan Tinggi," *JUPITER (Jurnal Penelit. Ilmu dan Teknol. Komputer)*, vol. 7, no. 1, pp. 39–50, 2015.
[24] D. Setiawati, I. Taufik, J. Jumadi, and W. B. Zulfikar, "Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile," *J. Online Inform.*, vol. 1, no. 1, p. 24, 2016, doi: 10.15575/join.v1i1.7.
[25] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018, doi: 10.33480/pilar.v14i2.926.
[26] E. T. Susdarwono and A. Setiawan, "PENERAPAN TEORI KEPUTUSAN DALAM MODEL PENGAMBILAN KEPUTUSAN TERKAIT MASALAH EKONOMI PERTAHANANKONSEP POHON KEPUTUSAN," vol. 11, no. November, pp. 243–257, 2020.
[27] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, pp. 802–810, 2018.
[28] B. Sugara and A. Subekti, "Penerapan Support Vector Machine (Svm) Pada Small Dataset Untuk Deteksi Dini Gangguan Autisme," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 177–182, 2019, doi: 10.33480/pilar.v15i2.649.
[29] W. Purnami, A. M. Regresi, and L. Ordinal, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," *J. Sains Dan Seni Its*, vol. 1, no. 1, 2012.
[30] C. Anam and H. B. Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," *J. Ilm. Ilmu-Ilmu Tek.*, vol. 8, no. 1, pp. 13–19, 2018, [Online]. Available: https://ejournal.upm.ac.id/index.php/energy/article/view/111/449.
[31] O. Arifin and T. B. Sasongko, "Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2018*, vol. 6, no. 1, pp. 67–72, 2018.
[32] M. L. Sibuea and A. Safta, "Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustring," *Jurteksi*, vol. 4, no. 1, pp. 85–92, 2017, doi: 10.33330/jurteksi.v4i1.28.
[33] S. Fischer, R. Klinkenberg, I. Mierswa, and O. Ritthoff, "Yale: Yet Another Learning Environment -

Tutorial," no. December 2001, 2002, doi: 10.17877/DE290R-15309.
[34] M. Z. Jovanovic and M. Vukicevic, "Chapter 24 Using RapidMiner for Research :," no. October, 2014.
[35] A. Arimond, C. Kofler, and F. Shafait, "Distributed Pattern Recognition in RapidMiner," *RapidMiner Community Meet. Conf. RapidMiner Community Meet. Conf. (RCOMM-10), Sept. 13-16, Dortmund, Ger.*, 2010, [Online]. Available: http://www.dfki.de/web/forschung/publikationen/renameFileForDownload?filename=Arimond-DisPaRe-RCOMM10.pdf&file_id=uploads_782.
[36] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q. Manag. Inf. Syst.*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.