PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA

THESIS

Oleh:
AHMAD LATIF QOSIM
NIM. 19841007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2021

PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA

THESIS

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
AHMAD LATIF QOSIM
NIM. 19841007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2021

PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA

THESIS

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Magister Komputer (M.Kom)

Oleh:
AHMAD LATIF QOSIM
NIM. 19841007

PROGRAM STUDI MAGISTER INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2021

PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA

THESIS

Oleh :
AHMAD LATIF QOSIM
NIM. 19841007

Telah diperiksa dan disetujui untuk diuji:
Tanggal: 26 Desember 2021

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Dr. Usman Pagalay, M.Si
NIP. 19680519 200312 1 001	NIP. 19650414 200312 1

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

PERBANDINGAN METODE KLASIFIKASI SUPPORT VECTOR MACHINE (SVM) DAN NAIVE BAYES CLASSIFIER (NBC) UNTUK MENENTUKAN KUALITAS UDARA

THESIS

Oleh:
AHMAD LATIF QOSIM
NIM. 19841007

Telah Dipertahankan di Depan Dewan Penguji Thesis
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Magister Komputer (M.Kom)
Tanggal: 29 Desember 2021

Susunan Dewan Penguji			**Tanda Tangan**
Penguji Utama	:	Dr. Totok Chamidy, M.Kom NIP 19691222 200604 1 001	( )
Ketua Penguji	:	Dr. M. Faisal, M.T NIP 19740510 200501 1 007	( )
Sekertaris Penguji	:	Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji	:	Dr. Usman Pagalay, M.Si NIP. 19650414 200312 1 001	( )

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Cahyo Crysdian
NIP. 19740424 200901 1 008

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Ahmad Latif Qosim
NIM	: 19841007
Program Studi	: Magister Informatika
Fakultas	: Sains dan Teknologi

Menyatakan dengan sebenarnya bahwa Thesis yang saya tulis ini benar-banar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan Thesis ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 14 Desember 2021
Yang membuat pernyataan,

Ahmad Latif Qosim
NIM. 19841007

MOTTO

“Irhamu man fil ardli yarhamkum man fis sama”

Sayangilah semua yang ada di bumi, maka semua yang ada di langit akan menyayangimu.

(HR. Abu Dawud dan Timidzi)

"Tujuan mempelajari ilmu ialah untuk semakin mengetahui keagungan Allah SWT"

HALAMAN PERSEMBAHAN

Terimakasih kepada dosen pembimbing yang telah sabar mendampingi saya. Dosen Pembimbing yang telah mengarahkan saya dalam melakukan penulisan karya ilmiah ini.

Karya ini saya persembahkan untuk kedua orang tua saya. Orang Tua yang telah mendukung secara maksimal dalam penulisan karya ilmiah ini.

Terima Kasih pada keluarga saya yang telah mendukung saya untuk menyelesaikan karya tulis ilmiah ini.

KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Syukur alhamdulillah penulis hanturkan kehadirat Allah SWT yang telah melimpahkan Rahmat dan Hidayah-Nya, sehingga penulis dapat menyelesaikan studi di Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang sekaligus menyelesaikan Thesis ini dengan baik.

Selanjutnya penulis haturkan ucapan terima kasih seiring do'a dan harapan jazakumullah ahsanal jaza' kepada semua pihak yang telah membantu terselesaikannya Thesis ini. Ucapan terima kasih ini penulis sampaikan kepada:

1. Bapak Prof. Dr. Suhartono, M.Kom dan Bapak Dr. Usman Pagalay, M.Si selaku dosen pembimbing Thesis, yang telah banyak memberikan pengarahan dan pengalaman yang berharga.
2. Segenap sivitas akademika Program Studi Magister Informatika, terutama seluruh Bapak/ Ibu dosen, terima kasih atas segenap ilmu dan bimbingannya.
3. Ayahanda dan Ibunda tercinta yang senantiasa memberikan doa dan restunya kepada penulis dalam menuntut ilmu.
4. Kakak dan adik penulis yang selalu memberikan semangat kepada penulis untuk menyelesaikan Thesis ini.
5. Semua pihak yang ikut membantu dalam menyelesaikan Thesis ini baik berupa materiil maupun moril.

Penulis menyadari bahwa dalam penyusunan Thesis ini masih terdapat kekurangan dan penulis berharap semoga Thesis ini bisa memberikan manfaat kepada para pembaca khususnya bagi penulis secara pribadi. *Amin Ya Rabbal Alamin*.

Wassalamu'alaikum Wr. Wb.

Malang, 29 Desember 2021
Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Qosim, Ahmad Latif. 2021. **Perbandingan Metode Klasifikasi Support Vector Machine (SVM) Dan Naive Bayes Classifier (NBC) Untuk Menentukan Kualitas Udara**. Thesis Program Studi Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom.(II) Dr. Usman Pagalay, M.Si.

Kata kunci: klasifikasi, naïve bayes, svm, kualitas udara.

Udara sehat sangat penting bagi kehidupan manusia dan kelangsungan hidup organisme lain sehingga masyarakat dan pemerintah perlu mengetahui seberapa baik dan tercemarnya udara tersebut, maka diperlukan pengelompokan terhadap data kualitas udara menggunakan klasifikasi yaitu menganalisa dan menentukan keadaan udara ke dalam kelas tertentu menggunakan metode *supervised learning* untuk menghasilkan akurasi dalam pengklasifikasian, dalam penelitian ini dilakukan perbandingan antara klasifikasi dengan metode *Support Vector Machine (SVM)* dan metode *Naïve Bayes Classifier (NBC)*. Hasil pengujian yang dilakukan pada data kualitas udara Jakarta 2021, metode *Support Vector Machine (SVM*) memperoleh rata-rata akurasi sebesar 95.44%, Precision 87.05%, Recall 92.09% adapun metode *Naïve Bayes Classifier (NBC)* menghasilkan rata-rata akurasi sebesar 89.11%, Precision 90.44%, Recall 78.22%, sehingga *Support Vector Machine (SVM)* menghasilkan rata-rata akurasi lebih tinggi dibanding *Naïve Bayes Classifier (NBC)*. Hasil akurasi akan berbeda tergantung dari banyaknya data dan prosentase pembagian data training dan testing.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Permasalahan lingkungan terutama masalah pencemaran udara saat sekarang ini menjadi permasalahan seluruh dunia khususnya di negara Indonesia. Pencemaran udara memang perlu untuk mendapatkan perhatian serta penanganan baik dari pemerintah maupun masyarakat secara khusus, karena kalau dibiarkan dampaknya akan berpengaruh pada kesehatan mahluk hidup terutama manusia. Dalam mengatasi permasalahan pencemaran udara perlu adanya kerjasama dari berbagai pihak yaitu pemerintah, masyarakat, para pelaku industri dan transportasi terutama pada kota-kota besar seperti Jakarta. Pada tahun 2019 Dinas Lingkungan Hidup melaporkan bahwa adanya udara yang berkualitas buruk di kota Jakarta sangat dipengaruhi oleh banyaknya kegiatan industri, bertambahnya kendaraan bermotor dan kegiatan rumah tangga.

Allah SWT menerangkan dalam Al-Qur'an di ayat 41 surat Al-Rum tentang lingkungan.

ظَهَرَ ٱلْفَسَادُ فِي ٱلْبَرِّ وَٱلْبَحْرِ بِمَا كَسَبَتْ أَيْدِي ٱلنَّاسِ لِيُذِيقَهُم بَعْضَ ٱلَّذِي عَمِلُوا۟ لَعَلَّهُمْ يَرْجِعُونَ

"Telah nampak (nyata) kerusakan di darat dan di laut disebabkan perbuatan tangan manusia, supaya Allah merasakan kepada mereka sebagian dari (akibat) perbuatan mereka, agar mereka kembali (kejalan yang benar)".

Pada ayat di atas menunjukkan bahwa terjadinya kerusakan yang ada pada (ekologi) alam dan ekosistem disebabkan oleh prilaku manusia sehingga berakibat terjadi ketidak seimbangan dalam sistem kerja alam termasuk munculnya pencemaran udara.

Berdasarkan KEPMEN Negara Lingkungan Hidup NOMOR P.14 tahun 2020 Tentang Indeks Standar Pencemaran Udara yang terdiri dari Partikulat (PM10), Partikulat (PM25), Karbon monoksida (CO), Sulfur Dioksida (SO2), Nitrogen Dioksida (NO2), Ozon (O3), hidrokarbon (HC). Sedangkan Kategori nilai Rentang ISPU kualitas udara berdasarkan ketetapan pemerintah tersebut diantaranya kualitas udara kategori Baik, kualitas udara kategori Sedang, kualitas udara kategori Tidak Sehat, kualitas udara kategori Sangat Tidak Sehat, dan kualitas udara kategori Berbahaya. Penetapan status tingkat kualitas udara diharapkan akan memberikan kemudahan dan konsistensi informasi bagi masyarakat umum serta merupakan salah satu upaya pemerintah dalam memantau pencemaran udara. Dalam membantu pemerintah dan masyarakat untuk mengetahui kualitas udara yang layak ataupun tidak layak bagi kesehatan, maka diperlukan sebuah klasifikasi terhadap kualitas udara tersebut dengan menggunakan Teknologi Informasi. Bentuk kontribusi Teknologi Informasi dalam pengendalian pencemaran udara adalah dengan melakukan klasifikasi kualitas udara melalui pengolahan data menggunakan mechine learning.

Metode dalam pembelajaran mesin atau disebut juga dengan mechine learning banyak sekali untuk melakukan klasifikasi, diantaranya

metode Support Vector Machine, Naïve Bayes Classifier, KNN, Decision Tree Classifier, Random Forest Classifier dan ada beberapa metode yang selain disebutkan yang dapat digunakan untuk melakukan klasifikasi. Dari sekian banyak metode klasifikasi yang telah disebutkan di atas, tentunya perlu dilakukan perbandingan untuk mengetahui hasil akurasi mana yang lebih baik dalam pengklasifikasian.

Berdasarkan permasalahan tersebut maka disusunlah penelitian terhadap data kualitas udara yang diperoleh dari web resmi provinsi DKI Jakarta tahun 2021 dengan menggunakan model klasifikasi yaitu dengan menganalisa dan menentukan keadaan udara ke dalam kelas tertentu menggunakan metode supervised learning sesuai Standar Index Pencemar Udara yang telah ditetapkan oleh pemerintah. Metode yang akan diimpelentasikan pada penelitian ini adalah Support Vector Mechine & Naïve serta menganalisa hasil nilai rata-rata akurasi dari kedua metode tersebut.

1.2 Pernyataan Masalah

Dari latar belakang yang telah dibahas di atas maka disusunlah rumusan masalah yaitu Bagaimana model klasifikasi untuk menentukan kualitas udara dengan metode Support Vector Machine (SVM) dan Naïve Bayes Classifier (NBC), serta berapa tingkat akurasinya?

1.3 Tujuan Penelitian

Penelitian ini bertujuan membuat model klasifikasi kualitas udara dengan Support Vector Machine (SVM) dan Naïve Bayes Classifier (NBC), serta membandingkan rata-rata akurasi klasifikasi dari kedua metode.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini antara lain:

Dataset berasal dari data ISPU Provinsi DKI Jakarta Tahun 2021.

1.5 Manfaat Penelitian

Penelitian ini memiliki manfaat antara lain memberikan informasi tentang klasifikasi kualitas udara yang dapat digunakan untuk memberikan masukan kepada pemerintah dan masyarakat dalam pengendalian pencemaran udara.

1.6 Sistematika Penulisan

Sistematika penulisan Thesis ini sebagai berikut:

BAB I. PENDAHULUAN (*INTRODUCTION*)

Pada bab pendahuluan menguraikan tentang latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi dan sistematika penyusunan laporan. Uraian di bab ini memberikan gambaran kepada pembaca terkait maksud dan tujuan dalam penelitian yaitu tentang perbandingan metode Support Vector Machine (SVM) dan Naïve Bayes Classifier (NBC) dalam menentukan kualitas udara.

BAB II STUDI PUSTAKA (*LITERATURE REVIEW*)

Dalam tinjauan pustaka akan menguraikan penelitian teoritis dan ilmiah yang berkaitan dengan proses dan metode yang digunakan dalam penelitian yang diambil dari berbagai sumber referensi seperti buku, jurnal, dan sumber yang valid.

BAB III s/d *N*. METODOLOGI PENELITIAN (*RESEARCH METHODOLOGY*)

Pada metodologi penelitian ini terdapat menjelaskan tentang pola dan desain penelitian, bahan atau materi penelitian, alat, jalannya penelitian, dan analisis hasil penelitian**.**

BAB *N*+1. PEMBAHASAN (*DISCUSSION*)

Bab ini menjelaskan hasil yang diperoleh dalam bentuk penjelasan teoritis, kualitatif, kuantitatif atau statistik. Dalam hal ini hasil yang diperoleh dari ujicoba setiap metode untuk dikupas dan dibandingkan dengan hasil ujicoba metode lainnya serta melalui perspektif Informatika dan inspirasi Al-Qur’an dan Hadist.

BAB *N*+2. KESIMPULAN (*CONCLUSION*)

Berisi kesimpulan dari hasil penelitian dan saran terkait penelitian yang dilakukan untuk memperbaiki model yang dibangun dalam mencapai hasil yang lebih baik di masa yang akan datang.

BAB II

STUDI PUSTAKA (*LITERATURE REVIEW*)

2.1 Kualitas Udara

Pencemaran udara sangat berpengaruh terhadap kualitas udara. Diantara yang menyebabkan kualitas udara menjadi tercemar dipengaruhi oleh dua hal yaitu partikel dan polutan berupa gas. Partikel pencemar terdiri dari total suspende particulate yang memiliki diameter partikel mencapai 100µm disebut juga partikel tersuspensi total (TSP), sedangkan partikel lain seperti (PM10) memeiliki partikel lebih sedikit dari 10µm, dan lebih sedikit dari 2.5µm yaitu (PM2.5), adapun pencemar yang berupa gas-gas yaitu gas nitrogen dioksida (NO_2), gas karbon monoksida (CO), gas sulfur dioksida (SO_2), gas oksidan atau gas ozon permukaan (O_3) (Rita et al. 2016).

Berdasarkan Keputusan Menteri Negara Lingkungan Hidup NOMOR P.14 TH 2020 yang berisi tentang ISPU (Indeks Standar Pencemaran Udara) yang terdiri dari PM10, PM25, Karbon monoksida, Sulfur Dioksida, Nitrogen Dioksida, Ozon, hidrokarbon. Sedangkan kategori angka rentang ISPU kualitas udara berdasarkan ketetapan pemerintah tersebut diantaranya Baik, Sedang, Tidak Sehat, Sangat Tidak Sehat, dan Berbahaya (Peraturan Pemerintah RI 2020).

2.2 Mechine Learning

Mechine Learning merupakan metode ilmu data yang meberdayakan komputer untuk belajar dengan tidak perlu diprogram dengan aturan yang eksplisit dan memungkinkan pembuatan algoritma yang dapat

mempelajari dan membuat prediksi serta memanfaatkan akses yang lebih besar ke kumpulan data besar dan baru serta memiliki kemampuan untuk meningkatkan dan belajar dari pengalaman (Choy et al. 2018). Berdasarkan metode pembelajaran, jenis pembelajaran mesin (*machine learning*) dapat dibagi menjadi pembelajaran tidak terawasi (*unsupervised learning*), pembelajaran terawasi (*supervised learning*), pembelajaran semi-terbimbing (*semi supervised learning*), dan pembelajaran penguatan. Pembelajaran yang diawasi (*supervised learning*) adalah teknik pembelajaran mesin yang menggunakan kumpulan data (data pelatihan/training) yang diberi label sebagai (data yang ada labelnya) untuk melakukan pembelajaran mesin untuk memungkinkan mesin dapat mengenali label inputan berdasarkan karakteristik yang tersedia untuk memprediksi dan mengklasifikasikan lebih lanjut, sedangkan pembelajaran tanpa pengawasan adalah metode dengan menarik kesimpulan berdasarkan dataset yang merupakan input yang ditandai sebagai jawaban.

2.3 Klasifikasi

Klasifikasi merupakan proses mencari dan menganalisis suatu data dalam menentukan suatu model sehingga dapat mendefinisikan dan membedakan kelas data, kemudian digunakan dalam memperkirakan kelas pada suatu objek yang statusnya belum diketahui. Untuk membuat pemodelan dibutuhkan algoritma pembelajaran diantaranya Naïve Bayes Classifier, Support Vector Machine (SVM), KNN, Decision Tree Classifier, ANN dan lain-lain. Klasifikasi dipakai untuk memprediksi kelas berdasarkan data yang

berbentuk categorical (Anwar and Syafrullah 2016; Rakhmalia 2018). Klasifikasi umumnya berhubungan dengan prediksi dalam menentukan dan mengkategorikan nilai dalam kelas atribut dan memanfaatkan kelas data baru, mengklasifikasikan kategori dan mengklasifikasikan data, atau mengembangkan model berdasarkan data pelatihan. Beberapa klasifikasi sering digunakan antara lain persetujuan kredit, target pemasaran, diagnosa medis, dan analisis efektivitas keputusan (Simangunsong 2019).

2.4 Pengukuran Kinerja Klasifikasi

Dalam data mining, mengukur kinerja klasifikasi sangat penting untuk menentukan hasil kinerja dan menunjukkan seberapa akurat sistem tersebut dalam mengkategorikan data. Confusion matrix merupakan salah satu pendekatan untuk mengevaluasi kinerja metode klasifikasi yang berisi informasi tentang hasil klasifikasi. Ada empat macam informasi, yaitu true positive (TP), true negative (TN), false positive (FP) dan false negative (FN) (Febriantono, Herasmara, and Pangestu 2021). Untuk mengetahui kinerja dari sebuah model klasifikasi dengan 3 kelas kategori yaitu melalui *Confusion Matrix* 3x3 (Irmanda and Ria Astriratma 2020), seperti terlihat pada Tabel 2.1:

Tabel 2. 1 Confusion Matrix 3x3.

		Hasil Prediksi		
		A	**B**	**C**
Hasil Aktual	**A**	TA	FB1	FC1
	B	FA1	TB	FC2
	C	FA2	FB2	TC

Melihat Tabel 2.1 di atas mempunyai 2 bagian yaitu prediksi dan aktual. Bagian kolom aktual merupakan data yang diperoleh berdasarkan dari kenyataan. Pada bagian aktual dibagi menjadi tiga kelas: kelas A, kelas B, dan kelas C, adapun pada prediksi didasarkan pada klasifikasi yang dihasilkan oleh model klasifikasi yang mencakup tiga kelas: kelas A, kelas B, dan kelas C (Irmanda and Ria Astriratma 2020). Model pada kasus ini terdiri dari sembilan nilai antara lain : $TA, FA1, FA2, FB1, TB, FB2, FC1, FC2, dan\ TC$. TA adalah kelas A yang secara akurat dikategorikan benar, TB adalah kelas B yang diklasifikasikan dengan benar, dan TC adalah kelas C yang diklasifikasikan dengan benar. FA1 kelas B yang diklasifikasikan sebagai kelas A, FA2 merupakan kelas C yang diklasifikasikan sebagai kelas A. FB1 merupakan kelas A yang diklasifikasikan sebagai kelas B, FB2 merupakan kelas C diklasifikasikan sebagai kelas B. FC1 merupakan kelas A yang diklasifikasikan sebagai kelas C, FC2 adalah kelas B yang diklasifikasikan sebagai kelas C. Sedangkan untuk mengetahui nilai Akurasi, yaitu sebagai berikut:

$$Akurasi = \frac{T}{T + FA1 + FA2 + FB1 + FB2 + FC1 + FC2}\ x\ 100\%$$

Dimana T= penjumlahan TA+TB+TC. Sedangkan akurasi sendiri merupakan jumlah hasil pembagian semua data uji yang benar dengan data uji keseluruhan. Dalam mengevaluasi suatu model prediksi berhasil atau tidaknya maka perlu juga dilakukan model perhitungan Precision, recall (sensifitas) yaitu sebagai berikut :

$$Precision = \frac{TP}{TP + FP}\ x\ 100\%$$

Presisi dihitung dengan membandingkan jumlah data yang benar diprediksi positif dengan total data yang benar diprediksi positif. Pada keadaan ketika ada tiga klasifikasi akan diubah sesuai dengan Confusion Matrix pada Tabel 2.1, Kemudian di kelas A, TP nya adalah TA, dan FP nya adalah FA1+FA2, di kelas B, TP nya adalah TB, dan FP nya adalah FB1+FB2, pada kelas C, TP nya adalah TC, pada kelas C FP nya adalah $FC1 + FC2$. Sedangkan pada Recall ditulis dengan rumus sebagai berikut:

$$Recall = \frac{TP}{TP + FN}\ x\ 100\%$$

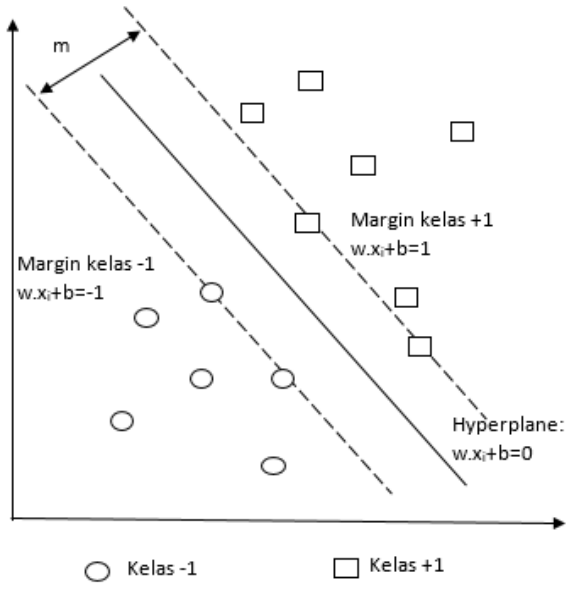
Dimana Recall atau sensitivitas dipakai diperoleh dengan membandingkan jumlah prediksi positif dengan jumlah keseluruhan kelas positif. Pada keadaan terdiri dari tiga kelas, disamakan dengan Confusion Matrix di Tabel 2.1, untuk kelas A TP nya adalah TA sedangkan FN nya adalah FB1+FC1. Pada kelas B TP nya adalah TB, sedangkan FN nya adalah FA1+FC2, di kelas C, TP nya adalah TC, FN adalah FA2+FB2.

2.5 Support Vector Machine

Support Vector Machine adalah pendekatan statistik yang dibangun dari teori pembelajaran yang menghasilkan hasil yang lebih baik dari pada metode tradisional (Anwar and Syafrullah 2016). Support Vector Machine (SVM) adalah bagian dari pendekatan klasifikasi yang menggunakan gagasan menempatkan hyperplane dengan margin terbesar terlebih dahulu. Hyperplane adalah garis yang menjadi pemisah data antar kelas atau kategori. Sedangkan

margin adalah jarak antara hyperplane dengan data terdekat di setiap kelas. Sedangkan Data yang terdekat dengan hyperplane dinamakan support vector (Irmanda and Ria Astriratma 2020). Ide awal dari SVM adalah untuk memaksimalkan batas hyperplane (margin hyperplane maksimum) dan bekerja dengan prinsip Structural Risk Minimization (SRM) dengan tujuan untuk menempatkan hyperplane secara optimal di ruang input yang memisahkan dua kelas (Anwar and Syafrullah 2016).

Pada SVM linear menyatakan bahwa tiap data training dideklarasikan dengan (x_i, y_i), dengan $i = 1,2, \dots N$, serta $x = \{x_i 1, x_i 2, \dots . x_{iq}\}^{\mathrm{T}}$ adalah data latih ke $-i$ yang merupakan atribut (fitur) set dan $y_i \in \{-1, +1\}$ menunjukkan label pada kelas. Pada Gambar 2.1 mencari hyperplane Klasifikasi Linear SVM .



Gambar 2. 1 Margin Hyperplane

Untuk mencari hyperplane pada klasifikasi Support Vector Machine Linear dilambangkan dalam rumus di bawah ini (Prasetyo 2012):

$$w.x_i + b = 0 \quad (1)$$

Pada rumus persamaan 1, dapat dijelaskan bahwa w serta b merupakan parameter model, dimana $w.x_i$ adalah inner-product dalam antara w dan x_i, sedangkan data x_i yang ada di dalam kelas -1 merupakan data yang sesuai dengan pertidaksamaan sebagai berikut (Prasetyo 2012):

$$w.x_i + b \leq -1 \quad (2)$$

Sedangkan data x_i yang ada di dalam kelas +1 adalah data yang sesuai dengan pertidaksamaan sebagai berikut (Prasetyo 2012):

$$w.x_i + b \leq +1 \quad (3)$$

Apabila ada data yang masuk dalam kelas -1 (seperti, x_a) yang terdapat pada hyperplane maka persamaan (1) akan terpenuhi. Pada data kelas -1 dinotasikan dengan

$$w.x_a + b = 0 \quad (4)$$

Sedangkan kelas +1 (seperti x_b) akan memenuhi persamaan

$$w.x_b + b = 0 \quad (5)$$

Dengan mengurangi persamaan (4) dengan (5) maka didapatkan

$$w.(x_b - x_a) \quad (6)$$

$x_b - x_a$ merupakan vector parallel pada posisi pada hyperplane dan diarahkan dari x_a ke x_b. Karena inner-product dalam bernilai nol, arah w harus tegak lurus terhadap hyperplane.

Dengan memberi label -1 pada kelas pertama dan +1 pada kelas kedua, sedangkan prediksi semua data uji akan memakai formula

$$y = \begin{cases} +1, \ jika\ w.z + b > 0 \\ -1, \ jika\ w.z + b < 0 \end{cases} \quad (7)$$

Pada Gambar 2.1 menyatakan bahwa hyperplane pada kelas -1 (garis yang terputus-putus) merupakan support vector yang memiliki persamaan

$$w.x_a + b = - \quad (8)$$

Sedangkan hyperplane pada kelas +1 (pada garis yang terputus-putus) menggunakan persamaan

$$w.x_a + b = +1 \quad (9)$$

Maka dari itu, margin dihitung dengan mengurangi persamaan (8) dengan (9) dan didapatkan persamaan

$$w.(x_b - x_a) = 2 \quad (10)$$

Jarak antara dua hyperplanes dari dua kelas menentukan margin hyperplane. Notasinya sebagai berikut

$$\|w\| \text{x } d = 2 \text{ atau } d = \frac{2}{\|w\|} \quad (11)$$

Klasifikasi pada kelas data SVM di persamaan (1) dan (2) dapat dijadikan satu dengan notasi sebagai berikut

$$(y_i(w.x_i + b) \geq 1, i = 1,2, \dots N. \quad (12)$$

Jarak antara hyperplane dan titik data terdekat dioptimalkan untuk menghasilkan margin yang ideal. Jarak tersebut dapat dirumuskan pada persamaan (11) dimana $\|w\|$ merupakan vector bobot w. Masalah tersebut

kemudian ditransformasikan menjadi masalah pemrograman kuadratik (QP) dengan meminimalkan invers persamaan (11), $\frac{1}{2}\|w\|^2$ dengan syarat $(y_i(w.x_i + b) \geq 1, i = 1,2, \ldots N.$

2.6 SVM Kelas Jamak

SVM awalnya dirancang untuk memecahkan masalah klasifikasi dua kelas, namun berkembang untuk menangani klasifikasi multi-kelas. Ada beberapa hyperplane dalam kategorisasi kasus multiclass. Salah satu dari strategi yang dipakai antara lain one against all (SLA). Pada SLA pada kasus klasifikasi k-class, temukan k hyperplanes di mana k adalah jumlah kelas dan ρ adalah hyperplane. Dalam metode ini (ℓ) diuji dengan semua data dari kelas dengan label +1, dan semua data dari kelas lain dengan label -1. Konsep dalam SLA misalnya dalam kasus ada tiga kelas, nilai 1, 2, dan 3. Jika (1) diuji, semua data di kelas 1 diberi label +1 dan data dari kelas 2 dan 3 diberi label -1. Dalam (2), semua data di kelas 2 ditugaskan label +1 dan data dari kelas 1 dan 3 diberi label -1. Begitu juga untuk (3), semua data di kelas 3 diberi label +1 dan data dari kelas 1 dan 2 diberi label -1. Kemudian cari hyperplane dengan algoritma SVM dua kelas. Kemudian, pada tiap-tiap kelas itu akan ditemukan hyperplane. Kelas data baru x kemudian dihitung menggunakan nilai terbesar hyperplane. (Anwar and Syafrullah 2016). SVM sekarang berkembang kepada masalah klasifikasi multi-kelas atau kelas jamak. SVM multi-kelas dapat digunakan sebagai pengklasifikasi multi-kelas dengan membuat beberapa pengklasifikasi biner. Metode ini bekerja mirip dengan ide regresi logistik multinomial, di mana kita membangun model logistik untuk setiap pasangan kelas dengan fungsi

dasar. Sepanjang baris yang sama, kita dapat membuat satu set SVM biner untuk melakukan klasifikasi multi-kelas (Ramasubramanian and Singh 2019). Langkah-langkah untuk mengimplementasikan ini adalah membuat terlebih dahulu pengklasifikasi biner (Antara satu kelas dan kelas lainnya, antara setiap pasangan kelas (semua kemungkinan pasangan)), Untuk setiap kasus baru, pengklasifikasi SVM mengadopsi pemenang-mengambil-semua strategi, di mana kelas dengan output tertinggi sebagai pemenang (Ramasubramanian and Singh 2019).

2.7 Naïve Bayes

Metode Nave Bayes adalah algoritma data mining yang mengklasifikasikan data menggunakan teori Bayes. Ketika berhadapan dengan kumpulan data yang sangat besar, Nave Bayes dapat diandalkan dan dapat mentolerir data yang tidak relevan. Vektor input yang menyimpan data dilambangkan dengan simbol X, dan label kelas dilambangkan dengan simbol Y (Helmi 2021). Berikut ini adalah persamaan teorema Bayes:

(13)

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

Penjelasan:

X = Label kelas yang belum dikenal.

Y = Hipotesis data X pada suatu terpisan atau kelas tersendiri

P(Y|X) = Kemungkinan hipotesis Y berdasar kondisi X (posteriori probabilitas).

P(Y) = Kemungkinan hipotesis Y (prior probabilitas)

P(X|Y) = Kemungkinan X berdasarkan kondisi pada hipotesis Y

P(X) =Kemungkinan X.

Hipotesis pada teorema Bayes adalah label pada kelas merupakan tujuan pemetaan dalam klasifikasi, dan yang menjadi bukti adalah atribut-atribut yang dimasukkan ke model klasifikasi. Nave Bayes direpresentasikan sebagai P(Y|X) jika X adalah vektor input dengan karakteristik dan Y adalah label kelas. Notasi ini menunjukkan bahwa setelah mengamati atribut X, probabilitas label kelas Y dihitung. Ini juga dikenal sebagai probabilitas akhir (probabilitas posterior) dari Y, pada P(Y) dikenal sebagai probabilitas awal (probabilitas sebelumnya) dari Y (Prasetyo 2012). Perlu diketahui bahwa untuk menggambarkan pendekatan Naive Bayes, diperlukan prosedur klasifikasi, yang melibatkan sejumlah instruksi untuk mengidentifikasi kelas mana yang sesuai untuk sampel yang sedang diperiksa. (Lishania, Goejantoro, and Nasution 2019). Sebagai hasilnya, pendekatan Nave Bayes dimodifikasi sebagai berikut:

$$P\left(Y_j \middle| X_{1,\dots,} X_n\right) = \frac{P(y)P(X_1, \dots., X_n|Y_j)}{P(X_1, \dots., X_n)} \quad (14)$$

Variable Y_j menunjukkan kelas, dan variabel $X_1, \dots., X_n$ menggambarkan ciri-ciri petunjuk yang pakai dalam melakukan klasifikasi. Dengan demikian formula tersebut menunjukkan peluang untuk ditempatkan pada sampel karakteristik tertentu pada kelas Y_j (Posterior) adalah kemungkinan munculnya

kelas Y_j (sebelum masuknya sampel, seringkali disebut prior), dikalikan dengan peluang munculnya atribut sampel di kelas Y_j (disebut likelihood), dibagi dengan peluang global dari atribut sampel yang terjadi (disebut evidence). Sehingga, rumus di atas juga dapat dinyatakan sebagai berikut (Helmi 2021):

$$Postorior = \frac{prior\ x\ likelihood}{evidence}$$

Dalam satu sampel, nilai bukti selalu sama untuk setiap kelas. Nilai posterior dibandingkan dengan posterior di kelas lain untuk menentukan sampel kelas mana yang akan ditugaskan ke suatu kelas.

Hubungan antara Naïve Bayes dan klasifikasi, korelasi hipotesis, dan bukti dengan klasifikasi dimana pada teorema Bayes bahwa hipotesis merupakan label kelas yang merupakan target pemetaan klasifikasi, sedangkan bukti merupakan fitur yang menjadi inputan untuk mengklasifikasi. Apabila X adalah vector masukan yang isinya fitur, sedangkan Y merupakan label kelas. Naïve Bayes disimbulkan P(Y|X), notasi ini menunjukkan bahwa propabilitas label pada kelas Y diperoleh setelah fitur-fitur X yang telah diamati. Notasi disebut juga dengan probabilitas akhir (*posterior probability*) bagi Y, sedangkan pada P (Y) disebut dengan probabilitas awal (*prior probability*) Y. Sewaktu proses training harus melalui pembelajaran probabilitas akhir (P(Y|X)) tergantung pada informasi yang dikumpulkan dari data pelatihan pada model untuk setiap X dan Y. Data uji x 'dapat dikategorikan menggunakan model ini dengan mengoptimalkan nilai p (y' | x ') yang didapat dengan menemukan nilai y'.

Formula pada Naïve Bayes untuk klasifikasi dinotasikan seperti di bawah ini (Prasetyo 2012)

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y))}{P(X)} \quad (15)$$

P(Y|X) merupakan peluang data vector X pada kelas Y, dimana P(Y) adalah probabilitas awal kelas Y.$\prod_{i=1}^{q} P(X_i|Y))$ sedangkan probabilitas independen kelas Y dari semua fitur dalam vector X. Nilai P(X) selalu tetap sehingga dalam perhitungan ramalan nanti kita hitung hanya bagian P(Y).$\prod_{i=1}^{q} P(X_i|Y))$ dengan menyeleksi yang terbesar sebagai kelas yang dipilih sebagai hasil prediksi. Adapun probabilitas independen $\prod_{i=1}^{q} P(X_i|Y)$ Pengaruh semua karakteristik data pada setiap kelas Y yang ditunjukkan oleh

$$P(X|Y = y) = \prod_{i=1}^{q} P(X_i|Y = y) \quad (16)$$

Pada tiap set fitur X=$\{X_1, X_2, X_3, \ldots., X_q\}$ terdiri atas q atribut (q dimensi).

Secara umum, Bayes mudah dihitung untuk karakteristik kategorikal, tetapi fitur numerik (berkelanjutan) memerlukan perlakuan khusus sebelum dimasukkan ke dalam Naive Bayes, yaitu dengan cara sebagai berikut (Prasetyo 2012):

1. Lakukan diskritisasi pada setiap fitur kontinu dan ganti nilai fitur kontinu dengan nilai interval diskrit. Hal ini dilakukan dengan mengubah fitur kontinu menjadi fitur ordinal.

2. Untuk fitur kontinu, dengan asumsi jenis distribusi probabilitas tertentu dan penghitungan parameter distribusi menggunakan data pelatihan. Biasanya, distribusi Gaussian digunakan untuk menggambarkan probabilitas kondisional karakteristik kontinu dalam kelas $P(X_i|Y)$, Sementar itu distribusi Gaussian ditentukan oleh dua parameter : mean, μ, dan varian, σ^2. Pada tiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i adalah

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2}\pi\sigma_{ij}} exp \frac{(x_i-\mu_{ij})^2}{2\pi\sigma_{ij}^2} \quad (17)$$

keterangan:

P = Kemungkinan
P_i = Variabel ke i
x_i = Nilai Variabel ke i
Y = Kelas yang sedang dicari
y_i = Sub kelas Y yang dicari
μ_{ij} =Rata-rata sampel dari data latih pada y_j
$\pi\sigma_{ij}$ =Jenis sampel data latih.

Parameter μ_{ij} dapat diperoleh dari rata-rata sample X_i ($\bar{x}$) pada semua data training yang terkait dengan kelas y_j sementara itu σ_{ij}^2 bisa dihitung berdasarkan variasi sample (S^2) data training.

2.8 Indeks Standar Pencemaran Udara

Indeks Standar Pencemaran Udara yang biasa disebut ISPU menunjukkan angka status kualitas udara daerah tertentu berdasar pengaruhnya

bagi kesehatan manusia, estetika, dan organisme. Stasiun Pemantau Kualitas Udara ambien (disingkat SPKUA) adalah seperangkat peralatan untuk memantau kualitas udara ambien yang berjalan secara kontinyu serta dapat memantau data secara langsung. Indeks polusi udara standar mencakup parameter berikut:

a. Parameter partikulat (PM10);

b. Parameter partikulat (PM2.5);

c. Parameter karbon monoksida (CO);

d. Parameter nitrogen dioksida (NO2);

e. Parameter sulfur dioksida (SO2);

f. Parameter ozon (O3); dan

g. Parameter hidrokarbon (HC).

dihitung menggunakan nilai berikut:

a. Nilai ISPU batas atas;

b. Nilai ISPU batas bawah;

c. Nilai ambien batas atas;

d. Nilai ambien batas bawah; dan

e. Nilai konsentrasi ambien hasil pengukuran.

Tabel 2. 2 Konversi Nilai Konsentrasi ISPU

ISPU	24 Jam partikulat (PM10) µg/m^3	24 Jam partikulat (PM2.5) µg /m^3	24 Jam sulfur dioksida (SO2) µg /m3	24 Jam karbon monoksida (CO) µg /m^3	24 Jam ozon (O3) µg /m^3	24 jam nitrogen dioksida (NO2) µg /m^3	24 Jam hidrokarbon (HC) µg /m^3
0 - 50	50	15,5	52	4000	120	80	45
51 - 100	150	55,4	180	8000	235	200	100
101 - 200	350	150,4	400	15000	400	1130	215
201 - 300	420	250,4	800	30000	800	2260	432
>300	500	500	1200	45000	1000	3000	648

Pada Tabel 2.2 menjelasakan tentang detail nilai parameter ISPU dan nilai konversinya terhadap kategari ISPU dimana jangka waktu pengamatannya 24 jam secara kontinyu dan hasilnya disampaikan setiap jam dengan waktu 24 jam serta hasil kalkulasinya diambil nilai paling tinggi pada ISPU.

Tabel 2. 3 Angka Rentang ISPU berdasarkan Kategori

Kategori	Status Warna	Angka Rentang
Baik	Hijau	1 – 50
Sedang	Biru	51 – 100
Tidak Sehat	Kuning	101 – 200
Sangat Tidak Sehat	Merah	201 - 300
Berbahaya	Hitam	>=301

Pada Tabel 2.3 adalah Detail rentang nilai parameter ISPU, status warna dan konversinya terhadap kategari ISPU. Sedangkan persamaan di bawah ini digunakan untuk mengkalkulasi Indeks Standar Pencemar Udara (ISPU). Berikut ini adalah prosedur penghitungan persamaan:

$$I = \frac{(I_a - I_b)}{(X_a - X_b)} (X_x - X_b) + I_b$$

I = Nilai ISPU yang dihitung

Ia = Nilai batas atas ISPU

Ib = Nilai batas bawah ISPU

Xa = Batas atas konsentrasi ambien (µg/m3)

Xb = Batas bawah Konsentrasi ambien (µg/m3)

Xx = Konsentrasi ambien terukur (µg/m3)

2.9 Penelitian Terkait

Penelitian yang telah dilakukan oleh (Hermawan 2019) melakukan prediksi kualitas udara Jakarta pada data ISPU tahun 2017 dan 2018 menggunakan metode Suppor Vector Machine dengan 5 jenis parameter diantaranya parameter PM10, parameter SO2, parameter CO, parameter O3, dan parameter NO2 dengan nilai akurasi paling tinggi yaitu 96.03%.

Hasil penelitian dilakukan oleh (Handayani et al. 2021) menggunakan pendekatan Support Vector Machine untuk mengklasifikasikan kualitas udara dengan parameter CO, parameter CO2, parameter HC, parameter debu/PM10 dan temperatur. Berdasarkan hasil uji coba yang dilakukan menghasilkan nilai akurasi pada klasifikasi sebesar 95,02%.

Penelitian yang dilakukan (Syihabuddin Azmil Umri 2021) yaitu membandingkan beberapa metode klasifikasi diantaranya *Support Vector Machine, K-Nearest Neighbors, Naive Bayes, Neural Network dan Decission Tree* dengan menguji 5 parameter pencemar diantaranya parameter monoksida (CO), parameter sulfur dioksida (SO2), parameter nitrogen dioksida (NO2), parameter ozon permukaan (O3), dan parameter partikel debu (PM10) dan

memperoleh akurasi tertinggi pada metode *Decission Tree* dengan akurasi sebesar 99.80% sedangkan metode pembanding menghasilkan akurasi 90%.

Penelitian yang dilakukan (Nurdin 2020) yaitu menggunakan pendekatan Support Vector Machine, mengklasifikasikan kualitas udara menggunakan pembacaan sensor dengan beberapa parameter kualitas udara seperti CO, CO2, HC, PM10, suhu, dan kelembaban. Hasil uji coba yang dilakukan menghasilkan nilai klasifikasi terbaik yaitu 99,33%.

Penelitian dilakukan (Wicahyo, Pudoli, and Kusumaningsih 2021) yaitu melakukan klasifikasi terhadap pengaruh pencemar udara dengan beberapa parameter diantaranya Partikulat (PM10), Sulfur Dioksida (SO2), Karbon Monoksida (CO), Ozon (O3), dan Nitrogen Dioksida (NO2) menggunakan metode Naive Bayes dari dataset kualitas udara Jakarta mulai bulan Januari 2018 sampai bulan Juni tahun 2020. Dari hasil klasifikasi yang dilakukan mendapatkan nilai akurasi sebesar 96% dengan data testing 129 dan data training sebesar 4000 data.

Penelitian yang dilakukan (Supriyatna and Mustika 2018) yaitu membandingkan hasil prediksi data imunoterapi pada kutil menggunakan algoritma Naive Bayes dan SVM sebanyak 90 record dan tujuh faktor, antara lain jenis kelamin, usia pasien, lama mengidap, jumlah, jenis, ukuran, dan keparahan. Berdasarkan confusion matrix prediksi menghsilkan akurasi yaitu 80% untuk SVM dan 100% untuk Naïve Bayes.

2.10 Menjaga Kualitas Udara menurut pandangan Al-Qur'an

Kualitas udara dijelaskan dalam surat al-Baqarah /2:164 sebagai berikut:

إِنَّ فِي خَلْقِ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ وَٱخْتِلَٰفِ ٱلَّيْلِ وَٱلنَّهَارِ وَٱلْفُلْكِ ٱلَّتِي تَجْرِي فِي ٱلْبَحْرِ بِمَا يَنفَعُ ٱلنَّاسَ وَمَآ أَنزَلَ ٱللَّهُ مِنَ ٱلسَّمَآءِ مِن مَّآءٍ فَأَحْيَا بِهِ ٱلْأَرْضَ بَعْدَ مَوْتِهَا وَبَثَّ فِيهَا مِن كُلِّ دَآبَّةٍ وَتَصْرِيفِ ٱلرِّيَٰحِ وَٱلسَّحَابِ ٱلْمُسَخَّرِ بَيْنَ ٱلسَّمَآءِ وَٱلْأَرْضِ لَأٓيَٰتٍ لِّقَوْمٍ يَعْقِلُونَ

"Sesungguhnya dalam penciptaan langit dan bumi, silih bergantinya malam dan siang, bahtera yang berlayar di laut membawa apa yang berguna bagi manusia, dan apa yang Allah turunkan dari langit berupa air, lalu dengan air itu Dia hidupkan bumi sesudah mati (kering)-nya dan Dia sebarkan di bumi itu segala jenis hewan, dan pengisaran angin dan awan yang dikendalikan antara langit dan bumi; sungguh (terdapat) tanda-tanda (keesaan dan kebesaran Allah) bagi kaum yang memikirkan".

Dalam ayat lain yang membahas tentang lingkungan, dalam surat Al-Rum ayat 41 Allah SWT berfirman:

ظَهَرَ ٱلْفَسَادُ فِي ٱلْبَرِّ وَٱلْبَحْرِ بِمَا كَسَبَتْ أَيْدِي ٱلنَّاسِ لِيُذِيقَهُم بَعْضَ ٱلَّذِي عَمِلُوا۟ لَعَلَّهُمْ يَرْجِعُونَ

"*Telah nampak (nyata) kerusakan di darat dan di laut disebabkan perbuatan tangan manusia, supaya Allah merasakan kepada mereka sebagian dari (akibat) perbuatan mereka, agar mereka kembali (kejalan yang benar)".*

Berdasarkan surat Al-Rum ayat 41 menunjukkan bahwa kerusakan yang ada pada alam (ekologi) dan ekosistem akibat dari perbuatan manusia sehingga

terjadi ketidak seimbangan dalam sistem kerja alam termasuk di dalamnya adalah pencemaran udara.

Dan tugas manusia adalah untuk menjaga lingkungan ini, dalam surat Al-Qur'an al-Baqarah/2: 30 Firman Allah sebagai berikut:

وَإِذْ قَالَ رَبُّكَ لِلْمَلَٰٓئِكَةِ إِنِّي جَاعِلٌ فِي ٱلْأَرْضِ خَلِيفَةً قَالُوٓاْ أَتَجْعَلُ فِيهَا مَن يُفْسِدُ فِيهَا وَيَسْفِكُ ٱلدِّمَآءَ وَنَحْنُ نُسَبِّحُ بِحَمْدِكَ وَنُقَدِّسُ لَكَ ۖ قَالَ إِنِّيٓ أَعْلَمُ مَا لَا تَعْلَمُونَ

"Ingatlah ketika Tuhanmu berfirman kepada para Malaikat: "Sesungguhnya Aku hendak menjadikan seorang khalifah di muka bumi". Mereka berkata: "Mengapa Engkau hendak menjadikan (khalifah) di bumi itu orang yang akan membuat kerusakan padanya dan menumpahkan darah, padahal kami senantiasa bertasbih dengan memuji Engkau dan mensucikan Engkau?" Tuhan berfirman: "Sesungguhnya Aku mengetahui apa yang tidak kamu ketahui".

Sebagai khalifah, manusia bertugas menegakkan ketertiban dan kedamaian di atas bumi memiliki tanggung jawab untuk mengarahkan dirinya sendiri dan secara efektif mengendalikan lingkungannya. Sehingga, dalam pandangan agama, menjaga lingkungan merupakan tugas utama umat manusia dalam menjamin eksistensinya.

Dalam ayat lain dalam al-A'raf/7:56, Allah menyatakan:

وَلَا تُفْسِدُواْ فِي ٱلْأَرْضِ بَعْدَ إِصْلَٰحِهَا وَٱدْعُوهُ خَوْفًا وَطَمَعًا ۚ إِنَّ رَحْمَتَ ٱللَّهِ قَرِيبٌ مِّنَ ٱلْمُحْسِنِينَ

Dan janganlah kamu membuat kerusakan di muka bumi, sesudah (Allah) memperbaikinya dan berdoalah kepada-Nya dengan rasa takut (tidak akan diterima) dan harapan (akan dikabulkan). Sesungguhnya rahmat Allah amat dekat kepada orang-orang yang berbuat baik.

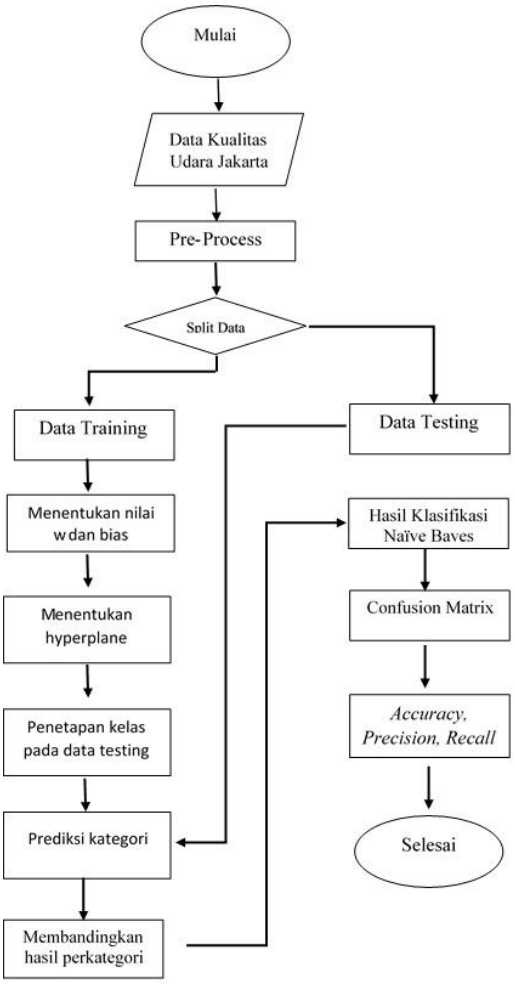
Ayat ini menjelaskan bahwa segala sesuatu yang Allah berikan pada manusia harus dipelihara menurut takaran yang telah Allah sediakan. Berdasarkan posisi manusia sebagai khalifah di bumi ini, maka manuasia memiliki kewajiban dan tanggung jawab terhadap lingkungan sebagai kemaslahatan, yang diberikan Allah yang patut disyukuri, dilestarikan, dan dijunjung tinggi serta kesadaran lingkungan harus ditumbuhkan (Perspektif and An 2019; Qomarullah n.d.; Rahmasari 2017).

BAB III

SUPPORT VECTOR MACHINE (SVM)

3.1 Desain Metode Support Vector Machine

Pada desain metode Support Vector Machine terdapat uraian tentang pola dan rancangan klasifikasi Support Vector Machine dapat dilihat pada Gambar 3.1:



Gambar 3. 1 Desain Metode Support Vector Machine

Pada Gambar 3.1 menjelaskan tentang rancangan desain metode Support Vector Machine dengan tahapan-tahapan diantaranya:

- **Seleksi Data/ Eksplorasi Data:** Pada desain penelitian dengan metode SVM yang dilakukan seperti dijelaskan pada Gambar 3.1, yaitu dimulai dari menganalisa data yang akan digunakan dalam proses klasifikasi SVM yaitu dataset yang berisi tentang Indeks Standar Pencemaran Udara (ISPU) dihitung menggunakan data dari lima stasiun pemantauan kualitas udara (SPKU) di Provinsi DKI Jakarta tahun 2021, data ini diambil pada website resmi https://data.jakarta.go.id/. Data kualitas udara ini memiliki variable sebagai berikut:

1.	TANGGAL	:	Tanggal pengukuran kualitas udara
2.	STASIUN	:	Lokasi pengukuran di stasiun
3.	PM10	:	Partikulat salah satu parameter yang diukur
4.	PM25	:	Partikulat salah satu parameter yang diukur
5.	SO2	:	Sulfida (dalam bentuk SO2) salah satu parameter yang diukur
6.	CO	:	Carbon Monoksida salah satu parameter yand diukur
7.	O3	:	Ozon salah satu parameter yang diukur
8.	NO2	:	NItrogen dioksida salah satu parameter yang diukur
9.	MAX	:	Nilai ukur paling tinggi dari seluruh parameter yang diukur dalam waktu yang sama
10.	CRITICAL	:	Parameter yang hasil pengukurannya paling tinggi
11.	CATEGORI	:	Kategori hasil perhitungan indeks standar pencemaran udara

- **Pre Proses Data:** Pada proses ini adalah melakukan penggabungan data kualitas udara tiap bulan yaitu mulai bulan Januari sampai bulan Oktober 2021 dijadikan satu. Data yang sudah dikumpulkan dan diseleksi kemudian melakukan proses penghapusan data yang isinya kosong, duplikat, terisi tetapi tidak lengkap, dan menghapus atribut yang tidak diperlukan. Data yang sudah bersih dari nilai kosong atau tidak lengkap kemudian mengubah tipe data dari

variable karakter di ubah menjadi numerik dan factor untuk data kategori agar bisa dilakukan proses klasifikasi dengan SVM.

- **Split Data:** Langkah selanjutnya adalah melakukan splitting data yaitu memisahkan data menjadi dua bagian, yaitu data training dan testing untuk dilakukan analisis model SVM.
- **Model Klasifikasi SVM:** Pada model dijelaskan tentang alur proses klasifikasi dengan model SVM dimana ada 6 fitrur (pm10, pm25, so2, co, o3, dan no2) maka w akan memiliki 6 fitur yaitu $(w_{1,}w_{2,}w_{3,}w_{4,}w_{5,}w_{6})$ formula yang digunakan dalam meminimalkan margin sebagai berikut:
  - $\frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2 + w_2^2 + w_3^2 \ldots . + w_6^2)$
  - dengan yi $(\mathrm{wi.xi} + \mathrm{b}) > 1, \mathrm{i} = 1,2,3,4,5 \ldots \mathrm{N}$
  - $\mathrm{yi}(\mathrm{w1.x1} + \mathrm{w2.x2} + \mathrm{w3.x3} + \mathrm{w4.x4} + \mathrm{w5.x5} + \mathrm{w6.x6} >= +1$ untuk kategori Baik)
  - $\mathrm{yi}(\mathrm{w1.x1} + \mathrm{w2.x2} + \mathrm{w3.x3} + \mathrm{w4.x4} + \mathrm{w5.x5} + \mathrm{w6.x6} >= -1$ untuk kategori Sedang)
  - $\mathrm{yi}(\mathrm{w1.x1} + \mathrm{w2.x2} + \mathrm{w3.x3} + \mathrm{w4.x4} + \mathrm{w5.x5} + \mathrm{w6.x6} >= -1$ untuk kategori Tidak Sehat).
  - Untuk memperoleh nilai dari tiap-tiap atribut, maka perlu adanya perhitungan linear dengan eliminasi pada setiap persamaan. Pada klasifikasi lebih dari dua kelas, maka konsepnya adalah sebagai berikut, pada klasifikasi ini terdiri Kelas “Baik”, “Sedang” dan “Tidak Sehat”, maka yang perlu dilakukan adalah jika kelas kategori “Baik” akan diuji maka diberi label +1 dan data kelas “Sedang” dan “Tidak Sehat” diberi label -1,

sedangkan jika kelas “Sedang” akan diuji maka dilabeli dengan +1, sedangkan kelas “Baik” dan “Tidak Sehat” diberi label -1. Langkah berikutnya adalah mencari hyperplane dengan menggunakan dua kelas, setelah diperoleh hyperplane pada tiap-tiap kelas, kelas dari data baru x kemudian ditentukan oleh nilai terbesar hyperplane. Selanjutnya penetapan kelas berdasarkan nilai terbesar dari hyperplane tersebut.

- **Confusion matrix:** mengukur perfomansi menggunakan Confusion matrix yaitu melakukan pengukuran seberapa akurat sistem tersebut dalam mengklasifikasi data dimana terdapat 4 jenis data yaitu true positif (TP), true negative (TN), false positif (FP) dan false negative (FN).

3.2 Implementasi Support Vector Machine

Adapun untuk mengimplementasikan Support Vector Machine sesuai dengan desain yang telah dikembangkan yaitu menggunakan pemrograman R Studio dan beberapa library pendukung yaitu library(tidyverse), library(tidyr), library(e1071), dan library(caret) dan menjalankan tahapan-tahapan sesuai perancangan yang telah dibuat sebelumnya yaitu sebagai berikut:

- **Seleksi Data/Eksplorasi Data:** Pertama kali yang dilakukan adalah mengambil data dari dari Indeks Standar Pencemar Udara (ISPU) yang diukur dari lima stasiun pemantau kualitas udara (SPKU) di Provinsi DKI Jakarta Tahun 2021 sebanyak 9 data yaitu bulan Januari, bulan Februari, bulan Maret, bulan April, bulan Mei, bulan Juni, bulan September, bulan Oktober sedangkan bulan Agustus, bulan Nopember, dan bulan Desember datanya kurang lengkap. Dataset yang digunakan untuk analisis ini adalah

data data kualitas udara sebanyak 1118 data dan 11 variabel, dimana 4 variabel kategorik, 1 target class dan 7 variabel bebas tetapi yang digunakan untuk klasifikasi hanya 6 variabel bebas. Dari data ini akan diprediksi apakah kualitas udara yang baru terdeteksi berkualitas Baik, Sedang, ataupun Tidak Sehat berdasarkan 6 variabel prediksi.

```r
JAN <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JANUARI2021.csv")
FEB <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/FEBRUARI2021.csv")
MARET <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/MARET2021.csv")
APRIL <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/APRIL2021.csv")
MEI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/MEI2021.csv")
JUNI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JUNI2021.csv")
JULI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JULI2021.csv")
SEPTEMBER <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/SEPTEMBER2021.csv")
OKTOBER <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/OKTOBER2021.csv")
```

Gambar 3. 2 Pembacaan Data

Pada Gambar 3.2 yaitu proses memasukkan data kualitas udara yang akan digunakan ke program R melalui fungsi read.csv, dimana data kulitas udara bulan Januari sampai dengan Oktober dieksport kedalam bentuk R. Dari beberapa data yang telah dibaca kemudian digabung menjadi satu seperti dijelaskan pada Gambar 3.3 dibawah ini.

```r
dataset <- rbind(JAN, FEB,MARET,APRIL,MEI,JUNI,JULI,SEPTEMBER,OKTOBER)
dataset
```

A tibble: 1,365 x 11

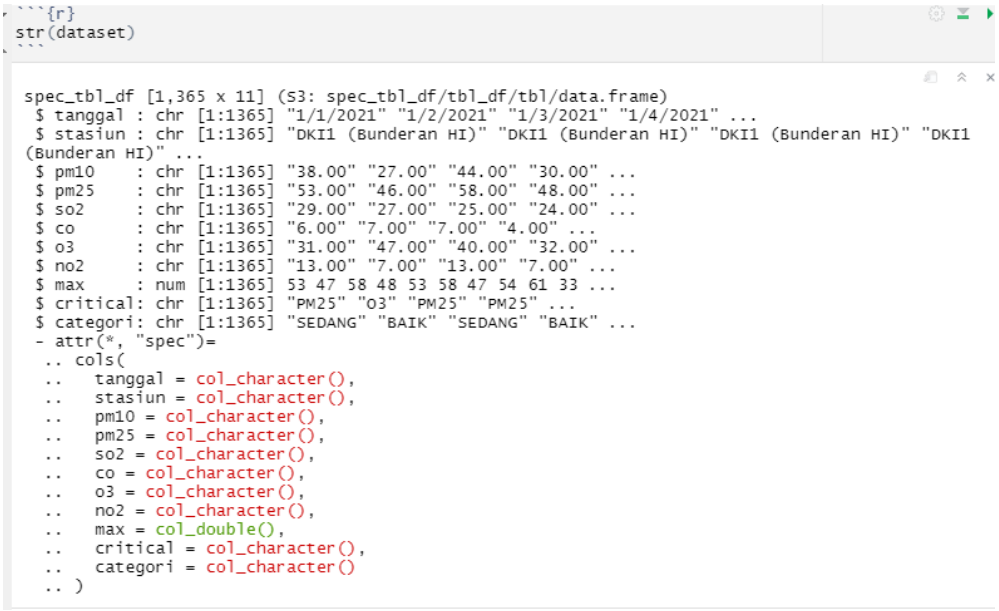
tanggal <chr>	stasiun <chr>	pm10 <chr>	pm25 <chr>	so2 <chr>	co <chr>	o3 <chr>	no2 <chr>	max <dbl>
1/1/2021	DKI1 (Bunderan HI)	38.00	53.00	29.00	6.00	31.00	13.00	53
1/2/2021	DKI1 (Bunderan HI)	27.00	46.00	27.00	7.00	47.00	7.00	47
1/3/2021	DKI1 (Bunderan HI)	44.00	58.00	25.00	7.00	40.00	13.00	58
1/4/2021	DKI1 (Bunderan HI)	30.00	48.00	24.00	4.00	32.00	7.00	48
1/5/2021	DKI1 (Bunderan HI)	38.00	53.00	24.00	6.00	31.00	9.00	53
1/6/2021	DKI1 (Bunderan HI)	41.00	58.00	23.00	13.00	46.00	13.00	58
1/7/2021	DKI1 (Bunderan HI)	35.00	47.00	22.00	6.00	39.00	10.00	47
1/8/2021	DKI1 (Bunderan HI)	37.00	54.00	26.00	16.00	17.00	10.00	54
1/9/2021	DKI1 (Bunderan HI)	47.00	61.00	16.00	27.00	22.00	12.00	61
1/10/2021	DKI1 (Bunderan HI)	23.00	25.00	16.00	11.00	33.00	8.00	33

1-10 of 1,365 rows | 1-9 of 11 columns Previous 1 2 3 4 5 6 ... 100 Next

Gambar 3. 3 Penggabungan Data

Pada Gambar 3.3 menunjukkan bahwa setelah data dibaca maka proses selanjutnya menggabungkan data menggunakan fungsi rbind di R, yaitu penggabungan vector, matriks dan data perbaris.

- **Pre Proses Data:** merupakan tahapan yang harus dilakukan untuk menormalisasi data agar bisa diproses dengan baik. Cleaning data atau menghapus baris yang kosong atau terdapat NA's untuk menghindari kesalahan pada perhitungan. Berikut ini tampilan pre proses data seperti diperlihatkan pada beberapa gambar di bawah ini.



```
```{r}
str(dataset)
```

spec_tbl_df [1,365 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ tanggal : chr [1:1365] "1/1/2021" "1/2/2021" "1/3/2021" "1/4/2021" ...
 $ stasiun : chr [1:1365] "DKI1 (Bunderan HI)" "DKI1 (Bunderan HI)" "DKI1 (Bunderan HI)" "DKI1
(Bunderan HI)" ...
 $ pm10   : chr [1:1365] "38.00" "27.00" "44.00" "30.00" ...
 $ pm25   : chr [1:1365] "53.00" "46.00" "58.00" "48.00" ...
 $ so2   : chr [1:1365] "29.00" "27.00" "25.00" "24.00" ...
 $ co   : chr [1:1365] "6.00" "7.00" "7.00" "4.00" ...
 $ o3   : chr [1:1365] "31.00" "47.00" "40.00" "32.00" ...
 $ no2   : chr [1:1365] "13.00" "7.00" "13.00" "7.00" ...
 $ max   : num [1:1365] 53 47 58 48 53 58 47 54 61 33 ...
 $ critical: chr [1:1365] "PM25" "O3" "PM25" "PM25" ...
 $ categori: chr [1:1365] "SEDANG" "BAIK" "SEDANG" "BAIK" ...
 - attr(*, "spec")=
  .. cols(
  ..   tanggal = col_character(),
  ..   stasiun = col_character(),
  ..   pm10 = col_character(),
  ..   pm25 = col_character(),
  ..   so2 = col_character(),
  ..   co = col_character(),
  ..   o3 = col_character(),
  ..   no2 = col_character(),
  ..   max = col_double(),
  ..   critical = col_character(),
  ..   categori = col_character()
  ..)
```

Gambar 3. 4 Bentuk Struktur data

Pada Gambar 3.4 menunjukkan struktur data atau keterangan data, dari data yang diperoleh menunjukkan bahwa data tersebut dibaca oleh R sebagai data chr atau character, sedangkan data yang dibutuhkan agar bisa diproses adalah menggunakan data numeric dan categorical. Oleh karena itu perlu melalui proses selanjutnya sperti diperlihatkan pada Gambar 3.5 di bawah ini:

```r
dataset <- na.omit(dataset)
dataset <- dataset[!apply(is.na(dataset) | dataset == "NA", 1, all),]
```

Gambar 3. 5 Bentuk Struktur data

Pada Gambar 3.5 melakukan proses penghapusan data yang tidak lengkap yang terdapat pada data kualitas udara menggunakan fungsi na.omit() di R dan menghapus data NA atau beberapa data yang hilang menggunakan fungsi (is.na). Agar data bisa diproses maka perlu mengubah tipe datanya, karena yang akan melakukan klasifikasi maka perlu ada data numerik dan categorical. Untuk mengubah tipe datanya maka perlu dilakukan proses seperti dijelaskan pada Gambar 3.6 di bawah ini.

```r
dataset$tanggal<- as.factor(dataset$tanggal)
dataset$stasiun<- as.factor(dataset$stasiun)
dataset$pm10<- as.numeric(dataset$pm10)
dataset$pm25<- as.numeric(dataset$pm25)
dataset$so2<- as.numeric(dataset$so2)
dataset$co<- as.numeric(dataset$co)
dataset$o3<- as.numeric(dataset$o3)
dataset$no2<- as.numeric(dataset$no2)
dataset$critical<- as.factor(dataset$critical)
dataset$categori<- as.factor(dataset$categori)
```

Gambar 3. 6 Mengubah tipe data

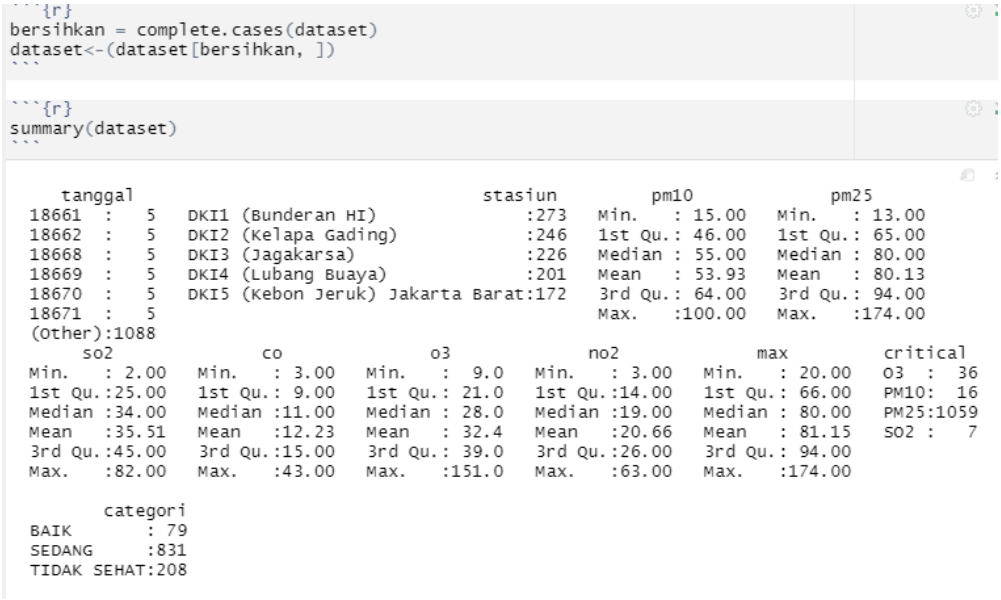
Pada Gambar 3. 6 menjelaskan proses mengubah data ke dalam bentuk variabel numeric dan kategorikal/faktor pada R, setelah diproses maka akan menghasilkan berupa data dengan variable numerik dan variable kategorikal. Untuk mengetahui nilai statistic dari data yang telah diproses dalam bentuk variable numerik dan kategorikal yaitu pada nilai minimum, Q1, Q2 atau median, Q3, maksimum dapat dilihat pada proses di bawah ini yang dijelaskan pada Gambar 3.7 dibawah ini

```r
summary(dataset)
```

```
      tanggal                                  stasiun         pm10            pm25       
 10/1/2021 :   5   DKI1 (Bunderan HI)             :273   Min.   : 15.0   Min.   : 13.00  
 10/10/2021:   5   DKI2 (Kelapa Gading)           :271   1st Qu.: 46.0   1st Qu.: 65.00  
 10/11/2021:   5   DKI3 (Jagakarsa)               :271   Median : 55.0   Median : 80.00  
 10/12/2021:   5   DKI4 (Lubang Buaya)            :237   Mean   : 53.8   Mean   : 79.88  
 10/13/2021:   5   DKI5 (Kebon Jeruk) Jakarta Barat:241   3rd Qu.: 63.0   3rd Qu.: 94.00  
 10/14/2021:   5                                          Max.   :100.0   Max.   :174.00  
 (Other)   :1263                                          NA's   :44      NA's   :15      
      so2              co               o3              no2              max        
 Min.   : 2.00   Min.   : 3.00   Min.   :  9.00   Min.   : 1.00   Min.   : 20.00  
 1st Qu.:25.00   1st Qu.: 9.00   1st Qu.: 22.00   1st Qu.:14.00   1st Qu.: 66.00  
 Median :34.00   Median :11.00   Median : 29.00   Median :19.00   Median : 80.00  
 Mean   :35.53   Mean   :11.94   Mean   : 32.71   Mean   :20.07   Mean   : 80.77  
 3rd Qu.:45.00   3rd Qu.:14.00   3rd Qu.: 39.00   3rd Qu.:25.00   3rd Qu.: 94.00  
 Max.   :82.00   Max.   :43.00   Max.   :151.00   Max.   :63.00   Max.   :174.00  
 NA's   :73      NA's   :18      NA's   :29       NA's   :7                       
 critical          categori    
 O3  :  49   BAIK       : 93  
 PM10:  30   SEDANG     :967  
 PM25:1204   TIDAK SEHAT:233  
 SO2 :  10                    
```

Gambar 3. 7 Nilai Statistic Dataset

Dari Gambar 3.7 dijelaskan hasil dari beberapa nilai statistik deskriptif pada tiap-tiap variabel yaitu pada variabel dengan tipe numerik maka yang dihasilkan adalah nilai Min(minimum), (1st Qu) quantil pertama, median, nilai rata-rata, quantil ketiga, nilai max dan ada beberapa missing value yaitu NA's, sedangan pada variabel kategorik maka dihasilkan data kelas. Karena hasil proses konversi masih ada beberapa data yang missing value atau NA's dan itu akan mempengaruhi pada proses klasifikasi maka perlu adanya proses penghapusan pada NA's dengan fungsi complete.cases() yang dijelaskan pada Gambar 3.8



```
```{r}
bersihkan = complete.cases(dataset)
dataset<-(dataset[bersihkan, ])
```

```{r}
summary(dataset)
```

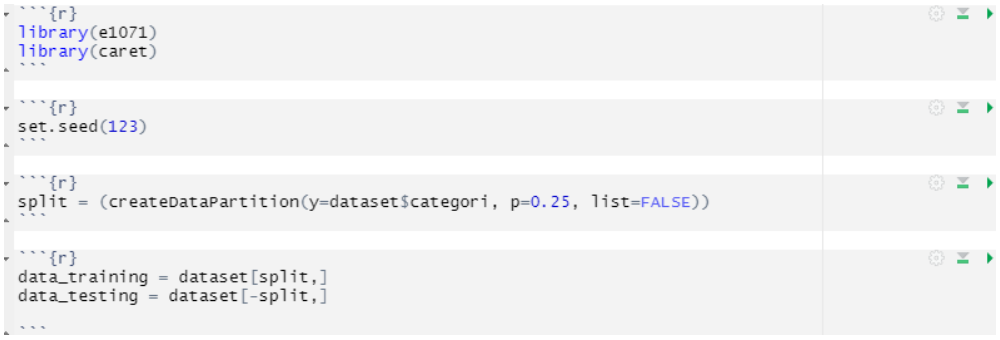
    tanggal   stasiun   pm10   pm25
 18661 :   5   DKI1 (Bunderan HI)   :273   Min.   : 15.00   Min.   : 13.00
 18662 :   5   DKI2 (Kelapa Gading)   :246   1st Qu.: 46.00   1st Qu.: 65.00
 18668 :   5   DKI3 (Jagakarsa)   :226   Median : 55.00   Median : 80.00
 18669 :   5   DKI4 (Lubang Buaya)   :201   Mean   : 53.93   Mean   : 80.13
 18670 :   5   DKI5 (Kebon Jeruk) Jakarta Barat:172   3rd Qu.: 64.00   3rd Qu.: 94.00
 18671 :   5   Max.   :100.00   Max.   :174.00
 (Other):1088
      so2   co   o3   no2   max   critical
 Min.   : 2.00   Min.   : 3.00   Min.   : 9.0   Min.   : 3.00   Min.   : 20.00   O3 : 36
 1st Qu.:25.00   1st Qu.: 9.00   1st Qu.: 21.0   1st Qu.:14.00   1st Qu.: 66.00   PM10: 16
 Median :34.00   Median :11.00   Median : 28.0   Median :19.00   Median : 80.00   PM25:1059
 Mean   :35.51   Mean   :12.23   Mean   : 32.4   Mean   :20.66   Mean   : 81.15   SO2 :   7
 3rd Qu.:45.00   3rd Qu.:15.00   3rd Qu.: 39.0   3rd Qu.:26.00   3rd Qu.: 94.00
 Max.   :82.00   Max.   :43.00   Max.   :151.0   Max.   :63.00   Max.   :174.00

        categori
 BAIK   : 79
 SEDANG   :831
 TIDAK SEHAT:208

Gambar 3. 8 Penghapusan Missing Value

Pada Gambar 3.8 adalah proses penghapusan missing value atau data yang hilang yang menyebabkan tidak akuratnya data, sehingga bisa dilihat dengan summary() data hasil penghapusan data sehingga terlihat hasil dari beberapa nilai statistik deskriptif pada tiap-tiap variabel yaitu pada variabel dengan tipe numerik maka yang dihasilkan adalah nilai Min(minimum), (1st Qu) quantil pertama, median, nilai rata-rata, quantil ketiga, nilai max dan ada beberapa missing value yaitu NA's, sedangan pada variabel kategorik maka dihasilkan data kelas.

- **Split Data** merupakan tahapan dalam mechine learning untuk membagi data training dan data testing, kali ini mencoba dengan data 25% untuk data training dan sisanya untuk data testing, seperti dijelaskan pada Gambar 3.9 di bawah ini.
```



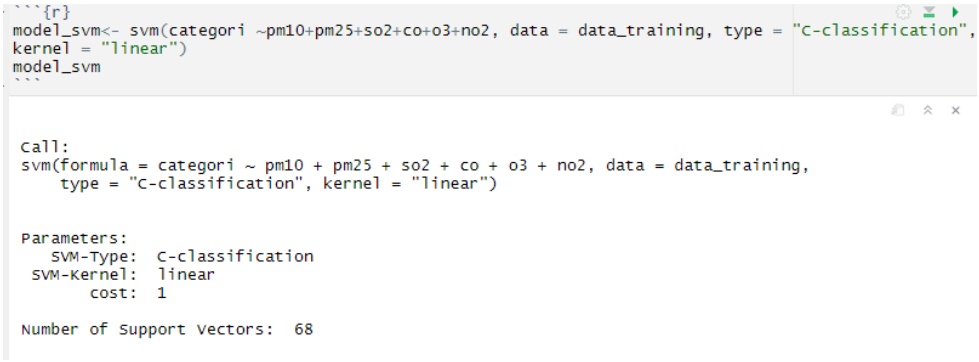
Gambar 3. 9 Pembagian data

Pada Gambar 3.9 adalah proses pembagian data dimana data pada kualitas udara jakarta yang kami gunakan adalah sebanyak 1118, kemudian data tersebut akan dibagi menjadi dua bagian yaitu 25% data training yang dipakai untuk membuat model di svm sedangkan sisanya data testing yang akan digunakan untuk mengevaluasi model.

- **Model Klasifikasi Support Vector Machine**

Pada model klasifikasi Support Vector Machine memasukkan data training untuk memprediksi kualitas udara masuk ke kategori Baik, Sedang, atau Tidak Sehat. Dari dataset yang dimiliki, maka yang menjadi target variabel untuk diprediksi adalah categori, sedangkan variabel pm10, pm25, so2, co, o3, dan no2. Adapun tahapan dalam membangun model diantaranya Setelah data melalui beberapa proses diantara seleksi data, praproses data, kemudian data kualitas udara dibagi menjadi dua bagian yaitu data training dan data testing. Untuk menghasilkan model maka diperlukan data training, sedangkan untuk data testing dimanfaatkan dalam melakukan evaluasi model yang telah dibuat. Pada penelitian ini untuk data training sebanyak 1118 baris. Data training akan diambil secara random, sedangkan sisanya yang tidak terpilih akan menjadi data

testing. Setelah data training dan data testing terbentuk kemudian akan dilakukan pemodelan dengan menggunakan fungsi SVM yang ada pada library pada R studio, library yang dibutuhkan untuk melakukan klasifikasi yaitu (e1071) yang berisi fungsi-fungsi yang akan digunakan dalam membentuk support vector. Setelah mengaktifkan fungsi library (e1071) maka proses selanjutnya adalah membentuk model Support Vector Machine seperti dijelaskan pada Gambar 3.10 di bawah ini.



```r
```{r}
model_svm<- svm(categori ~pm10+pm25+so2+co+o3+no2, data = data_training, type = "C-classification",
kernel = "linear")
model_svm
```
```

```
Call:
svm(formula = categori ~ pm10 + pm25 + so2 + co + o3 + no2, data = data_training,
    type = "C-classification", kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  68
```

Gambar 3. 10 Model Support Vector Machine

Pada Gambar 3.10 menjelaskan tentang model Support Vector Machine dengan nama Model_SVM yaitu menggunakan type ‘C-classification’ dan kernel = ‘Linear' dan nilai support vector sebagai hyperplane. Dari model svm selanjutnya digunakan untuk melakukan prediksi sebagai cara untuk mengklasifikasi menggunakan data testing seperti dijelaskan pada Gambar 3.11 di bawah ini.

```r
preds_svm <- predict(model_svm, newdata = data_testing)
```

```r
preds_svm
```

```
       1        2        3        4        5        6        7
SEDANG   SEDANG   SEDANG    BAIK   SEDANG   SEDANG   SEDANG
       8        9       10       11       12       13       14
SEDANG    BAIK   SEDANG   SEDANG    BAIK   SEDANG   SEDANG
      15       16       17       18       19       20       21
SEDANG   SEDANG   SEDANG    BAIK   SEDANG TIDAK SEHAT   SEDANG
      22       23       24       25       26       27       28
SEDANG    BAIK    BAIK   SEDANG    BAIK   SEDANG   SEDANG
      29       30       31       32       33       34       35
  BAIK   SEDANG   SEDANG   SEDANG   SEDANG   SEDANG   SEDANG
      36       37       38       39       40       41       42
SEDANG   SEDANG   SEDANG   SEDANG   SEDANG   SEDANG    BAIK
      43       44       45       46       47       48       49
SEDANG   SEDANG   SEDANG   SEDANG   SEDANG   SEDANG   SEDANG
      50       51       52       53       54       55       56
```

Gambar 3. 11 Klasifikasi pada Data Testing

Dari model svm yang akan digunakan untuk mempredikasi menggunakan data testing seperti pada pada Gambar 3.11 adalah hasil prediksi dari data testing dan menghasilkan urutan data yang digunakan. Sedangkan "Baik", "Sedang" dan "Tidak Sehat" merupakan kategori dari variabel dependen. Untuk melihat perbandingan hasil klasifikasi yaitu antara data hasil prediksi dengan data aktual dijelaskan pada seperti pada Gambar

```r
hasil=cbind(PREDIKSI=as.character(preds_svm),AKTUAL=as.character(data_testing$categori))
head(hasil, n = 15)
```

```
      PREDIKSI AKTUAL
 [1,] "SEDANG" "SEDANG"
 [2,] "SEDANG" "BAIK"
 [3,] "SEDANG" "SEDANG"
 [4,] "BAIK"   "BAIK"
 [5,] "SEDANG" "SEDANG"
 [6,] "SEDANG" "SEDANG"
 [7,] "SEDANG" "BAIK"
 [8,] "SEDANG" "SEDANG"
 [9,] "BAIK"   "BAIK"
[10,] "SEDANG" "SEDANG"
[11,] "SEDANG" "BAIK"
[12,] "BAIK"   "BAIK"
[13,] "SEDANG" "SEDANG"
[14,] "SEDANG" "SEDANG"
[15,] "SEDANG" "SEDANG"
```

Gambar 3. 12 Perbandingan Hasil Klasifikasi

Pada Gambar 3.12 menjelaskan hasil prediksi klasifikasi pada data testing dibandingkan dengan data aktual atau data sebenarnya. Untuk langkah selanjutnya yang dilakukan adalah melihat tingkat akurasi dari hasil klasifikasi yang dilakukan dengan menggunakan confusion matrix seperti dijelaskan pada Gambar 3.13 di bawah ini.

```
conf_matrix_svm <- table(preds_svm, data_testing$categori)
```

```{r}
confusionMatrix(conf_matrix_svm)
```

```
Confusion Matrix and Statistics


preds_svm     BAIK SEDANG TIDAK SEHAT
  BAIK          45      9           0
  SEDANG        14    608          39
  TIDAK SEHAT    0      6         117

Overall Statistics

               Accuracy : 0.9189
                 95% CI : (0.8983, 0.9364)
    No Information Rate : 0.7434
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7874

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: BAIK Class: SEDANG Class: TIDAK SEHAT
Sensitivity              0.76271        0.9759             0.7500
Specificity              0.98845        0.7535             0.9912
Pos Pred Value           0.83333        0.9198             0.9512
Neg Pred Value           0.98214        0.9153             0.9455
Prevalence               0.07041        0.7434             0.1862
Detection Rate           0.05370        0.7255             0.1396
Detection Prevalence     0.06444        0.7888             0.1468
```

Gambar 3. 13 Confusion Matrix

Pada Gambar 3.13 menjelaskan tentang hasil pengujian yang dilakukan dan menghasilkan beberapa informasi penting diantaranya hasil tingkat akurasi predikasi yaitu seberapa baik model melakukan klasifikasi, sedangkan precision merupakan perbandingan jumlah data yang diproyeksikan positif dengan jumlah total data yang diprediksi positif. Recall atau sensitivitas dipakai dalam

memperoleh perbandingan jumlah prediksi positif yang benar dibandingkan dengan jumlah total kelas positif.

3.3 Ujicoba Metode Support Vector Machine

Pada tahap ini dilakukan pengujian dengan mengatur pembagian data training dan data testing untuk mengetahui performan dari model Support Vector Machine yaitu nilai *Accuracy, Precision, Recall*. Pengujian ini dilakukan untuk mengetahui seberapa pengaruh banyaknya data training dan testing terhadap klasifikasi yang dilakukan oleh Support Vector Machine. pengujian dilakukan dengan menggunakan metode *Confusion Matrix*. Adapun model pengujiannya seperti ditampilkan pada Tabel 3.1.

Tabel 3. 1 Skenario Pengujian Pada Support Vector Machine

Pengujian	Training%	Testing%
1	25	75
2	35	65
3	45	55
4	55	45
5	65	35
6	75	25
7	85	15
8	90	10

Pada Tabel 3.1 akan dilakukan pengujian terhadap model Support Vector Machine yang telah dibuat dengan dengan prosentase pembagian data training dan testing pada waktu proses splitting data. Proses pengujiannya akan dijabarkan satu persatu di bawah ini yaitu:

- **Pengujian 1**

Pada Pengujian 1 pembagian data training 25% dan data testing 75% dengan data training sebanyak **280** serta data testing sebanyak **838**. Adapun hasil Confusion Matrix seperti dijelaskan pada Tabel 3.2 di bawah ini:

Tabel 3. 2 Confusion Matrix SVM Pengujian 1

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	46	18	0
	SEDANG	13	602	5
	TIDAK SEHAT	0	3	151

Dari Tabel 3.2 menampilkan tabel Confusion Matrix pada Pengujian 1 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.3

Tabel 3. 3 Accuracy, Precision, Recall pada SVM Pengujian 1

Accuracy	Precision	Recall
95.35%	90.46%	89.01%

Pada Tabel 3.3 diatas menunjukkan nilai accuracy sebesar 95.35%, nilai precision 90.46%, dan nilai recall 89.01%.

- **Pengujian 2**

Pada pengujian 2 pembagian data training 35% dan data testing 65% dengan training sebanyak 392, data testing sebanyak 726, serta Support Vector yang

sebanyak 87. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 3.4 di bawah ini:

Tabel 3. 4 Confusion Matrix SVM Pengujian 2

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	40	12	0
	SEDANG	11	525	9
	TIDAK SEHAT	0	3	126

Dari Tabel 3.4 menampilkan tabel Confusion Matrix pada Pengujian 2 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.5

Tabel 3. 5 Accuracy, Precision, Recall SVM pada Pengujian 2

Accuracy	Precision	Recall
95.18%	89.66%	90.31%

Pada Tabel 3.5 diatas menunjukkan nilai *accuracy* sebesar **95.18%**, nilai *precision* **89.66%**, dan nilai *recall* **90.31%**.

- **Pengujian 3**

Pada pengujian 3 pembagian data training 45% dan data testing 55% dengan training sebanyak 504, data testing sebanyak 614, serta Support Vector yang sebanyak 104. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 4.6 di bawah ini:

Tabel 3. 6 Confusion Matrix SVM Pengujian 3

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	31	3	0

	SEDANG	**12**	**450**	**18**
	TIDAK SEHAT	**0**	**4**	**96**

Dari Tabel 3.6 menampilkan tabel Confusion Matrix pada pengujian 3 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.7

Tabel 3. 7 Accuracy, Precision, Recall SVM pada Pengujian 3

Accuracy	Precision	Recall
95.77%	**89.72%**	**91.16%**

Pada Tabel 4.7 diatas menunjukkan nilai *accuracy* sebesar 95.77%, nilai *precision* 89.72%, dan nilai *recall* 91.16%.

- **Pengujian 4**

Pada pengujian 4 pembagian data training 55% dan data testing 45% dengan training sebanyak 617, data testing sebanyak 501, serta Support Vector yang sebanyak 109. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 4.8 di bawah ini:

Tabel 3. 8 Confusion Matrix SVM Pengujian 4

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	28	1	0
	SEDANG	6	290	15
	TIDAK SEHAT	0	1	64

Dari Tabel 3.8 menampilkan tabel Confusion Matrix pada pengujian 3 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.9

Tabel 3. 9 Accuracy, Precision, Recall SVM pada Pengujian 4

Accuracy	Precision	Recall
95.01%	**88.06%**	**91.10%**

Pada Tabel 3.9 diatas menunjukkan nilai *accuracy* sebesar **95.01%**, nilai *precision* **88.06%**, dan nilai *recall* **91.10%**.

- **Pengujian 5**

Pada pengujian 5 pembagian data training 65% dan data testing 35% dengan training sebanyak 729, data testing sebanyak 389, serta Support Vector yang sebanyak 135. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 4.10 di bawah ini:

Tabel 3. 10 Confusion Matrix SVM Pengujian 5

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	18	0	0
	SEDANG	9	289	6
	TIDAK SEHAT	0	1	66

Dari Tabel 4.10 menampilkan tabel Confusion Matrix pada pengujian 5 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.11

Tabel 3. 11 Accuracy, Precision, Recall SVM pada Pengujian 5

Accuracy	Precision	Recall
95.89%	**86.00%**	**97.86%**

Pada Tabel 3.11 diatas menunjukkan nilai *accuracy* sebesar 95.89%, nilai *precision* 86.00%, dan nilai *recall* 97.86%.

- **Pengujian 6**

Pada pengujian 6 pembagian data training 75% dan data testing 25% dengan training sebanyak 840, data testing sebanyak 278, serta Support Vector yang sebanyak 144. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 4.10 di bawah ini:

Tabel 3. 12 Confusion Matrix SVM Pengujian 6

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	12	1	0
	SEDANG	7	204	2
	TIDAK SEHAT	0	2	50

Dari Tabel 3.12 menampilkan tabel Confusion Matrix pada pengujian 6 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.13

Tabel 3. 13 Accuracy, Precision, Recall SVM pada Pengujian 6

Accuracy	Precision	Recall
95.68%	85.95%	94.75%

Pada Tabel 4.13 diatas menunjukkan nilai *accuracy* sebesar **95.68%**, nilai *precision* **85.95%**, dan nilai *recall* **94.75%**.

- **Pengujian 7**

Pada pengujian 7 pembagian data training 85% dan data testing 15% dengan training sebanyak 952, data testing sebanyak 166, serta Support Vector yang sebanyak 164. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 4.14 di bawah ini:

Tabel 3. 14 Confusion Matrix SVM Pengujian 7

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	7	1	0
	SEDANG	4	121	1
	TIDAK SEHAT	0	2	30

Dari Tabel 3.14 menampilkan tabel Confusion Matrix pada pengujian 7 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.15

Tabel 3. 15 Accuracy, Precision, Recall SVM pada Pengujian 7

Accuracy	Precision	Recall
95.18%	86.00%	92.43%

Pada Tabel 3.15 diatas menunjukkan nilai *accuracy* sebesar 95.18%, nilai *precision* 86.00%, dan nilai *recall* 92.43%.

- **Pengujian 8**

Pada pengujian 8 pembagian data training 90% dan data testing 10% dengan training sebanyak 1008, data testing sebanyak 110, serta Support Vector yang

sebanyak 169. Adapun hasil Confusion Matrix seperti dijelaskan pada Confusion Matrix seperti dijelaskan pada Tabel 3.16 di bawah ini:

Tabel 3. 16 Confusion Matrix SVM Pengujian 7

Support Vector Machine		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	3	1	0
	SEDANG	4	82	0
	TIDAK SEHAT	0	0	20

Dari Tabel 3.16 menampilkan tabel Confusion Matrix pada pengujian 8 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.17

Tabel 3. 17 Accuracy, Precision, Recall SVM pada Pengujian 8

Accuracy	Precision	Recall
95.45%	80.55%	90.12%

Pada Tabel 3.15 diatas menunjukkan nilai *accuracy* sebesar **95.45%**, nilai *precision* **80.55%**, dan nilai *recall* **90.12%**.

- **Kesimpulan**

Dari hasil pengujian yang telah dilakukan sebanyak 8 kali pengujian didapatkan nilai rata-rata *Accuracy, Precision, Recall* seperti ditampilkan pada Tabel 3.18 sebagai berikut:

Tabel 3. 18 Rata-rata Accuracy, Precision, Recall pada SVM

Pengujian	Training%	Testing%	Accuracy%	Precision%	Recall%
1	25	75	95.35%	90.46%	89.01%
2	35	65	95.18%	89.66%	90.31%

3	**45**	**55**	**95.77%**	**89.72%**	**91.16%**
4	**55**	**45**	**95.01%**	**88.06%**	**91.10%**
5	**65**	**35**	**95.89%**	**86.00%**	**97.86%**
6	**75**	**25**	**95.68%**	**85.95%**	**94.75%**
7	**85**	**15**	**95.18%**	**86.00%**	**92.43%**
8	**90**	**10**	**95.45%**	**80.55%**	**90.12%**
Rata-rata			95.44%	87.05%	92.09%

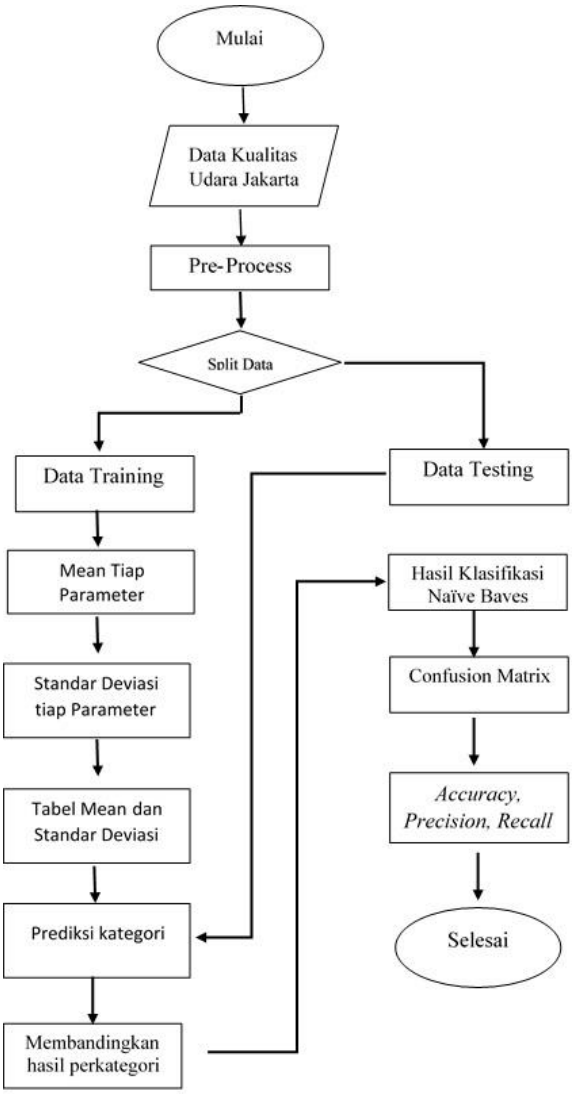
Pada Tabel 3.18 terlihat bahwa model Support Vector Machine mampu melakukan klasifikasi pada data kualitas udara Jakarta 2021. Dengan 8 kali percobaan dengan melakukan perubahan pada data training dan testing menunjukkan nilai rata-rata Accuracy sebesar **95.44%**, Precision sebesar **92.09%**, dan Recall sebesar **87.05%**.

BAB IV

METODE NAÏVE BAYES

4.1 Desain Klasifikasi Metode Naïve Bayes

Pada bab ini akan menjelaskan tentang pola dan rancangan klasifikasi dapat dilihat pada Gambar 4.1 sebagai berikut:



Gambar 4. 1 Bagan Desain Klasifikasi Metode Naïve Bayes

Pada Gambar 4.1 menjelaskan tentang rancangan desain metode naïve bayes dengan tahapan-tahapan sebagai berikut :

- **Seleksi Data/ Eksplorasi Data:** dimulai dari menganalisa data yang akan digunakan dalam proses klasifikasi naïve Bayes yaitu dataset yang berisi tentang Indeks Standar Pencemar Udara (ISPU) yang diukur dari 5 stasiun pemantau kualitas udara (SPKU) yang ada di Provinsi DKI Jakarta Tahun 2021, data ini diambil dari website resmi https://data.jakarta.go.id/.
- **Pre Proses Data:** Pada proses ini adalah melakukan penggabungan data kualitas udara tiap bulan mulai bulan Januari sampai bulan Juli 2021 dijadikan satu. Data yang sudah dikumpulkan dan diseleksi kemudian melakukan proses penghapusan data yang memiliki nilai kosong, duplikat, terisi namun tidak lengkap, dan penghapusan atribut yang tidak diperlukan. Data yang sudah bersih dari nilai kosong atau tidak lengkap kemudian mengubah tipe data dari variable karakter di ubah menjadi numerik dan factor untuk data kategori agar bisa dilakukan proses klasifikasi dengan Naïve Bayes.
- **Split Data:** Langkah selanjutnya adalah melakukan splitting data yaitu memisahkan data menjadi 2 bagian, training dan testing untuk dilakukan analisis model Naïve Bayes.
- **Model Klasifikasi Naïve Bayes:** Pada model dijelaskan tentang alur proses klasifikasi dengan model Naïve Bayes dengan melalui tahapan-tahapan. Adapun tahapan-tahapan untuk membuat model Naïve Bayes diantaranya sebagai berikut:

  1. Menghitung nilai mean disetiap atributnya berdasarkan kategori

2. Menghitung nilai standar deviasi dari setiap atribut perkategori
3. Menghitung nilai probabilitas tiap kategori
4. Mengalikan semua variable kategori
5. Menghitung rata-rata, simpangan baku, dan probabilitas dari nilai-nilai dalam tabel
6. Membandingkan hasil perkategori

- **Confusion matrix:** Melakukan pengukuran seberapa akurat sistem tersebut dalam mengklasifikasi data dimana memilki empat informasi yaitu true positif(TP), true negative (TN), false positif (FP) dan false negative (FN) .

4.2 Implementasi Metode Naïve Bayes

Adapun untuk implementasi metode Naïve Bayes yaitu menggunakan pemrograman R Studio dan beberapa library pendukung yaitu library(tidyverse), library(tidyr), library(e1071), dan library(caret) dan menjalankan tahapan-tahapan sesuai perancangan yang telah dibuat sebelumnya yaitu sebagai berikut:

- **Seleksi Data/Eksplorasi Data:** Pertama kali yang dilakukan adalah mengambil data dari dari 5 stasiun pemantau kualitas udara (SPKU) Provinsi DKI Jakarta Tahun 2021 yaitu bulan Januari sampai dengan bulan Juli. Dataset yang digunakan untuk analisis ini adalah data data kualitas udara sebanyak 992 data dan 11 variabel, dimana 4 variabel kategorik, 1 target class dan 7 variabel bebas tetapi yang digunakan untuk klasifikasi hanya 6 variabel bebas. Dari data ini akan diprediksi apakah kualitas udara yang baru terdeteksi berkualitas Baik, Sedang, ataupun Tidak Sehat berdasarkan 6 variabel prediksi. Berikut adalah

deskripsi struktur dari data kualitas udara Jakarta Tahun 2021 yang dijelaskan pada Gambar 4.2.

```
   {r}
str(dataset)
```

```
tbl_df [992 x 11] (S3: tbl_df/tbl/data.frame)
 $ tanggal : chr [1:992] "1/1/2021" "1/2/2021" "1/3/2021" "1/4/2021" ...
 $ stasiun : chr [1:992] "DKI1 (Bunderan HI)" "DKI1 (Bunderan HI)" "DKI1 (Bunderan HI)"
(Bunderan HI)" ...
 $ pm10    : chr [1:992] "38.00" "27.00" "44.00" "30.00" ...
 $ pm25    : chr [1:992] "53.00" "46.00" "58.00" "48.00" ...
 $ so2     : chr [1:992] "29.00" "27.00" "25.00" "24.00" ...
 $ co      : chr [1:992] "6.00" "7.00" "7.00" "4.00" ...
 $ o3      : chr [1:992] "31.00" "47.00" "40.00" "32.00" ...
 $ no2     : chr [1:992] "13.00" "7.00" "13.00" "7.00" ...
 $ max     : num [1:992] 53 47 58 48 53 58 47 54 61 33 ...
 $ critical: chr [1:992] "PM25" "O3" "PM25" "PM25" ...
 $ categori: chr [1:992] "SEDANG" "BAIK" "SEDANG" "BAIK" ...
 - attr(*, "na.action")= 'omit' Named int [1:68] 94 95 96 97 98 99 100 101 102 103 ...
  ..- attr(*, "names")= chr [1:68] "94" "95" "96" "97" ...
```

Gambar 4. 2 Struktur Dataset Kualitas Udara Jakara 2021

Pada Gambar 4.2 menjelaskan tetang struktur data dari kualitas udara Jakarta tahun 2021 dan semuanya masih bertipe data character, maka untuk proses selanjutnya perlu dilakukan pre proses data agar bisa dilakukan proses klasifikasi.

1. **Pre Proses Data:** merupakan tahapan yang harus dilakukan untuk menormalisasi data agar bisa diproses dengan baik. Adapun tahapannya sebagai berikut:

   - Penggabungan data kualitas udara tiap bulan yaitu mulai bulan Januari sampai bulan Oktober 2021 dijadikan satu menggunakan r bind pada R studio, seperti dijelaskan pada Gambar 4.3.

```
```{r}
JAN <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JANUARI2021.csv")
FEB <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/FEBRUARI2021.csv")
MARET <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/MARET2021.csv")
APRIL <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/APRIL2021.csv")
MEI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/MEI2021.csv")
JUNI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JUNI2021.csv")
JULI <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/JULI2021.csv")
SEPTEMBER <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/SEPTEMBER2021.csv")
OKTOBER <- read_csv("D:/SEKOLAH/TESIS/dataset 2021/OKTOBER2021.csv")
```
```

Gambar 4. 3 Pembacaan Data Kualitas Udara 2021

Pada Gambar 4.3 menjelaskan tentang pembacaan data kualitas udara tahun 2021 mulai bulan Januari sampai dengan bulan Oktober berupa file csv yang digabung menjadi satu dan diproses sebagai data training dan testing. Proses penggabungan data akan dijelaskan pada Gambar 4.4.

```{r}
dataset <- rbind(JAN, FEB,MARET,APRIL,MEI,JUNI,JULI,SEPTEMBER,OKTOBER)
dataset
```

A tibble: 1,365 x 11

tanggal <chr>	stasiun <chr>	pm10 <chr>	pm25 <chr>	so2 <chr>	co <chr>	o3 <chr>	no2 <chr>	max <dbl>
1/1/2021	DKI1 (Bunderan HI)	38.00	53.00	29.00	6.00	31.00	13.00	53
1/2/2021	DKI1 (Bunderan HI)	27.00	46.00	27.00	7.00	47.00	7.00	47
1/3/2021	DKI1 (Bunderan HI)	44.00	58.00	25.00	7.00	40.00	13.00	58
1/4/2021	DKI1 (Bunderan HI)	30.00	48.00	24.00	4.00	32.00	7.00	48
1/5/2021	DKI1 (Bunderan HI)	38.00	53.00	24.00	6.00	31.00	9.00	53
1/6/2021	DKI1 (Bunderan HI)	41.00	58.00	23.00	13.00	46.00	13.00	58
1/7/2021	DKI1 (Bunderan HI)	35.00	47.00	22.00	6.00	39.00	10.00	47
1/8/2021	DKI1 (Bunderan HI)	37.00	54.00	26.00	16.00	17.00	10.00	54
1/9/2021	DKI1 (Bunderan HI)	47.00	61.00	16.00	27.00	22.00	12.00	61
1/10/2021	DKI1 (Bunderan HI)	23.00	25.00	16.00	11.00	33.00	8.00	33

Gambar 4. 4 Proses Penggabungan Data

Pada Gambar 3.5 menjelaskan penggabungan data tiap bulan mulai Januari sampai dengan Oktober dan ditampilkan beberapa data setelah data digabung menjadi satu menjadi data set.

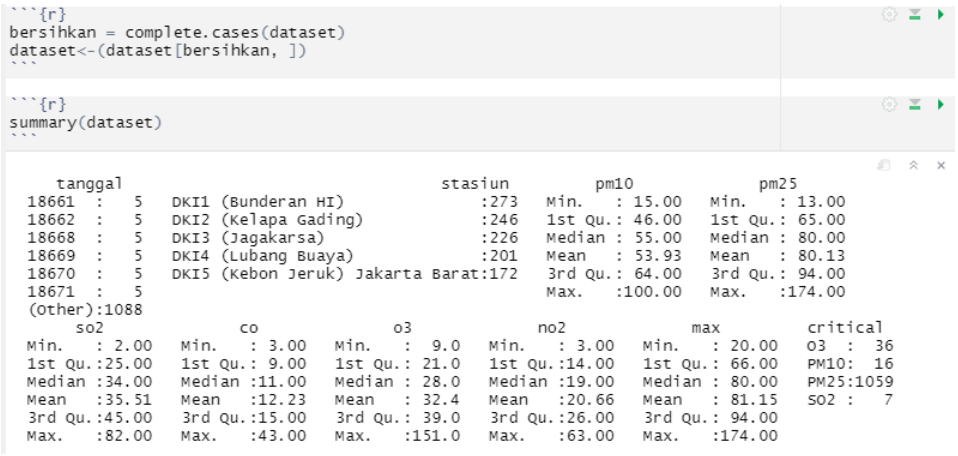
- Mengkonversi tipe data dari character ke numeric dan factor agar bisa diproses oleh metode Naïve Bayes. Adapun sintax dan hasil programnya seperti dijelaskan pada Gambar 4.5.

```{r}
dataset$tanggal<- as.factor(dataset$tanggal)
dataset$stasiun<- as.factor(dataset$stasiun)
dataset$pm10<- as.numeric(dataset$pm10)
dataset$pm25<- as.numeric(dataset$pm25)
dataset$so2<- as.numeric(dataset$so2)
dataset$co<- as.numeric(dataset$co)
dataset$o3<- as.numeric(dataset$o3)
dataset$no2<- as.numeric(dataset$no2)
dataset$critical<- as.factor(dataset$critical)
dataset$categori<- as.factor(dataset$categori)
```

Gambar 4. 5 Summary Data Setelah di konversi

Pada Gambar 4.5 menjelaskan ringkasan data setelah melalui konversi, tetapi disini terdapat NA’s yang artinya terjadi missing value dan harus dibersihkan, karena akan sangat mempengaruhi pada peritungan Naïve Bayes.

- Cleaning data atau menghapus baris yang kosong atau terdapat NA’s untuk menghindari kesalahan pada perhitungan. Data yang sudah melalui proses Cleaning kemudian dicek menggunakan summary dataset seperti diperlihatkan pada Gambar 4.6.



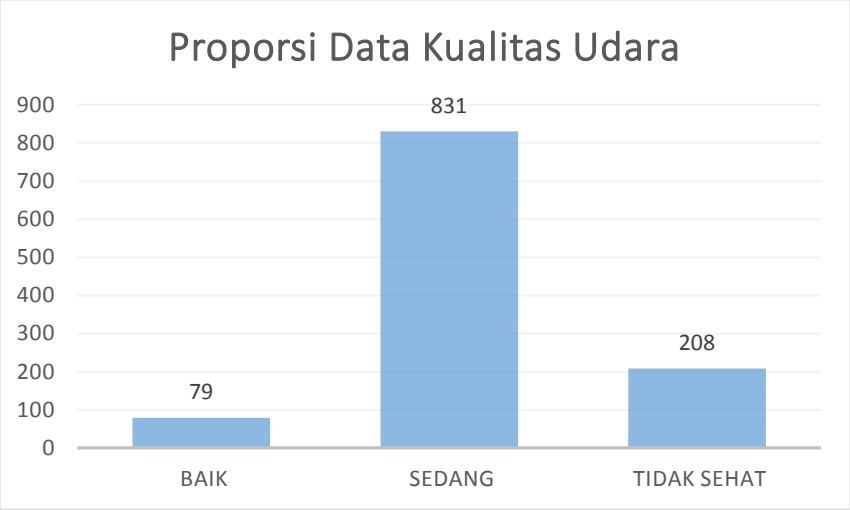
```{r}
bersihkan = complete.cases(dataset)
dataset<-(dataset[bersihkan, ])
```

```{r}
summary(dataset)
```

```
    tanggal                                  stasiun         pm10             pm25
 18661  :   5   DKI1 (Bunderan HI)              :273   Min.   : 15.00   Min.   : 13.00
 18662  :   5   DKI2 (Kelapa Gading)            :246   1st Qu.: 46.00   1st Qu.: 65.00
 18668  :   5   DKI3 (Jagakarsa)                :226   Median : 55.00   Median : 80.00
 18669  :   5   DKI4 (Lubang Buaya)             :201   Mean   : 53.93   Mean   : 80.13
 18670  :   5   DKI5 (Kebon Jeruk) Jakarta Barat:172   3rd Qu.: 64.00   3rd Qu.: 94.00
 18671  :   5                                          Max.   :100.00   Max.   :174.00
 (Other):1088
       so2              co               o3              no2              max          critical
 Min.   : 2.00   Min.   : 3.00   Min.   :  9.0   Min.   : 3.00   Min.   : 20.00   O3  :  36
 1st Qu.:25.00   1st Qu.: 9.00   1st Qu.: 21.0   1st Qu.:14.00   1st Qu.: 66.00   PM10:  16
 Median :34.00   Median :11.00   Median : 28.0   Median :19.00   Median : 80.00   PM25:1059
 Mean   :35.51   Mean   :12.23   Mean   : 32.4   Mean   :20.66   Mean   : 81.15   SO2 :   7
 3rd Qu.:45.00   3rd Qu.:15.00   3rd Qu.: 39.0   3rd Qu.:26.00   3rd Qu.: 94.00
 Max.   :82.00   Max.   :43.00   Max.   :151.0   Max.   :63.00   Max.   :174.00
```

Gambar 4. 6 Summary dataset Setelah Proses Data Cleaning

Pada Gambar 4.6 menjelaskan tentang data setelah melalui proses cleaning, dimana dapat dilihat bahwa data sudah bersih dari data missing value dan tidak terdapat NA’s, sedangkan proporsi datanya dijelaskan pada Gambar 4.7 di bawah ini.



Gambar 4. 7 Proporsi Data Kualitas Udara

Pada Gambar 4.7 menjelaskan tentang proporsi data terdiri atas kategori Baik sebanyak 79 data, Sedang sebanyak 831 data, dan Tidak Sehat sebanyak 208 data.

2. **Split Data** merupakan tahapan dalam mechine learning untuk membagi data training dan testing, kali ini data akan dibagi 75% untuk data training dan sisanya untuk data testing seperti diperlihatkan pada Gambar 4.8.

```
```{r}
split = (createDataPartition(y=dataset$categori, p=0.75, list=FALSE))
```

```{r}
training_set = dataset[split,]
test_set = dataset[-split,]
```

```{r}
dim(training_set)
```

[1] 840 11

```{r}
dim(test_set)
```

[1] 278 11
```

Gambar 4. 8 Spliting Data

Pada Gambar 4.8 mengimplementasikan proses splitting atau pembagian data yaitu data training dan data testing, hasil pembagiannya 840 data untuk training dan 278 untuk data testing, dengan 11 parameter.

3. Model Klasifikasi Naïve Bayes

Pada model klasifikasi Naïve Bayes memasukkan data training untuk memprediksi kualitas udara masuk ke kategori Baik, Sedang, atau Tidak Sehat. Dari dataset yang dimiliki, maka yang menjadi target variabel untuk diprediksi adalah categori, sedangkan variabel pm10, pm25, so2, co, o3, dan no2. Adapun script dan tampilan output pada R studio seperti diperlihatkan pada Gambar 4.9.

```
```{r}
#Pembuatan Model NaiveBayes
model_naive<- naiveBayes(categori ~pm10+pm25+so2+co+o3+no2, data = training_set)
print(model_naive)
```
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
       BAIK      SEDANG TIDAK SEHAT
 0.07142857  0.74285714  0.18571429

Conditional probabilities:
             pm10
Y                [,1]      [,2]
  BAIK         27.66667  6.024432
  SEDANG       51.98558 10.885577
  TIDAK SEHAT  70.37179  8.318823

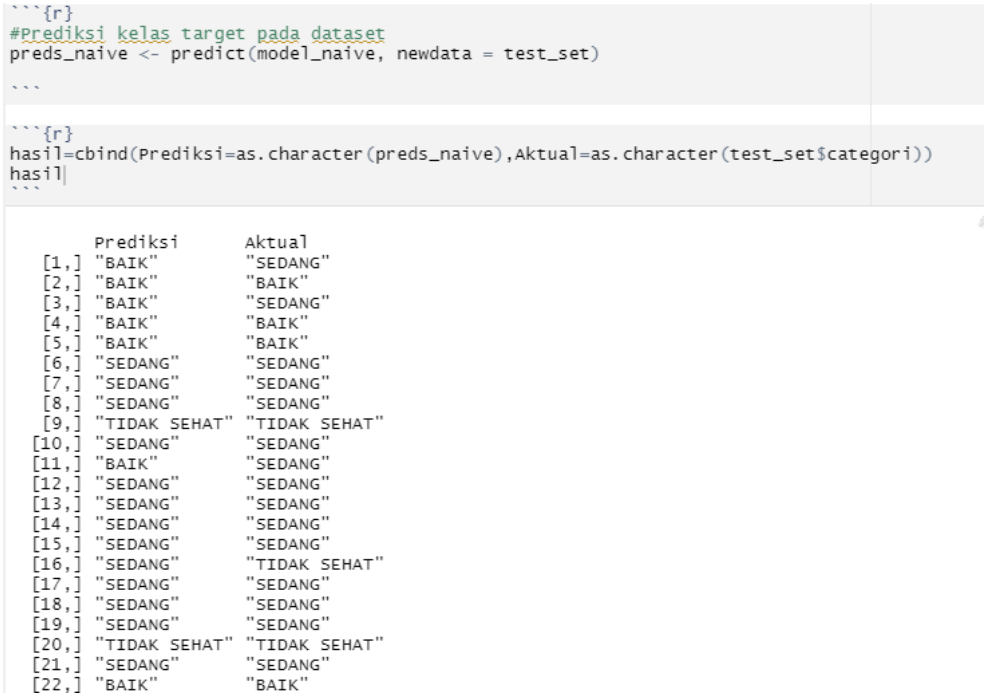
             pm25
Y                 [,1]      [,2]
  BAIK          40.93333  8.740839
  SEDANG        74.96795 14.886373
  TIDAK SEHAT  114.38462 12.877318

             so2
Y                [,1]     [,2]
  BAIK         25.21667 10.20318
  SEDANG       35.29808 13.14950
```

Gambar 4. 9 Hasil Model Naïve Bayes

Pada Gambar 4.8 menampilkan model Naïve Bayes dan hasil perhitungan nilai probabilitas atau peluang dari setiap kategorinya. Dari hasil model Naïve Bayes

selanjutnya akan digunakan untuk mempredikasi menggunakan data testing. Adapun model prediksi diperlihatkan pada Gambar 3.9 di bawah ini.



```r
#Prediksi kelas target pada dataset
preds_naive <- predict(model_naive, newdata = test_set)
```

```r
hasil=cbind(Prediksi=as.character(preds_naive),Aktual=as.character(test_set$categori))
hasil
```

```
      Prediksi      Aktual
 [1,] "BAIK"        "SEDANG"
 [2,] "BAIK"        "BAIK"
 [3,] "BAIK"        "SEDANG"
 [4,] "BAIK"        "BAIK"
 [5,] "BAIK"        "BAIK"
 [6,] "SEDANG"      "SEDANG"
 [7,] "SEDANG"      "SEDANG"
 [8,] "SEDANG"      "SEDANG"
 [9,] "TIDAK SEHAT" "TIDAK SEHAT"
[10,] "SEDANG"      "SEDANG"
[11,] "BAIK"        "SEDANG"
[12,] "SEDANG"      "SEDANG"
[13,] "SEDANG"      "SEDANG"
[14,] "SEDANG"      "SEDANG"
[15,] "SEDANG"      "SEDANG"
[16,] "SEDANG"      "TIDAK SEHAT"
[17,] "SEDANG"      "SEDANG"
[18,] "SEDANG"      "SEDANG"
[19,] "SEDANG"      "SEDANG"
[20,] "TIDAK SEHAT" "TIDAK SEHAT"
[21,] "SEDANG"      "SEDANG"
[22,] "BAIK"        "BAIK"
```

Gambar 4. 10 Model Prediksi Naïve Bayes

Pada Gambar 4.10 menjelaskan tentang prediksi model Naïve Bayes terhadap terhadap data test dan ditampilkan 22 data kondisi hasil prediksi dengan data actual. Dari sini menunjukkan bahwa metode Naïve Bayes dapat memprediksi kategori secara baik.

4.3 Ujicoba Metode Naïve Bayes

Melakukan pengujian dengan cara mengatur pembagian data training dan data testing untuk mengetahui performan dari model Naïve Bayes yaitu nilai *Accuracy, Precision, Recall*. Pengujian ini dilakukan untuk mengetahui seberapa pengaruh banyaknya data training dan testing terhadap klasifikasi yang dilakukan oleh Naïve Bayes. Adapun pengujian dilakukan dengan menggunakan

metode *Confusion Matrix*. Adapun model pengujiannya sebagai seperti di tampilkan pada Tabel 4.1.

Tabel 4. 1 Skenario Pengujian Model Naïve Bayes

Pengujian	Training%	Testing%
1	25	75
2	35	65
3	45	55
4	55	45
5	65	35
6	75	25
7	85	15
8	90	10

Pada Tabel 4.1 akan dilkukan pengujian terhadap model Naïve Bayes yang telah dibuat dengan dengan prosentase pembagian data training dan testing pada waktu proses splitting data. Proses pengujiannya akan dijabarkan satu persatu di bawah ini yaitu:

- **Pengujian 1**

Pada Pengujian 1 pembagian data training 25% dan data testing 75% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 4.2 di bawah ini:

Tabel 4. 2 Confusion Matrix Naïve Bayes Pengujian 1

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	52	38	0
	SEDANG	7	558	17
	TIDAK SEHAT	0	27	139

Dari Tabel 4.2 menampilkan tabel Confusion Matrix pada Pengujian 1 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.3

Tabel 4. 3 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 1

Accuracy	Precision	Recall
89.38%	79.13%	88.93%

Pada Tabel 4.3 diatas menunjukkan nilai accuracy sebesar 89.38%, nilai precision 79.13%, dan nilai recall 88.93%.

- **Pengujian 2**

Pada pengujian 2 pembagian data training 35% dan data testing 65% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 3.4 di bawah ini:

Tabel 4. 4 Confusion Matrix Naïve Bayes Pengujian 2

Naïve Bayes		Prediction		
		BAIK	**SEDANG**	**TIDAK SEHAT**
Aktual	**BAIK**	**48**	**42**	**0**
	SEDANG	**3**	**476**	**18**
	TIDAK SEHAT	**0**	**22**	**117**

Dari Tabel 3.4 menampilkan tabel Confusion Matrix pada Pengujian 2 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 3.5

Tabel 4. 5 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 2

Accuracy	Precision	Recall
88.29%	**89.64%**	**77.76%**

Pada Tabel 3.5 diatas menunjukkan nilai *accuracy* sebesar **88.29%**, nilai *precision* **89.64%**, dan nilai *recall* **77.76%**.

- **Pengujian 3**

Pada pengujian 3 pembagian data training 45% dan data testing 55% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 4.6 di bawah ini:

Tabel 4. 6 Confusion Matrix Naïve Bayes Pengujian 3

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	41	31	0
	SEDANG	1	305	17
	TIDAK SEHAT	0	20	79

Dari Tabel 4.6 menampilkan tabel Confusion Matrix pada pengujian 3 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.7

Tabel 4. 7 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 3

Accuracy	Precision	Recall
89.58%	91.25%	79.04%

Pada Tabel 4.7 diatas menunjukkan nilai *accuracy* sebesar 89.58%, nilai *precision* 91.25%, dan nilai *recall* 79.04%.

- **Pengujian 4**

Pada pengujian 4 pembagian data training 55% dan data testing 45% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 4.8 di bawah ini:

Tabel 4. 8 Confusion Matrix Naïve Bayes Pengujian 4

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	32	24	0
	SEDANG	3	335	12
	TIDAK SEHAT	0	14	81

Dari Tabel 3.8 menampilkan tabel Confusion Matrix pada pengujian 4 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.9

Tabel 4. 9 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 4

Accuracy	Precision	Recall
89.42%	89.45%	79.37%

Pada Tabel 4.9 diatas menunjukkan nilai *accuracy* sebesar 89.42%, nilai *precision* 89.45%, dan nilai *recall* 79.37%.

- **Pengujian 5**

Pada pengujian 5 pembagian data training 65% dan data testing 35% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 4.10 di bawah ini:

Tabel 4. 10 Confusion Matrix Naïve Bayes Pengujian 5

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	26	18	0
	SEDANG	1	257	7
	TIDAK SEHAT	0	15	65

Dari Tabel 4.10 menampilkan tabel Confusion Matrix pada pengujian 5 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.11

Tabel 4. 11 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 5

Accuracy	Precision	Recall
89.46%	91.73%	79.11%

Pada Tabel 4.11 diatas menunjukkan nilai *accuracy* sebesar 89.46%, nilai *precision* 79.11%, dan nilai *recall* 91.73%.

- **Pengujian 6**

Pada pengujian 6 pembagian data training 75% dan data testing 25% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 3.10 di bawah ini:

Tabel 4. 12 Confusion Matrix Naïve Bayes Pengujian 6

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	18	14	0
	SEDANG	1	186	4
	TIDAK SEHAT	0	7	48

Dari Tabel 4.12 menampilkan tabel Confusion Matrix pada pengujian 6 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.13

Tabel 4. 13 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 6

Accuracy	Precision	Recall
90.65%	92.30%	80.30%

Pada Tabel 4.13 diatas menunjukkan nilai *accuracy* sebesar 90.65%, nilai *precision* 92.30%, dan nilai *recall* 80.30%.

- **Pengujian 7**

Pada pengujian 7 pembagian data training 85% dan data testing 15% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 3.14 di bawah ini:

Tabel 4. 14 Confusion Matrix Naïve Bayes Pengujian 7

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	10	10	0
	SEDANG	1	106	1
	TIDAK SEHAT	0	8	30

Dari Tabel 4.14 menampilkan tabel Confusion Matrix pada pengujian 7 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.15

Tabel 4. 15 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 7

Accuracy	Precision	Recall
87.95%	91.06%	75.70%

Pada Tabel 4.15 diatas menunjukkan nilai *accuracy* sebesar 87.95%, nilai *precision* 91.06%, dan nilai *recall* 75.70%.

- **Pengujian 8**

Pada pengujian 8 pembagian data training 90% dan data testing 10% dihasilkan Confusion Matrix seperti dijelaskan pada Tabel 3.16 di bawah ini:

Tabel 4. 16 Confusion Matrix Naïve Bayes Pengujian 7

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	6	7	0
	SEDANG	1	72	1
	TIDAK SEHAT	0	4	19

Dari Tabel 4.16 menampilkan tabel Confusion Matrix pada pengujian 8 yang menghasilkan nilai *Accuracy, Precision, Recall* yaitu seperti dijelaskan pada Tabel 4.17

Tabel 4. 17 Accuracy, Precision, Recall Naïve Bayes pada Pengujian 8

Accuracy	Precision	Recall
88.18%	89.15%	75.35%

Pada Tabel 4.17 diatas menunjukkan nilai *accuracy* sebesar 88.18%, nilai *precision* **89.15%**, dan nilai *recall* **75.35%.**

- **Kesimpulan**

Dari hasil pengujian yang telah dilakukan sebanyak 8 kali pengujian didapatkan nilai rata-rata *Accuracy, Precision, Recall* seperti ditampilkan pada Tabel 4.18 sebagai berikut:

Tabel 4. 18 Rata-rata Accuracy, Precision, Recall pada Naïve Bayes

Pengujian	Training%	Testing%	Accuracy%	Precision%	Recall%
1	25	75	89.38%	88.93%	79.13%
2	35	65	88.29%	89.64%	77.76%
3	45	55	89.58%	91.25%	79.04%
4	55	45	89.42%	89.45%	79.37%
5	65	35	89.46%	91.73%	79.11%
6	75	25	90.65%	92.30%	80.30%
7	85	15	87.95%	91.06%	75.70%
8	90	10	88.18%	89.15%	75.35%
Rata-rata			89.11%	90.44%	78.22%

Pada Tabel 4.18 terlihat bahwa model Naïve bayes mampu melakukan klasifikasi pada data kualitas udara Jakarta 2021. Dengan 8 kali percobaan dengan melakukan perubahan pada data training dan testing menunjukkan nilai rata-rata Accuracy sebesar **89.11%**, Precision sebesar 90.44%, dan Recall sebesar 78.22%.

BAB V

PEMBAHASAN

Pada pembahasan ini akan dilakukan perbandingan dan analisa hasil klasifikasi dari dua metode yaitu Support Vector Machine (SVM) dan Naïve Bayes (NBC) terhadap data kualitas udara yang diperoleh dari website resmi https://data.jakarta.go.id/ mulai bulan Januari sampai bulan Oktober. Untuk dapat diimplementasikan pada metode Support Vector Machine dan Naïve Bayes maka data yang diperoleh tidak bisa langsung diproses klasifikasi. Oleh karena itu data harus diolah atau melalui pre proses terlebih dahulu dikarenakan data yang didapat dari website bertipe karakter. Pre proses yang dilakukan antara lain menghilangkan data yang kosong, menghilangkan data NA, dan mengkonversi tipe data menjadi numeric dan factor kemudian baru dapat diproses oleh kedua metode. Dataset yang digunakan untuk analisis ini adalah data data kualitas udara setelah melalui preprocessing sebanyak 1118 data dan 11 variabel, dimana 4 variabel kategorik, 1 target class dan 7 variabel bebas tetapi yang digunakan untuk klasifikasi hanya 6 variabel, sedangkan proporsi datanya terdiri atas kategori Baik sebanyak 79 data, Sedang sebanyak 831 data, dan Tidak Sehat sebanyak 208 data.

Dalam menentukan kualitas udara model yang digunakan dalam klasifikasi yaitu Support Vector Machine (SVM) dan Naïve Bayes, serta mencari berapa tingkat akurasinya dari kedua metode ini yaitu dari data ini akan diprediksi apakah kualitas udara yang baru terdeteksi berkualitas Baik, Sedang, ataupun Tidak Sehat berdasarkan 6 variabel prediksi.

5.1 Hasil Pengujian Klasifikasi dengan Support Vector Machine

Pada pengujian klasifikasi menggunakan metode SVM dilakukan dengan 8 kali percobaan dan menghasilkan nilai akurasi yang paling tinggi yaitu pada pengujian ke 5 menggunakan 1118 data kualitas udara Jakarta dengan pembagian data training 65% dan data testing 35% yaitu 729 data training dan 389 data testing. Berikut ini hasil Confusion matrix pada hasil dengan akurasi tertinggi dijelaskan pada Tabel 5.1

Tabel 5. 1 Konfusion Matrix SVM

SVM		**Prediction**		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	18	0	0
	SEDANG	9	289	6
	TIDAK SEHAT	0	1	66

Pada hasil klasifkasi pada pengujian ke lima yaitu menggunakan perbandingan prosentase data training 65% dan data testing 35% dengan perbandingan data training sebanyak 729 data serta data testing sebanyak 389 data. Pada Tabel 5.1 terdapat 18 data yang diklasifikasi “Baik” dan 289 data yang diklasifikasikan “Sedang” serta 66 data yang diklasifikasin “Tidak Sehat” serta ada beberapa kesalahan yaitu 9 data yang seharusnya “Sedang” diklasifikasin “Baik” dan 6 data diklasifikasin “Tidak Sehat” yang seharusnya “Sedang”. Selanjutnya dihitung nilai akurasi klasifikasi dari model Support Vector Machine yaitu sebagi berikut:

Tabel 5. 2 Confusion Matrix 3x3

		Hasil Prediksi		
		BAIK	SEDANG	TIDAK SEHAT
Hasil Aktual	BAIK	TA	FB1	FC1
	SEDANG	FA1	TB	FC2
	TIDAK SEHAT	FA2	FB2	TC

Hasil dari model klasifikasi dengan 3 kelas kategori yaitu melalui *Confusion Matrix* 3x3 maka diperoleh nilai akurasi, precision, dan recall seperti dijabarkan pada rumus di bawah ini:

$$Akurasi = \frac{T}{T + FA1 + FA2 + FB1 + \mathrm{FB2} + FC1 + FC2}\ x\ 100\%$$

Dimana T= penjumlahan TA+TB+TC

$$Akurasi = \frac{18 + 289 + 66}{(373 + 9 + 0 + 0 + 1 + 0 + 6)}\ x\ 100\%$$

$$Akurasi = \frac{373}{389}\ x\ 100\%$$

$Akurasi$ =95.89%

Sementara itu untuk nilai precisionnya adalah sebagai berikut:

$$Precision = \frac{TP}{TP + FP}\ x\ 100\%$$

Dimana:

TP kelas BAIK = 18, TP kelas SEDANG = 289, TP kelas TIDAK SEHAT = 66, FP untuk kelas BAIK yaitu FA1+FA2, FP kelas BAIK=9+10, FP untuk kelas SEDANG yaitu FB1+FB2, FP kelas SEDANG=0+1, FP untuk kelas TIDAK SEHAT yaitu FC1+ FC2, FP kelas TIDAK SEHAT=0+6.

Terlebih dahulu harus dicari nilai precision masing-masing kelas, kemudian hasilnya dijumlahkan dan dicari nilai rata-ratanya

$$Precision\ kelas\ Baik = \frac{18}{18+9}\ x\ 100\%$$

$Precision\ kelas\ Baik$ =66,67%

$$Precision\ kelas\ Sedang = \frac{289}{289+1}\ x\ 100\%$$

$Precision\ kelas\ Sedang$ = 99,66%

$$Precision\ kelas\ Tidak\ Sehat = \frac{66}{66+1}\ x\ 100\%$$

$Precision\ kelas\ Tidak\ Sehat$ =91.67%

$Jumlah\ semua\ Precision\ kelas$ = (66,67%+ 99,66%+ 91.67%) / 3

$Jumlah\ semua\ Precision\ kelas$ = (257.99%) / 3

$Jumlah\ semua\ Precision\ kelas$ =86.00%

Sementara itu untuk nilai Recall atau Sensitivity adalah sebagai berikut:

$$\text{Recall atau Sensitivity} = \frac{TP}{TP+FN}\ x\ 100\%$$

Dimana:

TP kelas BAIK = 18, TP kelas SEDANG = 289, TP kelas TIDAK SEHAT = 66, FN untuk kelas BAIK yaitu FB1+FC1, FN kelas BAIK=0+0, FN untuk kelas SEDANG yaitu FA1+FC2, FN kelas SEDANG=9+6, FN untuk kelas TIDAK SEHAT yaitu FA2+FB2, FN kelas TIDAK SEHAT=0+1.

Terlebih dahulu harus dicari nilai Recall atau Sensitivity masing-masing kelas, kemudian hasilnya dijumlahkan dan dicari nilai rata-ratanya

Recall $kelas\ Baik = \frac{18}{18+0}\ x\ 100\%$

Recall $kelas\ Baik$ =100,00%

Recall $kelas\ Sedang = \frac{289}{289+15}\ x\ 100\%$

Recall $kelas\ Baik$ = 95,07%

Recall $kelas\ Tidak\ Sehat = \frac{66}{66+1}\ x\ 100\%$

Recall $kelas\ Tidak\ Sehat$ =98.51%

$Jumlah\ semua$ Recall $kelas$ = (100,00% + 95,07% + 98.51%) / 3

$Jumlah\ semua$ Recall $kelas$ = (293.57%) / 3

$Jumlah\ semua$ Recall $kelas$ =97.86%

Untuk hasil keseluruhan pengujian klasifikasi yang dilakukan sebanyak 8 kali pengujian dalam menentukan kualitas udara yang dilakukan dengan menggunakan Algoritma Suport Vector Machine dijelaskan pada Tabel 5.1 berikut ini

Tabel 5. 3 Hasil Pengujian SVM

Pengujian	Training	Testing	Data Training	Data Testing	Jml Vector	Accuracy	Presisi	Recall
1	25%	75%	280	838	107	95.35%	90.46 %	89.01 %
2	35%	65%	392	726	144	95.18%	89.66 %	90.31 %
3	45%	55%	504	614	170	95.77%	89.72 %	91.16 %
4	55%	45%	617	501	192	95.01%	88.06 %	91.10 %
5	65%	35%	729	389	223	95.89%	86.00 %	97.86 %
6	75%	25%	840	278	246	95.68%	85.95 %	94.75 %
7	85%	15%	952	166	261	95.18%	86.00 %	92.43 %
8	90%	10%	1008	110	272	95.45%	80.55 %	90.12 %
Rata-rata						95.44%	87.05 %	92.09 %

Pada Tabel 5.3 dapat diketahui bahwa dalam penelitian ini Algortima Support Vector Machine telah berhasil melakukan prediksi dan klasifikasi dalam menetukan kualitas udara Jakarta tahun 2021. Klasifikasi yang dilakukan dengan Algortima Support Vector Machine dengan fungsi kernel linear mendapatkan hasil dari uji coba menyatakan bahwa prosentase pembagian data training dan data testing sangat mempengaruhi hasil akurasi, presisi, dan recall dalam proses pengujian. Uji coba dilakukan sebanyak 8 kali percobaan dengan scenario pembagian data training dan data testing secara berbeda dalam setiap tahap percobaannya. Pada hasil pengujian yang dilakukan menunjukkan bahwa nilai akurasi tertinggi didapat pada pengujian ke 5 dengan prosentase 65% data training yaitu sebanyak 729 data, sedangkan 35% yaitu sebanyak 389 data testing dengan menghasilkan akurasi **95.89%,** presisi **86.00%,** Recall **97.86%**, serta rata-rata akurasi **95.44%.**

5.2 Hasil Pengujian Naïve Bayes

Pada Pengujian Naïve Bayes dilakukan dengan 8 kali dan menghasilkan nilai akurasi yang paling tinggi yaitu pada pengujian ke 6 dengan perbandingan data training 75% dan data testing 25% atau 840 data training dan 278 data testing. Berikut ini hasil Confusion matrix pada hasil dengan akurasi tertinggi dijelaskan pada Tabel 5.1

Tabel 5. 4 Confusion Matrix Naïve Bayes

Naïve Bayes		Prediction		
		BAIK	SEDANG	TIDAK SEHAT
Aktual	BAIK	18	14	0
	SEDANG	1	186	4
	TIDAK SEHAT	0	7	48

Pada hasil klasifkasi pada pengujian ke enam yaitu menggunakan perbandingan prosentase data training 75% dan data testing 25% dengan perbandingan data training sebanyak 840 data serta data testing sebanyak 278 data. Pada Tabel 5.1 terdapat 18 data yang diklasifikasi “Baik” dan 186 data yang diklasifikasikan “Sedang” serta 48 data yang diklasifikasin “Tidak Sehat” serta ada beberapa kesalahan yaitu 6 data yang seharusnya “Sedang” diklasifikasin “Baik” dan 6 data diklasifikasin “Tidak Sehat” yang seharusnya “Sedang”. Selanjutnya dihitung nilai akurasi klasifikasi dari model Support Vector Machine yaitu sebagi berikut:

Tabel 5. 5 Confusion Matrix 3x3

		Hasil Prediksi		
		BAIK	SEDANG	TIDAK SEHAT
Hasil Aktual	BAIK	TA	FB1	FC1
	SEDANG	FA1	TB	FC2
	TIDAK SEHAT	FA2	FB2	TC

Berdasarkan model klasifikasi dengan 3 kelas kategori yaitu melalui *Confusion Matrix* 3x3 maka diperoleh nilai akurasi, precision, dan recall seperti dijabarkan pada rumus di bawah ini:

$$Akurasi = \frac{T}{T + FA1 + FA2 + FB1 + FB2 + FC1 + FC2}\ x\ 100\%$$

$$Akurasi = \frac{18 + 186 + 48}{(252 + 1 + 0 + 14 + 7 + 0 + 4)}\ x\ 100\%$$

$$Akurasi = \frac{252}{278}\ x\ 100\%$$

$$Akurasi = 90.65\%$$

Sementara itu untuk nilai precisionnya adalah sebagai berikut:

$$Precision = \frac{TP}{TP + FP}\ x\ 100\%$$

Dimana:

TP kelas BAIK = 18, TP kelas SEDANG = 186, TP kelas TIDAK SEHAT = 48, FP untuk kelas BAIK yaitu FA1+FA2, FP kelas BAIK=1+0, FP untuk kelas SEDANG yaitu FB1+FB2, FP kelas SEDANG=14+7, FP kelas TIDAK SEHAT yaitu FC1+FC2, FP kelas TIDAK SEHAT =0+4.

Terlebih dahulu harus dicari nilai precision masing-masing kelas, kemudian hasilnya dijumlahkan dan dicari nilai rata-ratanya

$$Precision\ kelas\ Baik = \frac{18}{18+1}\ x\ 100\%$$

$$Precision\ kelas\ Baik = 94.74\%$$

$$Precision\ kelas\ Sedang = \frac{186}{186+21}\ x\ 100\%$$

$$Precision\ kelas\ Sedang = 89.86\%$$

$$Precision\ kelas\ Tidak\ Sehat = \frac{48}{48+4}\ x\ 100\%$$

$$Precision\ kelas\ Tidak\ Sehat = 92.31\%$$

$$Jumlah\ semua\ Precision\ kelas = (94.74\% + 89.86\% + 92.31\%) / 3$$

$$Jumlah\ semua\ Precision\ kelas = (276.90\%) / 3$$

$$Jumlah\ semua\ Precision\ kelas = 92.30\%$$

Sementara itu untuk nilai Recall atau Sensitivity adalah sebagai berikut:

$$\text{Recall atau Sensitivity} = \frac{TP}{TP+FN}\ x\ 100\%$$

Dimana:

TP kelas BAIK = 18, TP kelas SEDANG = 186, TP kelas TIDAK SEHAT = 48, FN kelas BAIK yaitu FB1+FC1, FN kelas BAIK=14+0, FN kelas SEDANG yaitu FA1+FC2, FN kelas SEDANG =1+4, FN kelas TIDAK SEHAT yaitu FA2+FB2 , FN kelas TIDAK SEHAT=0+7.

Terlebih dahulu harus dicari nilai Recall/Sensitivity masing-masing kelas, kemudian hasilnya dijumlahkan dan dicari nilai rata-ratanya

Recall $kelas\ Baik = \frac{18}{18+14}\ x\ 100\%$

Recall $kelas\ Baik = 56.25\%$

Recall $kelas\ Sedang = \frac{186}{186+5}\ x\ 100\%$

Recall $kelas\ Baik = 97.38\%$

Recall $kelas\ Tidak\ Sehat = \frac{48}{48+7}\ x\ 100\%$

Recall $kelas\ Tidak\ Sehat = 87.27\%$

$Jumlah\ semua$ Recall $kelas =$ (56.25%+ 97.38%+ 87.27%) / 3

$Jumlah\ semua$ Recall $kelas =$ (240.90%) / 3

$Jumlah\ semua$ Recall $kelas$ =80.30%

Untuk hasil keseluruhan pengujian klasifikasi dalam menentukan kualitas udara yang dilakukan dengan menggunakan Algoritma Naïve Bayes dijelaskan pada Tabel berikut ini

Tabel 5. 6 Hasil Pengujian Naïve Bayes

Pengujian	Training	Testing	Data Training	Data Testing	Accuracy	Presisi	Recall
1	25%	75%	280	838	89.38%	88.93%	79.13%
2	35%	65%	392	726	88.29%	89.64%	77.76%
3	45%	55%	504	614	89.58%	91.25%	79.04%
4	55%	45%	617	501	89.42%	89.45%	79.37%
5	65%	35%	729	389	89.46%	91.73%	79.11%
6	75%	25%	840	278	90.65%	92.30%	80.30%
7	85%	15%	952	166	87.95%	91.06%	75.70%
8	90%	10%	1008	110	88.18%	89.15%	75.35%
Rata-rata					89.11%	90.44%	78.22%

Pada Tabel 5.6 tentang perbandingan hasil pengujian menyatakan bahwa dari hasil pengamatan uji coba yang telah dilakukan 8 kali percobaan yaitu dengan merubah prosentase data training mendapatkan hasil yang menyatakan bahwa prosentase pembagian data training dan data testing sangat mempengaruhi hasil akurasi, presisi, dan recall dalam proses pengujian. Pada hasil pengujian yang dilakukan menunjukkan bahwa nilai akurasi tertinggi didapat pada pengujian ke 6 dengan prosentase 75% data training yaitu sebanyak 840 data, sedangkan 25% yaitu sebanyak 278 data testing dengan menghasilkan akurasi 90.65%**,** presisi 92.30%**,** Recall 80.30%, serta rata-rata akurasi **95.44%** mendapatkan hasil dari uji coba menyatakan bahwa prosentase pembagian data training dan data testing sangat mempengaruhi hasil akurasi, presisi, dan recall dalam proses pengujian. Uji coba dilakukan sebanyak 8 kali percobaan dengan scenario pembagian data training dan data testing secara berbeda dalam setiap tahap percobaannya. Pada hasil pengujian yang dilakukan menunjukkan bahwa nilai akurasi tertinggi didapat pada pengujian ke 5 dengan prosentase 65% data training yaitu sebanyak 729 data, sedangkan 35% yaitu sebanyak 389 data testing dengan menghasilkan akurasi **95.89%,** presisi **86.00%,** Recall **97.86%**, serta rata-rata akurasi 89.11%**.**

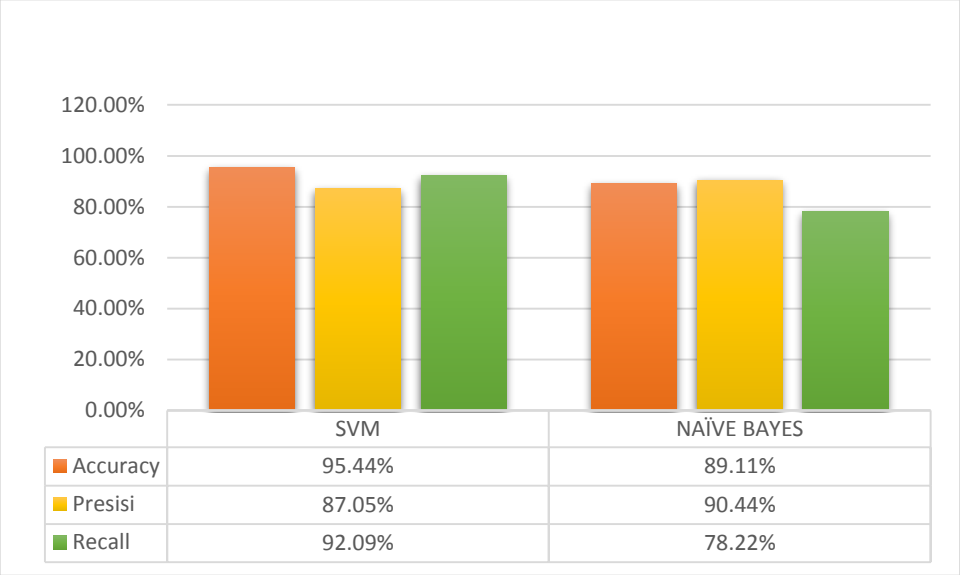
5.3 Hasil Perbandingan SVM dan Naïve Bayes

Dari hasil pengujian masing-masing metode yang dilakukan sebanyak 8 kali pengujian terhadap masing-masing metode, maka selanjutkan dilakukan perbandingan antara Support Vector Machine dengan metode Naïve Bayes seperti pada Tabel 5.7 di bawah ini.

Tabel 5. 7 Hasil Perbandingan Pengujian SVM dan Naïve Bayes

Training	Testing	SVM			NAÏVE BAYES		
		Accuracy	Presisi	Recall	Accuracy	Presisi	Recall
25%	**75%**	**95.35%**	**90.46%**	**89.01%**	89.38%	88.93%	79.13%
35%	**65%**	**95.18%**	89.66%	90.31%	88.29%	89.64%	77.76%
45%	**55%**	**95.77%**	89.72%	91.16%	89.58%	91.25%	79.04%
55%	**45%**	**95.01%**	88.06%	91.10%	89.42%	89.45%	79.37%
65%	**35%**	**95.89%**	86.00%	97.86%	89.46%	91.73%	79.11%
75%	**25%**	**95.68%**	85.95%	94.75%	90.65%	92.30%	80.30%
85%	**15%**	**95.18%**	86.00%	92.43%	87.95%	91.06%	75.70%
90%	**10%**	**95.45%**	80.55%	90.12%	88.18%	89.15%	75.35%
Rata-rata		**95.44%**	**87.05%**	**92.09%**	**89.11%**	**90.44%**	**78.22 %**

Berdasarkan Gambar 5.7 menunjukkan bahwa perbandingan data training sangat mempengaruhi nilai akurasi, presisi dan recall dari kedua metode. Dari hasil ujicoba yang dilakukan bahwa tingkat akurasi klasifikasi kualitas udara Jakarta tahun 2021 antara algoritma Support Vector Machine dan Naïve Bayes bahwa nilai rata- rata akurasi Support Vector Machine mengungguli dari Naïve Bayes yaitu 95.44% sedangkan Nilai rata-rata akurasi Naïve Bayes sebesar 89.11%, sedangkan untuk nilai rata-rata presisi algoritma Naïve Bayes mengungguli Support Vector Machine yaitu 90.44% sedangkan Support Vector Machine 87.05%. dari hasil pengujian menunjukkan bahwa semakin bertambah data training maka semakin besar pula nilai akurasinya. Hasil akurasi, presisi, recall keseluruhan dapat dilihat pada grafik seperti pada Gambar 5.1 di bawah ini



Gambar 5. 1 Grafik Perbandingan Akurasi, Presisi, Recall

Pada Gambar 5.1 menjelaskan tentang hasil nilai rata-rata Accuracy pada metode Support Vector Machine menghasilkan nilai rata-rata 95.44%, Precision 87.05%, Recall 92.09%, sedangkan metode Naïve Bayes rata-rata akurasi 89.11%, Precision 90.44%, Recall 78.22%. Pada Gambar 5.1 dapat dilihat bahwa dalam melakukan klasifikasi kualitas udara Jakarta tahun 2021 Accuracy metode Support Vector Machine (SVM) lebih tinggi dibandingkan dengan hasil Naïve Bayes, ini menunjukkan bahwa dalam kasus penelitian klasifikasi kualitas udara Jakarta tahun 2021 ini lebih metode Support Vector Machine (SVM) lebih unggul.

BAB VI

KESIMPULAN

7.1 Kesimpulan

Pada perbandingan Support Vector Machine dan Nave Bayes dalam mengkategorikan kualitas udara Jakarta pada tahun 2021, adalah sebagai berikut:

1. Model pada prediksi dengan klasifikasi kualitas udara Jakarta tahun 2021 berhasil diterapkan pada Support Vector Machine dan Naïve Bayes menggunakan 6 parameter.
2. Akurasi pada klasifikasi Naïve Bayes hasilnya lebih rendah dibandingkan SVM.
3. Support Vector Machine (SVM) dapat digunakan sebagai salah satu metode prediksi kualitas udara Jakarta tahun 2021 dengan memperoleh rata-rata prosentase akurasi 95.44%, Precision 87.05%, Recall 92.09%
4. Pada penelitian Prediksi untuk klasifikasi dalam menentukan kualitas udara Jakarta tahun 2021, metode SVM memiliki performa ketepatan prediksi lebih baik dibanding dengan metode Naive Bayes yang hanya dapat memperoleh rata-rata akurasi 89.11%, Precision 90.44%, Recall 78.22%.

7.2 Saran

Saran penelitian ini dapat dikembangkan lagi sehingga hasilnya lebih baik diantaranya:

1. Pentingnya mengetahui struktur dataset yang digunakan dalam penelitian karena data tersebut dapat mengubah keakuratan hasil.

2. Untuk pengembangan selanjutnya perlu adanya data yang lebih besar dan proporsi data yang seimbang dalam menghasilkan akurasi lebih baik.

3. Pada penelitian berikutnya dapat dilakukan dengan metode klasifikasi yang lainnya.

4. Pada penelitian kedepannya dapat dikembangkan dengan objek yang lain seperti pada kualitas air, kualitas tanah dan lain-lain.

DAFTAR PUSTAKA

Anwar, Syaiful, and M. Syafrullah. 2016. "Klasifikasi Kerusakan Kawasan Konservasi Dengan Metode Support Vector Machine (Svm) Menggunakan Kernel Gaussian : Studi Kasus the Nature Conservancy." *Jurnal TELEMATIKA MKOM* 8(2):89–96.

Choy, Garry, Omid Khalilzadeh, Mark Michalski, Synho Do, Anthony E. Samir, Oleg S. Pianykh, J. Raymond Geis, Pari V. Pandharipande, James A. Brink, and Keith J. Dreyer. 2018. "Current Applications and Future Impact of Machine Learning in Radiology." *Radiology* 288(2):318–28. doi: 10.1148/radiol.2018171820.

Febriantono, M. Aldiki, Ridho Herasmara, and Gusti Pangestu. 2021. "Cost Sensitive Tree Dan Naïve Bayes Pada Klasifikasi Multiclass." *Jurnal Informatika Polinema* 7(2):57–64. doi: 10.33795/jip.v7i2.533.

Handayani, Ade Silvia, Sopian Soim, Theresia Enim Agusdi, and Nyayu Latifah Husni. 2021. "Air Quality Classification Using Support Vector Machine." 10(1):55–69.

Helmi, Muhammad Iqbal Yunan. 2021. "DIAGNOSIS PENDERITA PENYAKIT KANKER PARU MENGGUNAKAN SUPPORT VECTOR MACHINE DAN NAÏVE BAYES."

Hermawan, Aditya. 2019. "NASKAH PUBLIKASI SPKU : SISTEM PREDIKSI KUALITAS UDARA (STUDI KASUS : DKI JAKARTA) SPKU : SISTEM PREDIKSI KUALITAS UDARA (STUDI KASUS : DKI JAKARTA)."

Irmanda, Helena Nurramdhani, and Ria Astriratma. 2020. "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM)." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 4(5):915–22. doi: 10.29207/resti.v4i5.2313.

Lishania, Irene, Rito Goejantoro, and Yuki Novia Nasution. 2019. "Perbandingan Klasifikasi Metode Naive Bayes Dan Metode Decision Tree Algoritma (J48) Pada Pasien Penderita Penyakit Stroke Di RSUD Abdul Wahab Sjahranie Samarinda." *Jurnal Eksponensial* 10(2):135–42.

Nurdin, Ade Silvia Handayani; Sopian Soim; Theresia Enim Agusdi; Rumiasih; Ali. 2020. "KLASIFIKASI KUALITAS UDARA DENGAN METODE SUPPORT VECTOR MACHINE." *Jire* 3(1):48–57.

Peraturan Pemerintah RI. 2020. *Peraturan Menteri Lingkungan Hidup Dan Kehutanan Republik Indonesia No 14 Tahun 2020 Tentang Indeks Standar Pencemaran Udara*.

Perspektif, Hidup, and Al-Q. U. R. An. 2019. "PENDIDIKAN PELESTARIAN LINGKUNGAN HIDUP PERSPEKTIF AL- QUR'AN Syahrul Munir." I(2):199–213.

Prasetyo, Eko. 2012. *Data Mining Konsep Dan Aplikasi Menggunakan Matlab*. edited by N. WK. Yogyakarta: ANDI.

Qomarullah, Muhammad. n.d. "LINGKUNGAN DALAM KAJIAN AL-QUR ` AN : Krisis Lingkungan Dan Penanggulangannya Perspektif Al-Qur ` An."

Rahmasari, Bekti. 2017. "Kebersihan Dan Kesehatan Lingkungan Dalam Perspektif Hadis Oleh :"

Rakhmalia, Riza Indriani. 2018. "Perbandingan Hasil Metode Naïve Bayes

Classifier Dan Support Vector Machine Dalam Klasifikasi Cerah Hujan."

Ramasubramanian, Karthik, and Abhishek Singh. 2019. *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R*.

Rita, Diah Dwiana Lestiani, Esrom Hamonangan, Muhayatun Santoso, and Hernani Yulinawati. 2016. "Air Quality (PM10 Dan PM2.5) For Completing the Enviromental Quality Index." *Ecolab* 10(1):1–7.

Simangunsong, JUANTO. 2019. "Universitas Sumatera Utara."

Supriyatna, Adi, and Wida Prima Mustika. 2018. "Komparasi Algoritma Naive Bayes Dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil." *J-SAKTI (Jurnal Sains Komputer Dan Informatika)* 2(2):152. doi: 10.30645/j-sakti.v2i2.78.

Syihabuddin Azmil Umri, Syekh. 2021. "Analisis Dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di Dki Jakarta." *JIKO (Jurnal Informatika Dan Komputer)* 4(2):98–104. doi: 10.33387/jiko.v4i2.2871.

Wicahyo, Amri, Ahmad Pudoli, and Dewi Kusumaningsih. 2021. "Penggunaan Algoritma Naive Bayes Dalam Klasifikasi Pengaruh Pencemaran Udara." *Jurnal ICT : Information Communication & Technology* 20(1):103–8.