PEMBELAJARAN MESIN PADA DATA PREPROCESSING DENGAN METODE PRINCIPAL COMPONENT ANALYSIS DAN SMOTE

Giovano Trihade Putra
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
giovanoputra@student.telkomuniversity.ac.id

Steven Mulya Manik
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
stevenmulyamanik@student.telkomuniversity.ac.id

Ig. Prasetya Dwi Wibawa
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
prasdwibawa@telkomuniversity.ac.id

Meta Kallista
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
metakallista@telkomuniversity.ac.id

Abstrak*— Pencemaran udara merupakan dampak negatif dari aktivitas manusia terhadap lingkungan. Udara yang tercemar oleh partikel yang berbahaya dapat membahayakan lingkungan, menyebabkan masalah pernafasan pada manusia, pemanasan global dan mempengaruhi metabolisme lingkungan. Indonesia menduduki peringkat ke-17 pada tahun 2022 dalam indeks kualitas udara (AQI) Internasional. Untuk menurunkan tingkat pencemaran udara, diperlukan adanya kesadaran masyarakat tentang pencemaran udara. Kota Bandung adalah kota besar di Indonesia yang berlokasi dekat ibu kota negara. Lokasi geografis yang strategis membuat kota ini cocok untuk ditempati, namun juga memiliki tingkat populasi dan aktivitas industri yang tinggi, yang menyebabkan tingginya tingkat pencemaran udara. Walaupun Kota Bandung sudah memiliki sistem pemantauan kualitas udara, sistem ini hanya memantau parameter umum seperti ozon, monoksida karbon, dan partikulat. Oleh karena itu, perlu adanya kesadaran masyarakat tentang tingkat pencemaran udara dengan menggunakan Indeks Standar Pencemaran Udara (ISPU) sebagai acuan. Dalam penelitian ini, penulis mengumpulkan data terkini sesuai dengan kebutuhan ISPU. Preprocessing digunakan untuk menentukan akurasi pembacaan kualitas udara dan memberikan prediksi kualitas udara kepada masyarakat. Dengan informasi ini, masyarakat dapat mengambil tindakan pencegahan untuk mengurangi produksi pencemaran.

Kata kunci*— Polusi Udara, ISPU, Preprocessing, Akurasi, Prediksi.

I. PENDAHULUAN

Pencemaran merupakan dampak negatif dari suatu aktivitas terhadap lingkungan. Udara yang terkontaminasi oleh partikel tersebut dapat memberikan buruk bagi lingkungan, contohnya masalah pernafasan manusia, pemanasan global dan efek pada metabolisme lingkungan. Di Indonesia sendiri mendapatkan peringkat no.17 pada 2022 pada AQI (Air Quality Index) Internasional. Untuk menurunkan tingkat polusi udara ini diperlukan adanya kesadaran masyarakat tentang polusi udara.

Kota Bandung merupakan kota besar Indonesia yang posisinya berdekatan dengan ibu kota negara. Letak geografis yang strategis menjadikan kota ini cocok ditempati oleh masyarakat. Namun, hal ini juga menyebabkan banyak industri dan penduduk yang padat, yang berdampak pada tingginya tingkat polusi udara. Di kota Bandung sudah memiliki indicator pemantau kualitas udara, namun untuk parameter yang ditinjau hanya sebatas parameter umum seperti Ozon, Karbon Monoksida, dan Partikulat. Dengan hal ini dibutuhkan adanya kesadaran masyarakat dalam tingkat polusi udara dengan harapan masyarakat dapat mengurangi sedikit demi sedikit produksi polusi yang dihasilkan. Untuk parameter indicator yang digunakan berdasarkan Indeks Standar Polusi Udara (ISPU).

Pada penelitian ini penulis mengambil sampel dengan waktu terkini sesuai dengan kebutuhan pada ISPU. Untuk menemukan tingkat akurasi pembacaan ini dibutuhkan adanya pengolahan preprocessing. Dengan preprocessing dapat menigkatkan kinerja dari pembelajaran semon, sehingga dapat Hmenentukan kadar kualitas udara dan dapat memberikan prediksi kualitas udara kepada masyarakat untuk dapat berwaspada.

II. KAJIAN TEORI

Pada pengerjaan Tugas Akhir ini akan dirancang sebuah sistem monitoring polusi udara dengan menggunakan beberapa sensor, yaitu; MQ7, MQ131, MICS6814, MQ136 dan Sharp GP2Y1010AU0F. Tugas Akhir ini akan fokus membahas rancang bangun sistem monitoring polusi udara dengan sistem dengan konsentrasi pada pre-processing. dengan hasil preprocessing akan memberikan dampak yang lebih baik papda hasil pengolahan pembelajaran mesin. Penggunaan pengolahan preprocessing juga akan memberikan data yang lebih teratur, serta dengan penggunaan preprocessing memberikan hasil pengolahan metode yang akan memberikan hasil prediksi yang jelas dengan label pada setiap nilai yang di dapatkan.

A. Kualitas Udara

Kualitas udara merupakan suatu takaran keadaan udara pada atmosfer di mana manusia hidup dan beraktivitas. Contoh dari gas oksigen (O2) dan karbon dioksida (CO2) dengan konsentrasi tertentu akan sangat penting untuk perkembangan kehidupan. Rata-rata penduduk Indonesia dapat diperkirakan akan kehilangan 1,2 tahun ekspektasi harapan kehidupan dengan tingkat polusi saat ini, menurut Air Quality Life Index (AQLI) karena kualitas udara gagal memenuhi pedoman Organisasi Kesehatan Dunia (WHO) [7][8]. Ketika pemerintah menyadari masalah kualitas udara, AQLI menunjukkan bahwa Indoensia memiliki peluang untuk memperoleh keuntungan yang sangat besar untuk memperhatikan kualitas udara di Indonesia [8].

Pemerintah menerapkan suatu cara untuk membuat regulasi untuk mengatur keadaan kualitas udara yang terbaik

untuk kehidupan manusia melalui Keputusan Mentri Negara Lingkungan Hidup Nomor : KEP 45/MENLH/1997 Tentang Indeks Standar Pencemaran Udara telah memiliki standar yang disebut ISPU (Indeks Standar Pencemaran Udara) [9] . Situasi yang tidak memenuhi standar yang ditetapkan dianggap sebagai kondisi yang dapat memiliki berbagai efek yang tidak diinginkan pada banyak organisme di planet ini, termasuk manusia dan tumbuhan.

B. Indeks Standar Pencemaran Udara

Indeks Standar Pencemaran Udara merupakan suatu angka tersusun yang tidak mempunyai satuan yang menggambarkan kondisi udara ambien di lokasi dan waktu dan waktu tertentu yang didasarkan kepada dampak terhadap kesehatan manusia, nilai estetika dan mahluk hidup lainnya.[19] Secara resmi Indeks tersebut telah diatur dalam Keputusan Mentri Negara Lingkungan Hidup Nomor KEP-45/MENLH/10/1997 tentang Indeks Standar Pencemar Udara.[19] Pada Pasal 2 ayat 2 dalam peraturan tersebut menjelaskan bahwa Indeks Standar Pencemaran Udara (ISPU) ditetapkan dengan cara mengubah kadar pencemar udara yang terukur menjadi suatu angka yang tidak berdimensi. [19] Rentang Indeks Standar Pencemar Udara, dapat dilihat pada tabel.

Tabel 1 Rentang Indeks Standar Pencemaran Udara [19]

Kategori	Rentang	Keterangan
Baik	0-50	Tingkat kualitas udara yang tidak memberikan efek bagi kesehatan manusia atau hewan dan tidak berpengaruh pada tumbuhan, bangunan atau nilai estetika.
Sedang	51-100	Tingkat kualitas udara yang tidak berpengaruh pada kesehatan manusia ataupun hewan tetapi berpengaruh pada tumbuhan yang sensitif, dan nilai estetika.
Tidak sehat	101-199	Tingkat kualitas udara yang bersifat merugikan pada manusia ataupun kelompok hewan yang sensitif atau bisa menimbulkan kerusakan pada tumbuhan ataupun nilai estetika.
Sangat tidak sehat	200-299	Tingkat kualitas udara yang dapat merugikan kesehatan pada sejumlah segmen populasi yang terpapar.
Berbahaya	300-Lebih	Tingkat kualitas udara berbahaya yang secara umum dapat merugikan kesehatan yang serius.

Parameter yang digunakan sesuai ISPU untuk mengukur tingkat pencemaran udara adalah:

a. Partikulat (PM10)
b. Karbondioksida (CO)
c. Sulfur dioksida (SO2).
d. Nitrogen dioksida (NO2).
e. Ozon (O3)

Tabel 2 Batas Indeks Standar Pencemar Udara Dalam SI [19]

Indeks Standar Pencemar Udara	24 jam PM10 $\mu g/m^3$	24 jam SO_2 $\mu g/m^3$	8 jam CO $\mu g/m^3$	1 jam O_3 $\mu g/m^3$	1 jam NO_2 $\mu g/m^3$
50	50	80	5	120	80
100	150	365	10	235	365
200	350	800	17	400	1130
300	420	1600	34	800	2260
400	500	2100	46	1000	3000
500	600	2620	57.5	1200	3750

Keterangan :
(1) Pada 25 C dan 760 mmHg.
(2) Tidak ada isndeks yang dapat dilaporkan pada konsentrasi rendah dangan jangka pemaparan yang pendek.

C. *Data Preprocessing*

Data merupakan suatu nilai, hasil, angka yang dapat diolah menjadi sesuatu. Data yang dimaksud merupakan suatu nilai hasil dari pembacaan sensor yang telah dilakukan pengujian pembacaan sensor. Data preprocessing merupakan hal yang penting karena dapat mengurangi data spike karena nilai yang diberikan sensor kadangkala dapat berubah terlalu terlalu rendah sepersekian detik akibat dari pengaruh internal maupun ekternal dari komponen. Akibat dari efek spike ini flow data yang dibaca oleh skematik sensor tidak sempurna. Dari preprocessing ini menggunakan dua metode yaitu dengan cara mereduksi data dan upscaling data. Reduksi merupakan suatu cara yang digunakan untuk mengurangi nilai yang tidak diinginkan sehingga hasil yang diberikan akan terlihat lebih sempurna. Upscaling merupakan suatu cara memberikan data yang kosong atau tidak terbaca dengan memasukkan nilai random dari jangkauan pembacaan sensor.

D. *Principal Component Analysist*

Analisis komponen utama (PCA) merupakan suatu teknik analisis data multivariable yang digunakan untuk mengurangi dimensi data set. Hal ini dilakukan dengan menemukan suatu set variabel baru yang tidak berkorelasi, yang disebut dengan komponen utama, yaitu komponen yang menjelaskan mayoritas variasi dalam data asli. Komponen utama pertama (first principal component) merupakan arah dimana data paling bervariasi. Komponen utama kedua (second principal component) merupakan arah yang menjelaskan varian paling banyak yang orthogonal (tegak lurus) dengan komponen utama pertama, dan seterusnya. Komponen utama ini dapat

dianggap sebagai perangkat sumbu baru yang sejajar dengan arah varian maksimum dalam data.

PCA suatu metode pengembangan dari teori Karhunen-Loève Transform (KLT) yang merupakan transformasi linear pada pengolahan citra. Pada praktiknya, metode PCA dan KLT dapat dikatakan sebagai metode yang sama[20]. Perhitungan proses PCA dalam prosesnya menggunakan persamaan berikut:

$$x_{gen} = x + \propto . (x' - x) \quad (1)$$

Dimana:
X = matriks dimensi N x M dengan X = [x1,…,xm]T
Θ = N x M matriks orthogonal berbasis vector
Y = N x M koefisien weighting dari matriks
N = jumlah sampel data
M = jumlah fitur data [15].

E. *Standar Scaller*

Standard scaller merupakan suatu library pada sklearn preprocessing yang bertujuan sebagai normalisasi data untuk menghasilkan akurasi yang akurat. Metode ini akan mengubah setiap fitur pada data menjadi skala yang sama dengan membuat rata-rata setiap fitur menjadi 0 dan standar deviasi menjadi 1. Hal ini dilakukan dengan mengurangi setiapiliaai fitur dengan rata-rata fitur dan kemudian membagi dengan standar deviasi fitur. Metode ini berguna untuk mengatasi perbedaan skala pada fitur-fitur yang berbeda sehingga algoritma machine learning dapat bekerja lebih baik. Variabel standard scaler akan memanggil kelas StandardScaler(), fungsi scaler.fit(X_train) akan digunakan untuk melatih data pada parameter X_train.

F. *Oversampling (LoRa)*

Oversampling adalah teknik yang digunakan dalam pembelajaran mesin untuk menangani dataset yang tidak seimbang. Dalam dataset yang tidak seimbang, kelas minoritas (kelas dengan contoh yang lebih sedikit) sering kali lebih banyak daripada daripada kelas mayoritas. Hal ini dapat menyebabkan bias pada model, karena model akan dilatih untuk memprediksi kelas mayoritas dengan lebih akurat. Oversampling mengatasi masalah ini dengan meningkatkan jumlah contoh kelas minoritas dalam dataset. Hal ini dapat dilakukan dengan menduplikasi contoh kelas minoritas yang sudah ada, atau dengan membuat contoh sintetis dari kelas minoritas menggunakan teknik seperti SMOTE (Synthetic Minority Over-sampling Technique).

Tujuan dari oversampling adalah untuk menyeimbangkan dataset, sehingga model dilatih pada distribusi kelas yang lebih merata. Hal ini dapat meningkatkan performa pada kelas minoritas, karena model akan terpapar pada lebih banyak contoh kelas tersebut selama pelatihan. Namun, oversampling juga dapat menyebabkan overfitting, karena model dapat menjadi terlalu terbiasa dengan contoh kelas minoritas yang terlalu banyak diambil, dan berkinerja buruk pada contoh-contoh baru yang belum pernah dilihat..

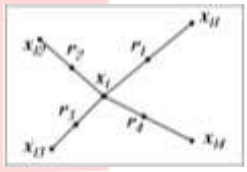
G. *SMOTE*

Synthetic Minority Oversampling Technieque (SMOTE) merupakan suatu algoritma pada preprocessing yang biasa dilakukan pada suatu data yang tidak seimbang. Hal ini disebabkan oleh kesederhanaan dalam desain prosedur dan ketangguhan ketika diterapkan pada berbagai jenis masalah. Smote telah terbukti berhasil dalam berbagai aplikasi dari beberapa domain yang berbeda. SMOTE juga telah menginspirasi beberapa pendekatan untuk mengatasi masalah ketidakseimbangan kelas, dan juga secara signifikan berkontribusi pada paradigma pembelajaran baru yang diawasi, termasuk klasifikasi multilabel, pembelajaran in-cremental, pembelajaran semi-supervised, pembelajaran multi-instance, dan lain-lain.

Perhitungan dilakukan dengan rumus berikut[21]:

(2)

xgen = data yang dibuat
x = data pada kelas minoritas
x1 = data nearest neightbour dari x
α = nilai random antara 0 hingga 1.



Gambar 1 Pembuatan Data Sintesis SMOTE

Nilai k untuk menentukan nearest neighbour harus ditentukan karena memiliki pengaruh yang besar terhadap data sintesis. Semakin besar nilai k, maka kemungkinan membuat data sintesis yang noisy akan semakin tinggi, namun di sisi lain, nilai k yang semakin kecil akan menyebabkan data sintesis hanya berkumpul pada beberapa titik tertentu.

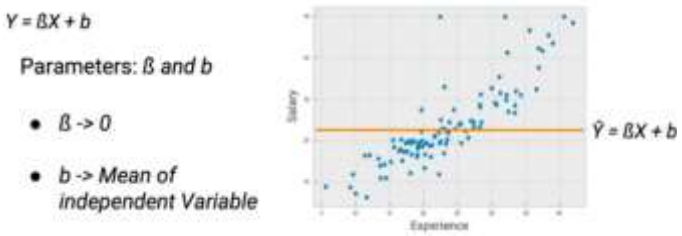
H. *Supervised Vector Machine (SVM)*

Support Vector Machine (SVM) adalah algoritma pembelajaran terawasi yang dapat digunakan untuk tugas klasifikasi dan regresi. Ide dasar di balik SVM adalah menemukan batas keputusan, atau hyperplane, yang memisahkan kelas-kelas yang berbeda dalam kumpulan data.

Konsep SVM dapat disederhanakan dengan tujuan utama yaitu mencari hyperlane terbaik yang berfungsi sebagai pemisah dua buah class. Hyperlane pemisah terbaik dapat ditemukan dengan mengukur margin hyperlane dan mencari titik maksimalnya. Margin ini merupakan jarak antar hyperlane dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai support vector.

I. *Gradient Descent*

Gradient descent merupakan suatu metode algoritma optimasi yang digunakan untuk meminimalkan fungsi kerugian dengan memperbarui parameter model secara berulang-ulang ke arah penurunan paling curam dari fungsi kerugian. Algoritma ini dimulai dengan satu set parameter awal dan menghitung gradien fungsi kerugian sehubungan dengan parameter tersebut. Gradien tersebut kemudian digunakan untuk memperbarui parameter dengan cara mengurangi nilai fungsi kerugian. Proses ini diulangi sampai nilai minimum dari fungsi kerugian tercapai atau kriteria penghentian terpenuhi.



Gambar 2 Rumus Gradient Descent pada Regresi Linear[23]

J. *Root Mean Squares Error*

Kesalahan kuadrat rata-rata adalah perbedaan kuadrat rata-rata antara nilai estimasi dan nilai aktual. uatu model prakiraan memiliki kualitas yang baik jika memiliki nilai RMSE yang rendah, yang menandakan bahwa perbedaan antara nilai yang diprediksi dan nilai yang sebenarnya sangat dekat. RMSE mengukur tingkat ketidakcocokan antara dua set nilai, dan semakin kecil nilai RMSE, semakin baik kualitas model prakiraan tersebut.

$$\mathrm{RMSE} = \left(\frac{\Sigma(y_i - \hat{y}_i)}{n}\right)^{1/2}$$

Keterangan:
RMSE = nilai root mean square error
y = nilai hasil observasi
ŷ = nilai hasil prediksi
i = urutan data pada database
n = jumlah data

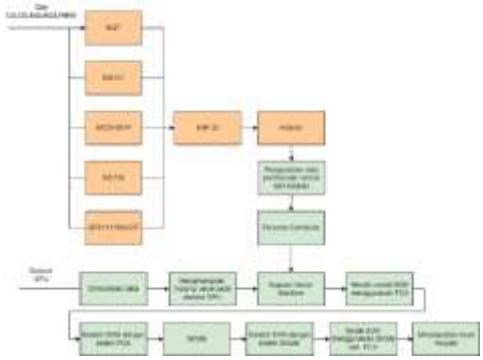
III. METODE

A. Desain Sistem

Penelitian tugas akhir akan dirancang untuk memonitoring polusi udara secara multi titik. Monitoring dilakukan dengan menggunakan sistem mikrokontroler berbasis ESP32 yang di program menggunakan perangkat lunak Arduino IDE. Sensor nirkabel akan terhubung dimana sensor mengumpulkan data dan data di akusisi menjadi data yang siap dikirim kepada jaringan. sensor memiliki sensor sebagai berikut; sensor GP2Y1010AU0F, MQ-136, MQ-131, MiCS 6814 dan MQ-7.

B. Diagram Blok

Sistematika kajian tugas akhir ini merupakan gabungan dari tiga tugas akhir. Tiga di antaranya adalah rancang bangun monitoring polusi udara dengan sistem IOT, klasifikasi data pengolahan polusi udara dan data preprocessing berdasarkan data yang diperoleh dari sensor. Tugas akhir ini akan fokus terhadap proses pra-data sensor untuk menemukan akurasi yang tepat. Hal ini dapat dilihat pada bagian hijau dari diagram blok pada Gambar 3.

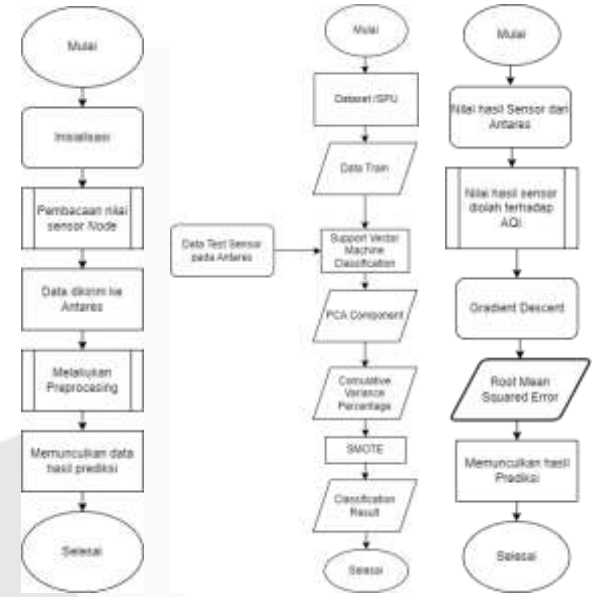


Gambar 3 Diagram Blok Monitoring Polusi Udara

C. Desain Perangkat Lunak

Perangkat lunak pada penelitian ini menggunakan beberapa aplikasi seperti Jupiter Notebook, Arduino IDE. Data sensor yang sudah di terbaca akan dikirimkan ke perangkat lunak Arduino IDE. Dalam perangkat lunak tersebut akan dimunculkan data hasil pembacaan sensor yang akan di eksport ke IoT. Hasil pada pembacaan sensor akan dibutuhkan kembali pada proses preprocessing dimana data tersebut akan di unduh. Setelah data hasil pembacaan sensor tersebut di dapatkan, maka selanjutnya data tersebut akan dilatih pada perangkat lunak Jupiter Notebook.

D. Diagram Alir



Gambar 4 Diagram Alir

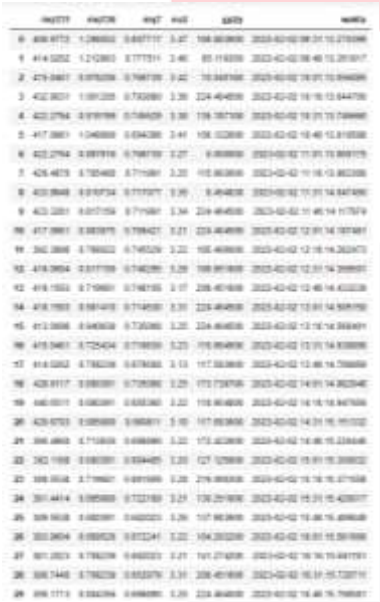
Gambar 5 merupakan diagram alir, dimulai dengan inisiasi seluruh port serial dari sensor yang digunakan. Sensor membaca nilai kualitas udara dan mengambil data yang dikirimkan ke IoT. Setelah data didapatkan dari IoT, maka dilakukan preprocessing untuk mendapatkan nilai prediksi. Data latih ISPU tersebut digunakan untuk mengolah data hasil pembacaan sensor yang menjadi data test pada proses ini. Hasil pembacaan data sensor akan diolah dengan metode Support Vector Machine bertujuan untuk klasifikasi yang akan mengeluarkan hasil akurasi dari metode ini. Dan selanjutnya data test pembacaan sensor akan diproses dengan metode Principal Component Analysis dan mengeluarkan Cumulative Variance Persentage. Selanjutnya dilakukan metode SVM yang dipadukan dengan SMOTE. Metode ini digunakan untuk melihat apakah hasil proses PCA tersebut

akan naik atau tidaknya. Data hasil pembacaan sensor didapatkan langsung dari IoT. Sistem ini terhubung langsung dengan menggunakan subscribe data langsung dari IoT. Data test tersebut diolah dengan standard AQI. Setelah mendapatkan hasil AQI, selanjutnya akan diproses dengan metode Gradient Descent untuk mendapatkan nilai optimasi dari pembacaan sensor. Setelah mendapatkan nilai optimalnya, sistem akan masuk ke proses RMSE (Root Mean Squared Error), yaitu suatu cara untuk mengevaluasi model regresi linear dengan mengukur tingkat akurasi hasil perkiraan nilai sensor.

IV. HASIL DAN PEMBAHASAN

A. Subscribe Software IoT

Data yang telah di proses pada pengujian analisis komponen sensor akan dikirimkan melalui perangkat lunak IoT. Pada Software IoT akan terlihat semua hasil pembacaan sensor dengan beberapa parameter diantaranya; mq131(O_3), mq136(SO_2), mq7(CO), mics6814(NO_2), dan sharp gp2y(PM_{10}).



Gambar 4 Data dari IoT

Pada tahap ini data hasil pengujian sensor akan dimasukkan untuk dapat diproses pada perbandingan dari hasil pengujian sensor dan juga hasil data train. Nilai pembacaan sensor ini didapatkan secara realtime dengan perangkat keras sistem monitoring. Dengan interval waktu pengujian selama 8 jam dengan interval pengambilan sampel setiap 15 menit.

B. Analisis Metode Support Vector Machine (SVM)

Tahap ini dilakukan pengujian hasil nilai sensor dengan metode Support Vector Machine. Tujuan dari penggunaan SVM ini yaitu sebagai cara untuk melihat apakah sistem Principal Componen Analysis ini dapat digunakan untuk menemukan akurasi yang lebih baik. SVM ini menggunakan parameter untuk menguji model pembelajaran mesin dengan parameter Gamma, Cost, dan Kernel RBF. Penggunaan data train yang digunakan sebanyak 200 data, data test yang digunakan sebanyak 30 data dengan atribut yang digunakan ada 5, yaitu PM10, Karbon Monoksida, Sulfur Dioksida, Ozon, Nitrogen Dioksida. Penggunaan 5 atribut tersebut menggunakan 3 kategori yaitu baik, tidak baik, dan sangat tidak baik.

```
from sklearn.svm import SVC
clf = SVC(kernel = 'rbf',gamma=1, C=1)
clf.fit(X_train, y_train)
```

```
SVC(C=1, gamma=1)
```

Gambar 5 Penggunaan Parameter pada SVM

Pada data hasil akurasi pembacaan sensor didapatkan nilai akurasi sebesar 0,7 atau 70% dari keseluruhan data. Untuk hasil akurasi tersebut dikategorikan baik untuk penggunaan metode pembelajaran mesin.

```
clf.score(X_test,y_test)
```

```
0.7
```

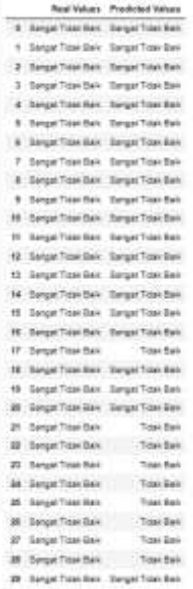
Gambar 6 Hasil Akurasi SVM

```
# using MAPE error metrics to check for the error rate and accuracy level
SVR_MAPE= MAPE(y_test,y_pred)
print("MAPE: ",SVR_MAPE)
```

```
MAPE:  50.0
```

Gambar 7 Rata-rata Persentase Kesalahan Absolut

Gambar 7 merupakan MAPE (Mean Absolut Percentage Error) yaitu rata-rata persentase kesalahan absolut yang digunakan untuk mengukur galat ramalan dibandingkan dengan nilai actual. Penggunaan ini untuk melihat prediksi akurasi semakin tinggi nilai MAPE maka akan semakin kecil nilai akurasinya. Karena MAPE yang di hasilkan memiliki nilai 30% maka prediksi dikategorikan layak.



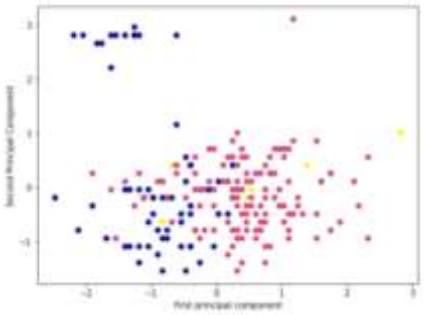
Gambar 8 Hasil Peramalan Metode Pembelajaran Mesin SVM

Hasil dari penggunaan metode SVM dapat dilakukan pemberian label pada setiap hasil pengolahannya, dengan fitur ini dapat memberikan hasil pembacaan yang terbaik yang digunakan pada tahap preprocessing.

C. Hasil Analisis menggunakan metode *Support Vector Machine* dengan *Principal Component Analysis*

Sistem menggunakan metode Principal Component Analysis untuk melihat apakah ada pengaruh hasil akurasi data dengan hanya menggunakan metode SVM saja. Proses ini dilakukan

dengan menggunakan beberapa Principal Component digabungkan dengan model SVM.



Gambar 9 Color Map Data ISPU

```
Cumulative Variances (Percentage):
[ 51.62838829  70.63835891  86.19103274  94.42078254 100.        ]
Number of components: 5
```

Gambar 10 Cumulative Variance

Pada gambar 9 merupakan cumulative variance yaitu merupakan persentase varian data yang digunakan dalam PCA. Pada penelitian ini menggunakan persentase varian 100%. Komponen PCA yang digunakan berjumlah 5. Hal ini digunakan untuk memperoleh nilai akurasi yang tinggi.

```
SVC_model.score(X_test_pca,y_test)
0.6666666666666666
```

Gambar 11 Hasil Persentase SVM dengan PCA

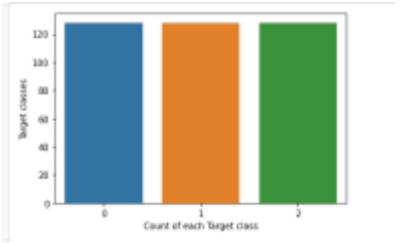
Terlihat pada gambar 11 pada penelitian ini menghasilkan akurasi yaitu 0.6 atau 66% dari keseluruhan sistem dengan persentase di 5 komponen. Untuk hasil ini dikategorikan cukup baik untuk penggunaan metode pembelajaran mesin.

Gambar 12 Hasil Peramalan Metode Pembelajaran Mesin SVM dengan PCA

Pembacaan dari prediksi hasil metode SVM dan PCA dapat dilihat pada gambar 12, untuk hasil prediksi cukup tidak berubah dari metode SVM dikarenakan metode hasil akurasi yang berbeda sedikit. Sehingga untuk keluaran prediksi dari metode SVM dan SVM dengan PCA tidak adanya perubahan.

D. Hasil Analisis Menggunakan SMOTE

Data yang diterima saat proses sebelumnya masih berupa data imbalance. Sehingga dibutuhkannya adanya proses SMOTE. Data imbalance ini memiliki nilai target kelas tertinggi di nilai 120 pada kelas "1" atau sedang, dapat dilihat pada gambar 13 berikut.



Gambar 13 Data Imbalance Sistem Polusi Udara

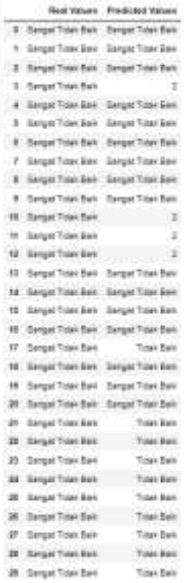
Metode SMOTE menggunakan data maksimum sebagai target yang akan diterapkan pada setiap variabel minoritas. Hasil yang sudah di proses pada SMOTE ini menghasilkan data yang balance, sehingga variabel satu sama lain berjumlah sama. Pengaruh imbalance data akurasi sebelum dan setelah SMOTE menyebabkan kenaikan pada persentase. Hal ini diakibatkan oleh semakin banyak jumlah data yang diolah semakin akurat nilai akurasi. Naik turunnya presentase tergantung dari sistem random data yang di proses.

Hasil yang sudah di proses pada SMOTE ini menghasilkan data yang balance, sehingga variabel satu sama lain berjumlah sama.

```
clf.score(X_test,y_test)
0.5666666666666667
```

Gambar 14 Hasil Persentase SVM dan SMOTE

Terlihat pada gambar 14 merupakan hasil proses model SVM digabungkan dengan SMOTE pada sensor. Hasil akurasi yang di dapatkan yaitu 0.56 atau 56%. Nilai yang didapatkan ketika semua data yang tidak balance tersebut menurun dari metode pembelajaran mesin sebelumnya. Hasil ini dapat dikategorikan buruk untuk penggunaan pada metode model pembelajaran mesin.



Gambar 15 Hasil Peramalan Metode Pembelajaran Mesin SVM dengan SMOTE

Pada hasil pembacaan metode SMOTE, hasil prediksi yang dikeluarkan berbeda dengan kedua metode sebelumnya, hal ini dapat dipastikan bahwa hasil dari akurasi yang cukup jauh dari metode SVM dan SVM dengan PCA. Sehingga hasil dari prediksi yang dihasilkan cukup jauh dari metode lainnya.

E. Perbandingan Pengujian Metode SVM, SVM dengan PCA dan SVM dengan SMOTE

Bagian ini akan melakukan perbandingan performa metode Support Vector Machine, Support Vector Machine dengan Principal Component Analysis, Support Vector Machine dengan SMOTE.

Tabel 3 Tabel Perbandingan Akurasi Setiap Metode

Metode *Machine Learning*	Hasil Akurasi Penggunaan Metode *Machine Learning*
Support Vector Machine	70%
Support Vector Machine dengan Principal Componen Analysis	66,6%
Support Vector Machine dengan SMOTE	56%

Pada metode Support Vector Machine memiliki tingkat akurasi senilai 70%, metode Support Vector Machine dan Principal Component Analysis memiliki tingkat akurasi senilai 66,6%, dan metode Support Vector Machine dengan SMOTE memiliki tingkat akurasi senilai 56%. Dari hasil penelitian ini dapat disimpulkan bahwa dengan metode Support Vector Machine merupakan hasil model pembelajaran mesin yang sangat baik.

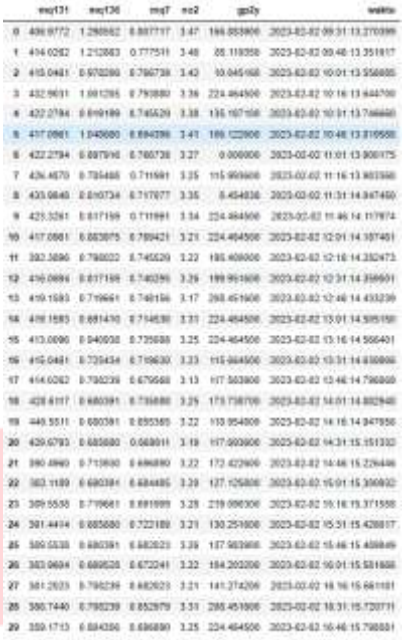
Jika penggunaan Support Vector Machine tersebut digunakan dengan metode Principal Component Analyst maka hasil akurasi yang diperoleh lebih rendah karena parameter yang digunakan akan semakin banyak. Sehingga hasil dari pengolahan akurasi akan menurun. Untuk penggunaan Support Vector Machine dengan SMOTE bisa menyebabkan overfitting dan hasil yang tidak optimal jika tidak digunakan dengan benar.

SMOTE (Synthetic Minority Over-sampling Technique) digunakan untuk membuat sintesis dari data minoritas (kelas minoritas) dengan mengambil sample dari data yang ada dan membuat sample baru dengan interpolasi antara sample. Sedangkan PCA (Principal Component Analysis) digunakan untuk melakukan dimensi reduksi pada data. Namun, jika data sudah terreduksi dengan baik, pemakaian PCA tidak akan memberikan perubahan signifikan pada hasil SVM.

Namun, jika menggunakan metode ketiga metode tersebut secara bersamaan tingkat akurasi yang didapatkan akan menurun, ini dipengaruhi oleh beberapa kasus:

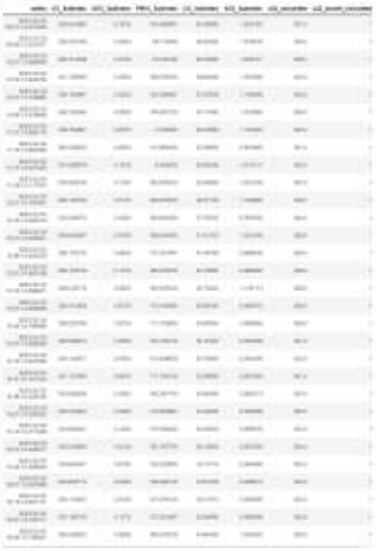
1) *Principal Component Analysis* dapat menghilangkan beberapa fitur yang penting dari data yang digunakan oleh SVM untuk membuat keputusan klasifikasi.
2) SMOTE dapat menambahkan data yang tidak sebenarnya dari kelas minoritas, yang dapat menyebabkan overfitting dan menurunkan tingkat akurasi.

F. Hasil Pengujian Prediksi Sensor



Gambar 16 Data Test Pembacaan Sensor

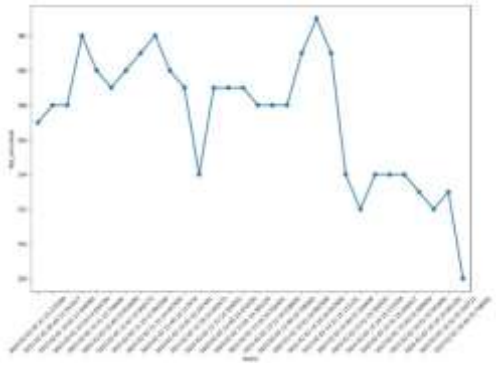
Pengujian sensor pada sensor dilakukan selama 8 jam dengan interval waktu setiap pengambilan sensor yaitu 15 menit. Dan target pengambilan data yaitu 30 data pada setiap sensor.



Gambar 17 Data Pengujian Sensor dengan AQI

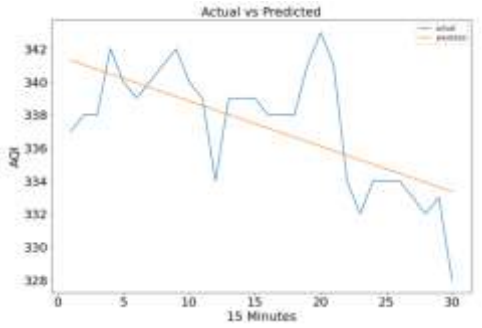
Pada pengujian prediksi dibutuhkan parameter yang telah diolah sebelumnya, yaitu O3 Subindeks, NO2 Subindeks, PM10 Subindeks, CO Subindeks, SO2 Subindeks, dan AQI calculation. Untuk mendapatkan parameter yang memenuhi kriteria penulis, maka penulis memberikan label subindeks pada setiap masing masing parameter. Kondisi pada AQI Bucketing digunakan untuk memberikan kondisi pada setiap parameter sensor. Hal ini digunakan untuk mencapai hasil klasifikasi yang dibutuhkan pada proses pre-processing. Dengan hasil kondisi pada AQI Bucketing ini memberikan peluang penulis pada proses selanjutnya, yaitu melakukan klasifikasi kepada setiap variabel. Pada AQI Bucket Calculate diambil berdasarkan sensor yang mempunyai nilai

maksimal. Pada data ini variabel yang bernilai maksimal adalah Ozon.



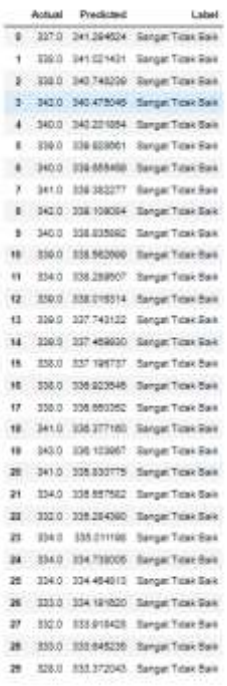
Gambar 18 Line Graph Data Percobaan Sensor

Dari hasil percobaan pada sensor, semakin lama udara pada pengujian sensor semakin baik karena kandungan kadar gas Ozon di udara semakin sedikit. Hal ini disebabkan oleh gas Ozon di udara pada pengujian sampel tergolong tidak sehat karena memiliki kandungan Ozon yang banyak. Lalu, setelah dilakukan percobaan selama 8 jam, ternyata kandungan gas ozon di udara semakin sedikit. Oleh karena itu, kualitas udara pada saat percobaan lebih baik dari data sebelumnya, tetapi kualitas udara masih tergolong sangat tidak sehat.



Gambar 19 Grafik Prediksi Sensor

Setelah dilakukan percobaan sensor menggunakan pembelajaran mesin. didapatkan hasil prediksi seperti gambar 19. Dapat dilihat terdapat garis biru pada grafik menandakan data pengujian langsung dan garis kuning merupakan data hasil prediksi. Pengambilan sampel setiap 15 menit di ruang terbuka, dengan rentang waktu 9 pagi hingga 5 sore. Didapatkan data pengujian sekitar 30 hasil pengujian. Hasil pembacaan sensor memiliki data yang cenderung menurun, ini dipengaruhi oleh semakin sedikitnya kadar gas yang menjadi pengujian sensor. Hasil prediksi ini menunjukkan kualitas udara yang semakin baik, begitu juga dengan data pengujian aslinya.



Gambar 20 Hasil Prediksi pada Pembelajaran Mesin

Dari pengolahan data pembacaan sensor pada dapat memberikan hasil prediksi yang sangat mendekati nilai aktual. Hal ini memberikan proses hasil prediksi yang sangat sempurna pada sistem monitoring.

Gambar 21 Hasil Prediksi pada IoT

Hasil pengolahan pada pembelajaran mesin berhasil di unggah ke IoT dengan data yang dapat dilihat seperti pada gambar 21. Hasil Prediksi ditampilkan dengan komponen variabel ‘aktual’ dan ‘predicted’ sehingga dapat dilihat bahwa perbandingan nilai dari kedua komponen tersebut tidak terlalu jauh. Dan fitur selanjutnya terdapat hasil prediksi dari pembacaan sensor tersebut yaitu pada variabel “label” yang dapat memberikan informasi kepada masyarakat bahwa prediksi polusi udara sekitar pengujian agar dapat mengantisipasi polutan tersebut.

V. KESIMPULAN

1) Penelitian ini menggunakan 5 parameter yang memenuhi kebutuhan data ISPU yaitu MQ131(O3), MQ136(SO2), MQ7(CO), MICS 6814(NO2), dan sharp GP2Y(PM10).

2) Dari hasil penelitian ini dapat disimpulkan bahwa dengan metode Support Vector Machine merupakan hasil model pembelajaran mesin yang sangat baik pada proses pembelajaran mesin dengan hasil akurasi 70% lebih baik dibandingkan dengan menggunakan metode Principal

Component Analysis dan SMOTE yang dapat mereduksi dan memberikan variabel baru segingga.

3) Jika penggunaan Support Vector Machine digunakan dengan metode Principal Component Analyst maka hasil akurasi yang diperoleh lebih rendah karena parameter yang digunakan akan semakin banyak. Sehingga hasil dari pengolahan akurasi akan menurun. Untuk penggunaan Support Vector Machine dengan SMOTE bisa menyebabkan overfitting dan hasil yang tidak optimal jika tidak digunakan dengan benar.

REFERENSI

[1]"Repository - Monitoring Polusi Udara Menggunakan Mikrokontroler Hemat Daya Multi Titik Dengan Jaringan Mesh Nirkabel." https://repository.telkomuniversity.ac.id/pustaka/179304/monitoring-polusi-udara-menggunakan-mikrokontroler-hemat-daya-multi-titik-dengan-jaringan-mesh-nirkabel.html (accessed Jun. 26, 2022).

[2] A. A. Dharmasaputro, I. Prasetya, D. Wibawa, and M. Kallista, "Lembar Pengesahan Proposal Tugas Akhir Monitoring Dan Klasifikasi Polusi Udara Menggunakan Random Forest Dengan Imbalanced Dataset (Air Pollution Monitoring and Classification Using Random Forest With Imbalanced Dataset)," pp. 2–9.

[3] A. Kurniawan, "Pengukuran Parameter Kualitas Udara (Co, No2, So2, O3 Dan Pm10) Di Bukit Kototabang Berbasis Ispu," *J. Teknosains*, vol. 7, no. 1, p. 1, 2018, doi: 10.22146/teknosains.34658.

[4] S. Kaivonen and E. C. H. Ngai, "Real-time air pollution monitoring with sensors on city bus," *Digit. Commun. Networks*, vol. 6, no. 1, pp. 23–30, Feb. 2020, doi: 10.1016/J.DCAN.2019.03.003.

[5] P. M. Mannucci and M. Franchini, "Health effects of ambient air pollution in developing countries," *Int. J. Environ. Res. Public Health*, vol. 14, no. 9, pp. 1–8, 2017, doi: 10.3390/ijerph14091048.

[6] A. Famili, W. M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intell. Data Anal.*, vol. 1, no. 1, pp. 3–23, 1997, doi: 10.3233/IDA-1997-1102.

[7] V. Roza, M. Ilza, and S. Anita, "Korelasi Konsentrasi Particulate Matter (PM10) di Udara dan Kandungan Timbal (Pb) dalam Rambut Petugas SPBU di Kota Pekanbaru," *Din. Lingkung. Indones.*, vol. 2, no. 1, p. 52, 2015, doi: 10.31258/dli.2.1.p.52-60.

[8] B. Michael Greenstone and Q. Fan, "Indonesia's Worsening Air Quality and its Impact on Life Expectancy".

[9] M. Negara Lingkungan Hidup, "Keputusan Menteri Negara Lingkungan Hidup No. 45 Tahun 1997 Tentang : Indeks Standar Pencemar Udara".

[10] M. Rashid, S. Yunus, R. Mat, S. Baharun, and P. Lestari, "PM10 black carbon and ionic species concentration of urban atmosphere in Makassar of South Sulawesi province, Indonesia," *Atmos. Pollut. Res.*, vol. 5, no. 4, pp. 610–615, Oct. 2014, doi: 10.5094/APR.2014.070.

[11] A. Aslam *et al.*, "Mitigation of particulate matters and integrated approach for carbon monoxide remediation in an urban environment," *J. Environ. Chem. Eng.*, vol. 9, no. 4, p. 105546, 2021, doi: 10.1016/j.jece.2021.105546.

[12] Y. Niaz, J. Zhou, A. Nasir, M. Iqbal, and B. Dong, "Comparitive study of particulate matter (PM10 and PM2.5) in Dalian-China and Faisalabad-Pakistan," *Pakistan J. Agric. Sci.*, vol. 53, no. 1, pp. 97–106, 2016, doi: 10.21162/PAKJAS/16.3623.

[13] P. Graphics Inc, "Ozone Therapy in Medicine and Dentistry," 2008.

[14] W. E. Cahyono, "Pengaruh Penipisan Ozon Terhadap Kesehatan Manusia," *Semnas Penelitian, Pendidik. dan Penerapan MIPA*, pp. 208–214, 2005.

[15] E. J. Emmett and M. C. Willis, "The Development and Application of Sulfur Dioxide Surrogates in Synthetic Organic Chemistry," *Asian J. Org. Chem.*, vol. 4, no. 7, pp. 602–611, 2015, doi: 10.1002/ajoc.201500103.

[16] T. M. Chen, J. Gokhale, S. Shofer, and W. G. Kuschner, "Outdoor air pollution: Nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects," *Am. J. Med. Sci.*, vol. 333, no. 4, pp. 249–256, 2007, doi: 10.1097/MAJ.0b013e31803b900f.

[17] M. M. Ballari, Q. L. Yu, and H. J. H. Brouwers, "Experimental study of the NO and NO 2 degradation by photocatalytically active concrete," vol. 161, no. 2, pp. 175–180, 2011, doi: 10.1016/j.cattod.2010.09.028.

[18] X. Congresso, N. Igiic, L. Stato, P. Garzolini, and T. Wassermann, "XVIII Congresso Nazionale IGIIC – Lo Stato dell'Arte 18 –Palazzo Garzolini di Toppo Wassermann, Udine, 29-31 ottobre 2020," pp. 29–31, 2020.

[19] "_PT. ERM INDONESIA".

[20] D. A. Ihsani, A. Arifin, and M. H. Fatoni, "Klasifikasi DNA Microarray Menggunakan Principal Component Analysis (PCA) dan Artificial Neural Network (ANN)," *J. Tek. ITS*, vol. 9, no. 1, 2020, doi: 10.12962/j23373539.v9i1.51637.

[21] A. Walford, "Raw Data:," *Mult. Nature-Cultures, Divers. Anthropol.*, pp. 65–80, 2020, doi: 10.2307/j.ctv1850gsb.8.

[22] D. A. Pisner and D. M. Schnyer, "Support vector machine," *Mach. Learn. Methods Appl. to Brain Disord.*, pp. 101–121, 2019, doi: 10.1016/B978-0-12-815739-8.00006-7.

[23] E. Bendersky, "Understanding gradient descent," 2016, [Online]. Available: https://eli.thegreenplace.net/2016/understanding-gradient-descent/

[24] C. Optical and D. Sensor, "GP2Y1010AU0F," pp. 1–9, 2006.

[25] ETC2, "MQ136 Semiconductor Sensor for Sulfur Dioxide," pp. 2–4.