Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

TUGAS AKHIR

Hady Satria
41518110117

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
Hady Satria
41518110117

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM : 41518110117

Nama : Hady Satria

Judul Tugas Akhir : Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 02 Agustus 2022

Hady Satria

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Hady Satria
NIM	:	41518110117
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 02 Agustus 2022

Hady Satria

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Hady Satria
NIM	:	41518110117
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

<table>
<tr><th>No</th><th>Luaran</th><th colspan="3">Jenis</th><th colspan="2">Status</th></tr>
<tr><td rowspan="6">1</td><td rowspan="4">Publikasi Ilmiah</td><td colspan="2">Jurnal Nasional Tidak Terakreditasi</td><td></td><td rowspan="2">Diajukan</td><td rowspan="2">✓</td></tr>
<tr><td colspan="2">Jurnal Nasional Terakreditasi</td><td>✓</td></tr>
<tr><td colspan="2">Jurnal International Tidak Bereputasi</td><td></td><td rowspan="2">Diterima</td><td rowspan="2"></td></tr>
<tr><td colspan="2">Jurnal International Bereputasi</td><td></td></tr>
<tr><td rowspan="4">Disubmit/dipublikasikan di :</td><td>Nama Jurnal</td><td colspan="3">: Jurnal Rekayasa Sistem dan Teknologi Informasi</td></tr>
<tr><td>ISSN</td><td colspan="3">: 2580-0760</td></tr>
<tr><td>Link Jurnal</td><td colspan="3">: http://jurnal.iaii.or.id/index.php/RESTI</td></tr>
<tr><td>Link File Jurnal Jika Sudah di Publish</td><td colspan="3">:</td></tr>
</table>

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 02 Agustus 2022

Hady Satria

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518110117
Nama	:	Hady Satria
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 02 Agustus 2022

(Muhammad Rifqi, S.Kom, M.Kom)

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518110117
Nama	:	Hady Satria
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 19 Agustus 2022

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518110117
Nama	:	Hady Satria
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 02 Agustus 2022

(Wawan Gunawan, S.Kom, MT)

LEMBAR PENGESAHAN

NIM	:	41518110117
Nama	:	Hady Satria
Judul Tugas Akhir	:	Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 02 Agustus 2022

Menyetujui,

(Eliyani, Dr. Ir)
Dosen Pembimbing

Mengetahui,

(Wawan Gunawan, S.Kom, MT)	(Ir. Emil R. Kaburuan, Ph.D., IPM.)
Koord. Tugas Akhir Teknik Informatika	Ka. Prodi Teknik Informatika

KATA PENGANTAR

Puji syukur kita panjatkan kehadirat Allah SWT, atas limpahan Rahmat dan Karunia-Nya, Sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul "Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression". Penulis membuat Tugas Akhir ini karena merupakan salah satu syarat yang harus dipenuhi oleh mahasiswa tingkat akhir untuk menyelesaikan studi dan untuk mendapatkan gelar Sarjana Ilmu Komputer pada Program Studi Teknik Informatika Universitas Mercu Buana.

Penulis menyadari bahwa tanpa bantuan dan bimbingan Dosen Pembimbing dan Berbagai Pihak, Tugas Akhir ini tidak dapat terselesaikan hingga saat ini dengan baik. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Kedua orang tua penulis, yang selalu memberikan doa dan dukungan penuh sehingga penulis dapat menyelesaikan Tugas Akhir ini dengan semangat, baik dan lancar.
2. Ibu Eliyani, Dr. Ir. selaku Dosen Pembimbing Tugas Akhir yang selaku memberikan bimbingan dan arahan dalam penulisan dan penyusunan Tugas Akhir ini.
3. Bapak Muhammad Rifqi, S.Kom, M.Kom selaku dosen pembimbing akademik yang selalu memberikan bimbingan akademik sehingga penulis dapat menyelesaikan studinya tepat waktu.
4. Bapak Ir. Emil R. Kaburuan, Ph.D., IPM. Selaku Kepala Program Studi Teknik Informatika
5. Bapak Wawan Gunawan, S.Kom, MT. selaku Koordinator Tugas Akhir Teknik Informatika
6. Seluruh Jajaran Dosen dan Staf Program Studi Teknik Informatika Universitas Mercu Buana
7. Kakak dan adik penulis yang selalu membantu memberikan dukungan moril maupun materil.
8. Kepada teman dan semua pihak yang terlibat dan tidak dapat disebutkan satu persatu, semoga Allah SWT selalu melindungi dan membalas kebaikan yang lebih besar.

9. Last but not least, I wanna thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having no days off.

Akhir kata, penulis berharap penulis menyadari bahwa Tugas Akhir ini masih jauh dari kata sempurna dikarenakan terbatasnya pengetahuan dan pengalaman yang dimiliki. Semoga melalui Tugas Akhir ini dapat memberikan manfaat dan menambah pengetahuan daripada penulis dan pembaca yang budiman.

Jakarta, 13 Juli 2022
Penulis

DAFTAR ISI

HALAMAN SAMPUL ..... i.
HALAMAN JUDUL ..... **ii**
LEMBAR PERNYATAAN ORISINALITAS ..... **iii**
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR ... **iv**
SURAT PERNYATAAN LUARAN TUGAS AKHIR ..... **v**
LEMBAR PERSETUJUAN PENGUJI ..... **vi**
LEMBAR PENGESAHAN ..... **ix**
ABSTRAK ..... **x**
ABSTRACT ..... **xi**
KATA PENGANTAR ..... **xii**
DAFTAR ISI ..... **xiv**
NASKAH JURNAL **1**
KERTAS KERJA **11**
BAB 1. LITERATUR REVIEW **16**
BAB 2. ANALISIS DAN PERANCANGAN **26**
BAB 3. SOURCE CODE **33**
BAB 4. DATASET **43**
BAB 5. TAHAPAN EKSPERIMEN **46**
BAB 6. HASIL SEMUA EKSPERIMEN **56**
DAFTAR PUSTAKA **68**
LAMPIRAN DOKUMEN HAKI **73**
LAMPIRAN KORESPONDENSI **75**

NASKAH JURNAL

Terakreditasi SINTA Peringkat 2
Surat Keputusan Direktur Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Nomor: 158/E/KPT/2021
masa berlaku mulai Volume 5 Nomor 2 Tahun 2021 sampai Volume 10 Nomor 1 Tahun 2026

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. x (2022) x - x ISSN Media Elektronik: 2580-0760

Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression

Streamlit Implementation to Classify Air Pollution Levels in DKI Jakarta Using Regression Algorithms

Hady Satria[1], Eliyani[2]
[1]Teknik Informatika, Fakultas Ilmu Komputer, Universitas MercuBuana
[2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas MercuBuana
[1]41518110117@student.mercubuana.ac.id, [2]eliyani@mercubuana.ac.id*

Abstract

The issue of air pollution is quite an important thing that relates to the growth of life in urban areas, this issue is commonly known as air pollution. Air conditions in Jakarta show that during the 2017-2020 period, the air is generally categorized as moderate to unhealthy. This is measured based on the Air Pollution Standard Index (ISPU) which classifies the air quality in an area based on the parameters of chemical substances and their effects on health. air quality classification ranging from good, moderate, unhealthy, very unhealthy, and dangerous categories, can be identified by using the regression type algorithm, namely Logistic Regression and Random Forest. This study intends to rely on the type of regression algorithm by designing a dashboard to classify the level of air pollution in DKI Jakarta using the Steamlit library. In addition, it functions as a validation of the algorithm that has been previously trained regarding the evaluation level and its accuracy. In the problem of classifying the air pollutant standard index in DKI Jakarta using a regression algorithm. The results of the modeling evaluation, for the appropriate value based on the proportion of split data 80%: 20% and 60%: 40%, show that the Random Forest algorithm is superior to the accuracy of 96.66% & 96.22%. With the smallest MEA, MSE, RMSE values, respectively, 0.03866, 0.04920, 0.2218. for split data 80% : 20%, then the value is 0.045, 0.0615, 0.2480, on split data 60% : 40%. For the evaluation value of r-square is greater, namely 0.89910 & 0.87777 which indicates the model issued by the regression is better. Validity using streamlit to test the latest dataset with results showing the inputted independent variables can be classified with the results of the regression algorithm modeling

Keywords: Regression Algorithm; Air Pollutant Standard Index; Streamlit

Abstrak

Isu pencemaran udara cukup menjadi hal penting yang mengaitkannya dengan pertumbuhan hidup di perkotaan, isu ini biasa dikenal dengan polusi udara. Kondisi udara di Jakarta menunjukkan selama periode 2017-2020 umumnya udara dikategorikan tingkat sedang hingga tidak sehat. Hal ini diukur berdasarkan Indeks Standar Pencemaran Udara (ISPU) mengklasifikasikan kualitas udara di suatu daerah berdasarkan parameter zat-zat kimia serta akibat terhadap kesehatan. klasifikasi kualitas udara mulai dari kategori baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya, dapat diketahui dengan pemanfaatan algoritma tipe regression yaitu Logistic Regression (Regresi Logistik) dan Random Forest. Penelitian ini bermaksud untuk menghandalkan tipe algortima regression dengan merancang dashboard untuk mengklasifikasikan tingkat pencemaran udara di DKI Jakarta dengan library Steamlit. Selain itu, difungsikan sebagai validasi terhadap algortima yang telah di training terlebih dahulu mengenai tingkat evaluation dan keakuratannya. Dalam permasalahan klasifikasi indeks standar pencemar udara di DKI Jakarta menggunakan algoritma regression. Hasil *evaluation modelling*, untuk nilai sesuai berdasarkan proposi data *split* 80% : 20% dan 60% : 40% menujukkan bahwa secara *accuracy* lebih unggul algoritma *Random Forest* sebesar 96,66% & 96,22%. Dengan nilai MEA, MSE, RMSE terkecil secara berurut 0,03866, 0,04920, 0.2218. untuk data *split* 80% : 20%, kemudian nilai 0,045, 0,0615, 0,2480, pada data *split* 60% : 40%. Untuk nilai evaluasi r-square lebih besar yaitu 0.89910 & 0.87777 yang menunjukkan model yang dikeluarkan oleh regresi lebih baik. Dilakukan validitas menggunakan streamlit untuk menguji dataset terbaru dengan hasil menunjukkan variabel *independent* yang diinput dapat diklasifikasikan dengan hasil modeling algortima *regresston*.

Kata kunci: Algoritma Regression; Indeks Standar Pencemar Udara; Streamlit

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

1. Pendahuluan

Dalam proses pencemaran udara terdapat komponen atau zat lain yang masuk ke dalam udara, kualitas udara yang mengandung karbon monoksida (CO), nitrogen dioksida (NO_2), ozon (O_2) dan lainnya masuk kedalam tubuh masih dapat ditolerin bila nilai dibawah batas wajar sesaui rekomendasi organisasi kesehatan dunia (WHO). Akan tetapi, bila melampaui batas wajar dapat menyebabkan masalah serius pada kesehatan termasuk menyebabkan stroke, penyakit jantung, asma, infeksi pernapasan dan penyakit paru obstruktif kronis **[1].** Penyumbang emisi polusi udara akibat dari kegiatan manusia seperti pembangkit listrik tenaga uap berbahan bakar batu bara, minyak dan gas yang menghasilkan jumlah pencemar yang besar ke atmosfer bumi, hingga kualitas udara daerah sekitarnya.

Berdasarkan pemantauan yang dilakukan oleh kedutaan besar AS di Jakarta menunjukkan selama periode 2017-2020 umumnya udara di ibukota DKI Jakarta dikategorikan tingkat sedang hingga tidak sehat. Dalam klasifikasi udara di Indonesia sendiri diatur oleh Keputusan Badan Pengendalian Dampak Lingkungan (Bapedal) Nomor KEP-107/Kabepedal/11/1997 Tentang Indeks Standar Pencemaran Udara (ISPU) guna mengkategorikan kualitas udara di suatu daerah berdasarkan parameter penyebabnya serta pengaruh terhadap kesehatan. Kategori ini dibagi menjadi 5 (lima) yaitu kategori baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya. sedangkan untuk parameter indeks zat yang mencemari kualitas udara ini sendiri dibagi menjadi partikel debu (PM_{10}), partikel halus ($PM_{2.5}$), sulfur dioksida (SO_2), karbon monoksida (CO), nitrogen dioksida (NO_2), ozon (O_3). Data Indeks Standar Pencemar Udara (ISPU) ini dapat mudah diklasifikasikan levelnya dengan menggunakan algoritma dari *machine learning*. *machine learning* sendiri salah satu kemampuan komputer dalam belajar mengenai data baru tanpa diprogram secara eksplisit. Kemudian *machine learning* dalam proses pembelajarannya melalui pengalaman *(experience)* terhadap tugas (*task)* dan mengukur peningkatan kinerja (*performance measure)* **[2].** *Machine learning* memiliki beberapa algoritma pembelajaran yang dibagi menjadi 3 (tiga) yaitu *supervised learning, umsupervised learning* dan *reinforcement learning*, dalam memudahkan proses klasifikasi umumnya menggunakan algoritma *supervised learning* seperti *K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, Decission Tree, Random Forest Classifier* dan lainnya.

Penggunaan algortima *supervised learning* dalam mengklasifikasikan indeks standar pencemar udara telah diimplementasikan pada beberapa penelitian sebelumnya. Dalam penelitian terdahulu misalnya yang dilakukan oleh **[3]** yaitu sistem prediksi kualitas udara berhasil dilakukan menggunakan metode *Support Vector Machine kernel RBF* dengan tingkat akurasi sebesar 96,03%. Tetapi kelemahan yang dimiliki pada penelitian ini hanya menggunakan satu metode serta hanya menggunakan data di 2 (dua) periode tahun 2017-2018 sehingga tingkat akurasi masih dapat berubah bila data yang digunakan lebih banyak atau dalam kurun waktu 5 tahun. Kemudian penelitian serupa dilakukan oleh **[4]** yaitu Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara di DKI Jakarta menggunakan algoritma *neural network backpropagation, support vector machine, k-nearest neighbors, naive bayes* dan *decission tree,* pada penelitian yang dilakukan didapat nilai akurasi tertinggi oleh algortima *decission tree* sebesar 99,80%, nilai kappa yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Namun pada penelitian ini tidak hanya melakukan traning data saja tanpa menguji atau mengevaluasi dengan dataset yang baru. hal ini menjadi kelemahan dari penelitian yang dilakukan untuk menguji nilai akurasi berdasarkan training data yang dimiliki tidak menguji dan memvalidasi dataset yang terbaru diinputkan juga tidak terdapat *deploy* dari hasil *training* dan tidak adanya pembagian antara data *training* dan data *testing*. Oleh karena itu pada penelitian ini, peneliti memanfaatkan *streamlit* guna menguji dan memvalidasi data hasil *training* dan *testing* data baru untuk melihat hasil klasifikasi dari data yang diinputkan berdasarkan variabel *dependent* atau target level dari Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta.

Dalam mengklasifikasikan tingkat pencemaran udara menggunakan algoritma tipe regression yaitu Logistic Regression (Regresi Logistik) dan Random Forest. Evaluasi menggunakan nilai Confusion Matrix Accuracy dan evaluation regression mulai dari Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared. Melakukan data split data training dan data testing, terakhir untuk deployment sistem menjadi dashboard menggunakan library python Streamlit. menguji proses modelling dengan algortima dapat mengklasifikasikan data yang baru diinputkan. Data inputan tersebut terdiri dari nilai karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan partikel debu (PM_{10}) yang menjadi variabel independent. Tujuan penelitian sendiri guna mengetahui kaakuratan algortima regression dalam mengklasifikasikan data tingkat pencemaran udara di DKI Jakarta dengan algoritma regression yaitu Logistic Regression (Regresi Logistik) dan Random Forests.

2. Metode Penelitian

Pada bagian artikel ini menjelaskan tahapan-tahapan secara rinci mengenai rancangan penelitian mulai dari pengumpulan data (*data acquisition)*, *training & evaluation model* hingga *deployment.*

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

A. Pengumpulan data (*Data Acquisition*)

Pengumpulan data bermaksud untuk mencari data-data yang terkait dengan penelitian, Dalam hal ini diperlukan beberapa spesifikasi pendataan yang sesuai dengan topik.

Berdasarkan proses ini, *dataset* didapatkan dari *website* resmi milik Pemerintah Provinsi DKI Jakarta yaitu Open Data Jakarta (data.jakarta.go.id) jenis dataset Indeks Standar Pencemaran Udara (ISPU) di Provinsi DKI Jakarta. Pengumpulan *dataset* dimulai dari periode 2013 hingga 2021. Dengan Jumlah dataset tiap tahunnya yang terkumpul sebanyak 96, Keseluruhan data sebanyak 2.913. Terdiri atas Tanggal,PM_{10},So_2,CO,O_3,NO_2,Max,Critical,Categori,Lokasi_spku.

B. Persiapan data (*Data Preprocessing)*

Perolehan data awal masih belum dapat diproses, maka perlu beberapa perubahan kolom mulai dari perubahaan tipe data sesuai kebutuhan. Proses pengolahan dataset dibagi menjadi beberapa tahap. Tahapan ini memiliki fungsionalitas masing-masing guna memudahkan proses pemahaman pada data. Pertama melakukan pembersihan data (*data cleaning)*, rekayasa fitur (*feature engineering)* dan pembagian data (*data split).* Tahap ini termasuk bagian persiapan data (*data preprocessing).*

Dalam bidang data science, memproses *dataset* memerlukan beberapa pembersihan data, hal ini dilakukan guna meminalisir hasil yang tidak sesuai. Sebab *dataset* yang diperoleh memiliki beberapa komponen attribute kolom yang tidak dibutuhkan. Sehingga kehadiran data *cleaning* merupakan hal penting dalam mendeteksi, memperbaiki dan menghapus data yang rusak, data ganda serta data yang tidak diperlukan dalam dataset, tabel atau basis data agar data berkualitas dan tepat untuk digunakan pada proses data *training*. Tujuan lain dari *data cleaning* ialah agar semua data konsisten, seragam, lengkap dan valid. Salah satu cara untuk melakukan data *cleaning* menggunakan statitika dalam mengisi *missing value,* yaitu dengan nilai rata-rata (*mean)* pada variabel *independent.* Secara matematis untuk memperoleh rata-rata (*mean)* dengan persamaan **(1) [5]**

$$\bar{x} = \frac{\sum x_i}{n} \qquad (1)$$

Dimana nilai rata-rata dari kumpulan data, x_i nilai data ke-i dan n banyaknya data. Hasil rata-rata ini hanya dapat dikalkulasi melalui data bertipe angka. Kemudian setelah ini, melakukan visualisasi data berdasarkan variabel *independent* dan *dependent*, hal ini dilakukan untuk mendapatkan informasi dari data dalam bentuk gambar diagram.

Normalisasi data pada tipe kategori ke dalam tipe numerik. Dalam hal ini untuk variabel *dependent* atau attribute target diubah menjadi skala angka mulai dari indek 0 sampai 3. Setelah melalui normalisasi yaitu dengan proses *feature engineering* merupakan proses pembentukan atau mengekstrak fitur-fitur, parameter-parameter dan atribut-atribut pada dataset untuk keperluan proses *training* **[2]**. Proses ini menghasilkan pembelajaran mesin yang mampu memecahkan masalah lebih cepat dan akurat. Dengan metode *feature extraction* yang membagi variabel berdasarkan nilai X dan y. X mewakili nilai *pm10, so2, co, o3, no2.* Sedangkan y menampung nilai *category*. Setelah melakukan proses *features engineering* melakukan data *split* suatu proses yang memecahkan data menjadi dua bagian berdasarkan tujuannya pertama data *training* (data latih) dan data *testing* (data pengujian). Data *training* digunakan untuk melatih dan membangun model, kalau data *testing* digunakan untuk melakukan pengujian sekaligus mengukur performa model yang dibangun. Untuk data *split* menggunakan metode *train test split* Berikut ini **Tabel 1** pembagian data *split.*

Tabel 1. *Splitting* dataset

Data *training*	Data *testing*	Jumlah data *training*	Jumlah data *testing*
80 %	20 %	2274	569
60 %	40 %	1705	1138

C.* Algortima *Regression

Supervised learning adalah pembelajaran tearah/terawasi, dadlam proses pembelajaran data *training* yang memiliki target atau label capaian. Termasuk dalam kelompok *machine learning.* Jenis *supervised learning* ini terbagi menjadi tipe algortima *regression*, umumnya digunakan untuk peramalan atau klasifikasi. Fokus dari algortima *regression* ini ialah memetakan variable *independent* ke variable *dependent*. *feature/predictor/attribute* tabel menjadi *independent variable,* serta untuk *dependent* ialah target dari dataset tersebut. Pembeda antara tipe *classification* dan *Regression* dari cara mengukur kinerja performa.

Untuk *classification* mengukur kinerja berdasarkan perbandingan jumlah dugaan benar dan salah. Untuk *regression* dinilai berdasarkan kedekatanya dengan nilai asli dari dugaannya. Kemudian dalam mengategorikan target, *regression* biasanya menduga sebuah bilangan angka (*continue*) sejumlah data, *classification* menduga berdasarkan kelompok/kategori/kelas dari data. Umumnya algoritma tipe *regression* ini dikelompokan menjadi *Logistic Regression* (Regresi Logistik) dan *Random Forests, Ordinary Least Squares Regression (OLSR), Regresi Stepwise, Multivariate Adaptive Regression Splines (MARS), Locally Estimated Scatterplot Smoothing (LOESS).*

Pembuatan model *training* dengan algoritma *regression* hal utama dalam proses penelitian ini, pemilihan model pertama ialah menggunakan *Logistic Regression* (Regresi Logistik) sebagai metode yang dapat digunakan dalam mencari hubungan variabel

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

respon yang bersifat dikotomus (bersakala ordinal atau nominal) atau polikotomus (memiliki skala nominal atau ordinal dengan lebih dari 2 kategori). Ketika variabel dependen memiliki skala yang bersifat polikotomus atau multinomial maka dapat digunakan regresi logistik multinomial **[6]**. Algoritma ini dipilih karena pada target termasuk kedalam kategori skala ordinal. Secara teknis *Logistic Regression* ialah *regression linear* yang di substitusi kedalam fungsi logistic, untuk itu representasi dari persamaan *Logistic Regression* yang disebut dengan fungsi aktivasi sigmoid **(2) [7]**.

$$f(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (2)$$

Dimana f(x) adalah output yang diprediksi, sedangkan teta adalah koefiesien, dan x adalah variabel *independent.*

Random Forest sebuah algoritma yang dikembangkan dari metode CART (*Classification and Regression Trees*) dalam hal ini ialah algoritma teknik pohon keputusan **[8]**. Metode ini memiliki beberapa kelebihan antara lain, menghasilkan hasil klasifikasi yang baik, menghasilkan error yang lebih rendah, secara efisien dapat mengatasi data training dengan jumlah data yang sangat besar dan menghasilkan satu set pohon acak **[9].** Semakin banyak pohon (*tree*) semakin besar pula akurasi yang didapatkan**_[10].** Berikut ini represntasi dari *Random Forest* **(3)-(3.1) [11]**

$$\textbf{Entropy (Y)} = -\sum \textbf{p (c|Y)i log2 p (c|Y)} \quad (3)$$

Dimana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$\textbf{Information Gain (Y, a)} = \boldsymbol{Entropy\ (Y)} - \sum \boldsymbol{veValues} \frac{|\boldsymbol{Yv}|}{|\boldsymbol{Ya}|} \boldsymbol{Entropy\ (YV)} \quad (3.1)$$

Dimana Values (a) semua nilai yang mungkin dalam himpunan kasus a. Yv subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya semua nilai yang sesuai dengan a.

D. Evaluation

Evaluation ialah proses untuk mengukur performa model yang telah di proses melalui *training.* Beragam macam metode untuk mengevaluasi model yang telah di bangun. Beberapa metrik yang digunakan *confusion matrix,* matrik ini menyajikan ringkasan semua hasil prediksi yang dihasilkan dengan membandingkan antara hasil periksi dan hasil yang diharapkan, identik dengan 4 kolom nilai *true dan false* untuk positif dan negatif. Matrik *confusion matrix* dapat menghitung beberapa matrik lainnya seperti *accuracy*, *recall, specificity, precision, false positive rate, false negative rate, f-1 score.* Pada penelitian ini menggunakan matrik *confusion matrix accuracy* yaitu membandingkan jumlah item yang di prediksi benar dengan total seluruh prediksi yang dilakukan. Adapun persamaan **(4)**

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

Keterangan:
TP: True Positif
TN: True Negatif
FP: False Positif
FN: False Negatif

Selain itu, untuk evaluasi jenis metode *regression* ini digunakan model evaluasi *matrix Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared. Matrix Mean Absolute Error (MAE)* mengukur tingkat keakuratan model prediksi Persamaan dari *MEA* ini **(5)[12].**

$$\boldsymbol{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{fi - yi}| \quad (5)$$

Nilai fi hasil peramalan, yi nilai sebenarnya, serta n adalah banyak data

Matrix Mean Squared Error (MSE) berfungsi untuk menghitung rata-rata selisih kuadrat antara nilai yang diramalkan dengan yang diamati, MSE akan mengevaluasi suatu metode yang mengkuadratkan hasil dari kesalahannya, berikut metode persamaan dari *MSE* **(6)[13]**

$$\boldsymbol{MSE} = \sum \frac{\boldsymbol{Et^2}}{\boldsymbol{n}} \quad (6)$$

Et^2 nilai dari galat kuadrat, dan n ialah jumlah data

Metode pengukuran dengan mengukur perbedaan nilai dari yang prediksi sebuah model dengan nilai observasi atau label. RMSE dihitung dari akar kuadrat *mean square error.* Nilai ini berkisar dari 0 hingga tak hingga. Dimana untuk nilai RMSE kecil maka model tersebut mendekati nilai observasinya atau dapat dikatakan akurat. Persamaan dari metode RMSE ini **(7)[14].**

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum(\boldsymbol{X - Y})^2}{\boldsymbol{n}}} \quad (7)$$

Koefisien determinasi (R Square atau R kuadrat) metode pengukuran pangaruh dari variabel *independent* terhadap variabel *dependent.* Nilai R suared ialah 0 sampai 1, semakin deket nilai 1 maka model yang dikeluarkan oleh algoritma *regression* tersebut semakin baik, R-Square memiliki persamaan **(8)[15].**

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

$$R = 1 - \frac{SSE}{SST} \quad (8)$$

Keterangan:
SSE: *Sum of Squares of the Regression* (Jumlah Kuadrat
Regresi)
SST: *Total Sum of Squares* (Total Jumlah Kuadrat)

E. Deployment Dengan Streamlit

Aktivitas akhir dari penelitian yaitu menerapkan model-model dari algoritma *regression,* dalam hal ini dilakukan untuk menjaga skalabilitas dari model *training.* Dalam prosesnya menggunakan tools dari *library* python yaitu Streamlit. Proses ini membuat sebuah dashboard yang bertujuan melakukan *deployment* pada model terbaik sekaligus sebagai validasi inputan pada masing-masing variabel *independent* untuk melihat hasil klasifikasi tingkat pencemar udara sesuai Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta. Selain itu, implementasi *deployment* ini dapat membandingkan hasil klasfikasi tersebut. *Deployment* yang dikembangkan memanfaatkan *library python* yaitu *Streamlit.*

3. Hasil dan Pembahasan

Pada bagian ini menjelaskan secara rinci hasil dan capaian yang telah dilakukan dalam permasalahan klasifikasi indeks standar pencemar udara di DKI Jakarta menggunakan algoritma *regression.* Kemudian hasil dari *training model* tersebut akan di implementasikan antarmuka berupa dashboard dengan memanfaatkan streamlit

A. Dataset

Berdasarkan hasil pengumpulan dan seleksi dataset indeks standar pencemar udara format .csv yang bersumber dari laman situs milik pemerintah provinsi DKI Jakarta. Kategori dataset diambil per Indeks standar pencemar udara di DKI Jakarta per bulan. Data ini dikumpulkan dari tahun 2013 – 2021, kalkulasi dataset dan atribut isi data berhasil dikumpulkan dalam **Tabel 2**:

Tabel 2. Kalkukasi dataset

Tahun	Jumlah Dataset	Jumlah Seluruh Data dalam Dataset
2021	12	365
2020	12	366
2019	12	345
2017	12	369
2016	12	366
2015	12	372
2014	12	365
2013	12	365
Jumlah	96	2.913

Jumlah dataset yang terkumpul sebanyak 96, dengan total data sebanyak 2.913. kemudian diolah menjadi satu dataset yang memiliki *columns* serupa dengan menggunakan *append* dataset tiap tahunnya. Hal ini dilakukan untuk mempermudah proses data *preprocessing.* Sebelum masuk ketahap tersebut, dilakukan proses membaca dataset melalui *library python pandas.* Untuk melihat secara keseluruhan atribut dalam dataset maka melakukan pemetaan data **Tabel 3.**

Tabel 3. Pemetaan dataset ISPU

Atribut	Diperlukan	Deskripsi
Unnamed: 0	Tidak	Nomor data
categori	Ya (Variabel Dependen)	Klasifikasi hasil perhitungan dari parameter ISPU
CO	Ya (Variabel Independent)	Karbon Monoksida
Max	Tidak	Nilai maksimal untuk parameter ISPU tertinggi
NO_2	Ya (Variabel Independent)	Nitrogen Dioksida
O_3	Ya (Variabel Independent)	Ozon
PM_{10}	Ya (Variabel Independent)	Partikel debu berukuran 10 mikron
SO_2	Ya (Variabel Independent)	Sulfur dioksida
Tanggal	Tidak	Waktu pengambilan data parameter

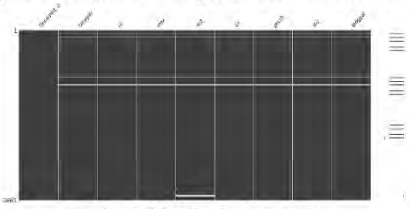
Setelah memeta dataset, bagian dari data *preprocessing* dilakukan proses *checking missing value* dataset. Proses ini akan menghasilkan *column* memiliki kekosongan data atau *inputan* yang tidak sesuai., Berikut ini rincian dari *missing value* pada dataset, Perhatikan **Tabel 4**.

Tabel 4. *Missing Value*

Atribut	*Missing value*
Unnamed: 0	0
Category	18
CO	18
Max	18
NO_2	49
O_3	18
PM_{10}	18
SO_2	18
Tanggal	18

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

Untuk melihat secara visualisasi *missing value* terdapat pada beberapa *row* ditunjukkan pada **Gambar 1**.



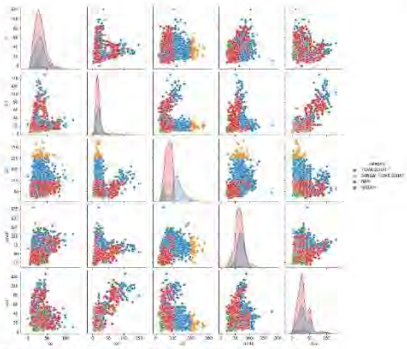
Gambar 1. Visualisasi *missing value*

Perlakuan yang dilakukan terhadap *missing value* ini dengan salah satu metode *measures of central tendency* yaitu memasukan nilai *mean* sesuai nilai atribut pada data. dan menghilangkan baris yang tidak dibutuhkan. Guna memudahkan pemahaman mengenai nilai yang terdapat dalam variabel *dependet* maka divisualisasikan melihat sebaran data berdasarkan category pada dataset. Hal ini dilakukan untuk memudahkan proses *feature engineering*. Perhatikan **Gambar 2**

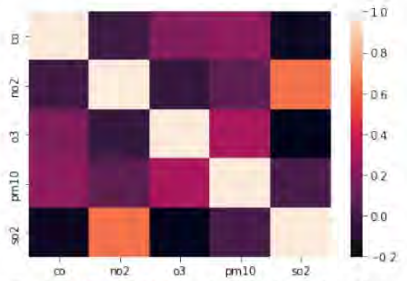
Gambar 2. Persentase jumlah *category* dataset

Berdasarkan sebaran nilai variabel *dependent* tersebut, terdapat nilai yang akan menjadi klasifikasi terhadap variabel *independent*. Nilai data terbanyak terklasifikasi sebagai kategori sedang, tidak sehat, baik dan sangat tidak sehat. Kemudian untuk memudahkan proses pembelajaran terhadap *goals* dilakukan *transformasi* tipe data pada parameter tersebut menjadi *integer*. Target atau variable *dependent* ini dikonversi dengan *label Encoding* menjadi index skala 0-3a, skala 0 mewakili kategori baik, nilai 1 untuk kategori sangat tidak sehat, nilai 2 mewakili sedang, 3 untuk tidak sehat, sehingga dari proses in jumlah dataset yang akan di lakukan *training model* sebanyak 2843 data. Pada proses visualisasi data dilakukan untuk melihat sebaran pengaruh antara parameter tatau variabel *independent*.

Gambar 3 penjelasan korelasi antara variabel *independent*. Secara probabilitas juga korelasi antara variabel *independent*, perhatikan **Gambar 4**.



Gambar 3. Korelasi tiap variabel *independent*



Gambar 4. Probabilitas korelasi variabel *independent*

Feature engineering langkah untuk membagi nilai variabel berdasarkan tujuannya, untuk variabel *independent* disatukan dalam variabel X, serta untuk variabel *dependent* diwakili variabel y. Terakhir dalam persiapan dataset ini dilakukan pembagian data berdasarkan data *training* dan data *testing* seperti tabel 1.

B. *Training dan Evaluation Model*

Membangun model *training* pada dataset digunakan algoritma *regression*, algoritma pertama *Logistic Regression* dengan parameter max_iter=200 juga mengaktifkan penggunaan StandardScaler, selanjutnya algoritma kedua menggunakan Random Forest dengan parameter n_estimators=100. Kemudian setelah proses *training* model menguji kedua algoritma tersebut dengan beberapa nilai evaluasi model untuk mengukur keakuratan data yang direpresentasikan dalam evaluasi

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

model regresi (*mean_absolute error, mean squared error, root mean squared error & r-square)* dan serta menggunakan model evaluasi *confusion matrix,* perhatikan Tabel 5 sampai Tabel 8. Kemudian untuk ringkasan dari hasil evaluasi dapat dilihat pada Tabel 9 & Tabel 10

Tabel 5. *Evaluation confusion matrix (80% : 20%)*

	PB	PSTS	PS	PTS	CR
TB	0.83	0	0.028	0	57,69%
TSTS	0	1	0	0.02	78,57 %
TS	0.17	0	0.82	0.14	92,75%
TTS	0	0	0.15	0.84	69,02%
CP	83,33 %	100%	82,47%	83,55 %	
FS	68,18 %	88,00 %	87,31 %	75,60 %	

Tabel 6. *Evaluation confusion matrix Random Forest (80% : 20%)*

	PB	PSTS	PS	PTS	CR
TB	0.93	0	0.0028	0	57,69%
TSTS	0	1	0	0	78,57 %
TS	0.074	0	0.96	0.017	92,75%
TTS	0	0	0.037	0.98	69,02%
CP	83.33 %	100 %	82,47 %	83,55 %	
FS	68,18 %	88,00 %	87,31 %	75,60 %	

Tabel 7. *Evaluation confusion matrix Logistic Regression (60% : 40%)*

	PB	PSTS	PS	PTS	CR
TB	0.82	0	0.028	0	63,79%
TSTS	0	1	0	0.021	65,00 %
TS	0.18	0	0.85	0.15	91,91%
TTS	0	0	0.13	0.83	74,46%
CP	82.22 %	100 %	84,69 %	83,28 %	
FS	71,84 %	78,79 %	88,15 %	78,62 %	

Tabel 8. *Evaluation confusion matrix Random Forest (60% : 40%)*

	PB	PSTS	PS	PTS	CR
TB	0.93	0	0.0028	0	57,69%
TSTS	0	1	0	0.0086	78,57 %
TS	0.067	0	0.96	0.017	92,75%
TTS	0	0	0.039	0.97	69,02%
CP	93.33 %	100 %	95,79 %	97,42 %	
FS	94,92 %	91,89 %	97,15 %	94,84 %	

Keterangan:
PB : *Predicted* Baik
PSTS : *Predicted* Sangat Tidak Sehat
PTS : *Predicted* Tidak Sehat
PTS : *Predicted* Sedang
TB : *True* Baik
TSTS : *True* Sangat Tidak Sehat
TS : *True* Sedang
TTS : *True* Tidak Sehat
CR : *Class Precision*
CP : *Class Recall*
FS : *F-1 Score*

Tabel 9. Ringkasan *Evaluation Model (80% : 20%)*

Algoritma	*Accuracy*	MEA	MSE	RMSE	R2
Logistic Regression	83,13 %	0,19859	0.25834	0.50827	0.47032
Random Forest	96,66 %	0,03866	0,04920	0.2218	0.89910

Tabel 10. Ringkasan *Evaluation Model (60% : 40%)*

Algoritma	*Accuracy*	MEA	MSE	RMSE	R2
Logistic Regression	84,36 %	0,18804	0.25131	0.50131	0.50061
Random Forest	96,22 %	0,04569	0,06151	0.24801	0.87777

Berdasarkan ringkasan *evaluation model,* yang terbagi menjadi dua nilai sesuai dengan proposi data *split* 80% : 20% dan 60% : 40% menujukkan bahwa secara *accuracy* lebih unggul algoritma *Random Forest* sebesar 96,66% & 96,22% menunjukkan algoritma ini lebih baik modelnya. Dengan nilai MEA, MSE, RMSE

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

terkecil secara berurut 0,03866, 0,04920, 0.2218 untuk data *split* 80% : 20%, kemudian nilai 0,045, 0,0615, 0,2480, dengan data *split* 60% : 40% yang menunjukkan keakuratan nilai error yang baik untuk nilai evaluasi r-square lebih besar yaitu 0.89910 & 0.87777 yang menunjukkan model yang dikeluarkan oleh regresi lebih baik. Sedangkan untuk algortima *logistic regression* untuk nilai *accuracy* sebesar 83,13% & 84,36%. Dengan nilai evaluasi MEA, MSE, RMSE secara berurut 0,19859, 0,25834, 0.5082 untuk data *split* 80% : 20%, kemudian nilai 0,18804, 0,25131, 0,5013 dengan data *split* 60% : 40%, sedangkan untuk nilai r-square yaitu 0,47032 & 0,50061.

C. *Deployment Model*

Proses *Deployment* dengan *Streamlit* ini akan memudahkan dalam memklasifikasi Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta. Juga digunakan untuk melihat hasil skalabilitas dari model *regression* yang telah digunakan sebelumnya., secara fungsionalitas akan terdapat pilihan jenis algoritma dan inputan manual ataupun berupa file jenis csv yang di *generate* otomatis. Untuk inputan itu sendiri terdapat variabel *independent* yaitu nilai dari pm10, so2, co, o3 dan no2. Kemudian dapat memilih jenis algortima yang diinginkan Secara otomatis berdasarkan algortima yang telah ada, akan memklasifikasikan tingkat pencemaran udara di DKI Jakarta berdasarkan Indeks Standar Pencemar Udara (ISPU). Beberapa hal yang dibutuhkan dalam menbangun *Streamlit* yakni menginstall dan running hasil code yang dibuat melalui npx localtunnel. Serta untuk menyimpan hasil algoritma yang telah *training* kedalam pickle. Perhatikan **Gambar 5**, hasil dari proses *deployment* mengimplementasikan *Streamlit*.

Gambar 5. Dashboard Streamlit

Kemudian untuk menguji tingkat keakuratan terhadap algoritma maka diinputkan secara manual melalui Langkah pada **Gambar 6** – **Gambar 7**. Kemudian sebagai validasi Kembali lihat dataset yang inputkan salah satu nilai sama pilih salah satu, **Gambar 8** – **Gambar 9,** hasil dari inputan data baru menghasilkan klasifikasi yang sama.

Gambar 6. Input data dengan algortima *Logistic Regression*

Gambar 7. Input data dengan algortima *Random Forest*

Gambar 8. Hasil klasifikasi algoritma *Logistic Regression*

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

Gambar 9. Hasil klasifikasi algoritma *Random Forest*

Berdasarkan data input, dashboard dapat berhasil mengklasifikasikan tingkat pencemaran udara di DKI Jakarta dengan kategori Tidak Sehat, untuk masing-masing inputan variabel untuk co, no2, o3, pm10, so2 sebesar 39,18,3,61,34. Namun ada beberapa yang tidak sesuai klasifikasinya untuk kedua algoritma tersebut.

4. Kesimpulan

Implementasi penggunaan Streamlit yang berfungsi sebagai pengujian validasi atas model algortima *regression* yaitu *Logistic Regression* dan *Random Forest* menghasilkan nilai yang sesuai dengan dataset, namun terdapat beberapa perbedaan klasifikasi oleh variabel *dependent* dari kedua algortima tersebut. Hal ini disebabkan dari dominasi nilai variabel *dependent* yang tidak seimbang jumlahnya, kemudian untuk nilai model evaluasi yang di hasilkan dari kedua model tersebut menjadi cukup berbeda mulai dari nilai *confusion matrix*, tingkat *accuracy* dan evaluasi model *regression*, jauh lebih unggul algortima Random Forest dengan nilai masing-masing evaluasi berdasarkan data split 80% : 20% dan 60% : 40% , accuracy sebesar 96,66% & 96,22% menunjukkan algoritma ini lebih baik modelnya. Dengan nilai MEA, MSE, RMSE terkecil secara berurut 0,03866, 0,04920, 0.2218 untuk data split 80% : 20%, kemudian nilai 0,045, 0,0615, 0,2480, dengan data split 60% : 40%. Untuk nilai evaluasi r-square yaitu 0.89910 & 0.87777. Sedangkan, algortima *Logistic Regression* menghasilkan nilai accuracy sebesar 83,13% & 84,36%. Dengan nilai evaluasi MEA, MSE, RMSE secara berurut 0,19859, 0,25834, 0.5082 untuk data split 80% : 20%, kemudian nilai 0,18804, 0,25131, 0,5013 dengan data split 60% : 40%, sedangkan untuk nilai r-square yaitu 0,47032 & 0,50061.

Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada seluruh teman-teman, staf, dan civitas akademika program studi teknik informatika universitas mercubuana yang telah mendukung dan membimbing penelitan ini.

Daftar Rujukan

1. L. Myllyvirta., I. Suarez., dan E. Uusivuori., 2020. Pencemaran Udara Lintas Batas di provinsi Jakarta, Banten, dan Jawa Barat.
2. M. Linawati., 2019. "Data AirVisual: Kualitas Udara DKI Jakarta Peringkat 1 Terburuk di Unri., dkk, Analisis dan Komparasi ... 104 Dunia - News Liputan6.com." Liputan 6, 23 September. Tersedia [https://www.liputan6.com/news/read/4069080 /data-airvisual-kualitas-udara-dki-jakartaperingkat-1-terburuk-di-dunia] diakses 1 Maret 2021.
3. Bapelda,. 1999 "PEDOMAN TEKNIS PERHITUNGAN DAN PELAPORAN SERTA INFORMASI INDEKS STANDAR PENCEMAR UDARA,".Tersedia [http://www.cetsuii.org/BML/Udara/ISPU/ISPU%20(Indeks%2 0Standar%20Pencemar%20Udara).htm.] diakses 1 Maret 2021.
4. Id, I. D. (2021). Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python. UNRI Press
5. Hermawan A. 2019. SPKU : Sistem Prediksi Kualitas Udara (Studi Kasus : DKI Jakarta).
6. Syekh S A, Firdaus M S, dkk. 2021. Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara Di DKI Jakarta. Jurnal Informatika dan Komputer, 4(2), 98-104. https://doi.org/10.33387/jiko
7. Ferdiansyah, P., Indrayani, R., dan Subektiningsih, S. (2020). Analisis Manajemen Bandwidth Menggunakan Hierarchical Token Bucket Pada Router dengan Standar Deviasi, Jurnal Nasional Teknologi dan Sistem Informasi, 6(1). 38–45. https://doi.org/10.25077/teknosi.v6i1.2020.38–45
8. Z. Siregar, "Implementasi Metode Regresi Linier Berganda Dalam Estimasi Tingkat Pendaftaran Mahasiswa Baru," KESATRIA : Jurnal Penerapan Sistem Informasi (Komputer & Manajemen), vol. 2, no. 3, pp. 133–137, 2021.
9. Junifer J P, Tanjaya H, and Kenichi. 2021. Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression. Information System Development. Vol. 6, no. 2. 2021
10. P. W. Fica Oktavia Lusiana*, IndriFatma, "Estimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Simalungun," Journal of Informatics Management and Information Technology, vol. 1, no. 2, pp. 79–84, 2021.
11. M. A. H. Laksamana, Amroni, and A. N. Toscany, "Penerapan Data Mining untuk Memprediksi Jumlah Total Produksi Hcl Pada Perusahaan PT . Lontar Papyrus Menggunakan Algoritma Regresi Linier Berganda," Jurnal Ilmiah Mahasiswa Teknik Informatika, vol. 3, no. 2, pp. 187–198, 2021.
12. N. Arminarahmah, A. D. GS, G. W. Bhawika, M. P. Dewi, and A. Wanto, "Mapping the Spread of Covid-19 in Asia Using Data Mining X-Means Algorithms," IOP Conf.

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

Series: Materials Science and Engineering, vol. 1071, no. 012018, pp. 1–7, 2021.

13. Wanika V.W, and Elvina I, M. 2018. Prediksi Harga Ponsel Menggunakan Metode Random Forest. Computer Science and ICT. Vol.4, No.1. 2018.

14. Adi A.S, Muqtadir A. 2019. Penerapan Metode Mean Absolute Error (MEA) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi. SAINTEKBU: Jurnal Sains dan Teknologi. Vol.11, No.1, 2019

15. Yulia R.H. 2017. Peramalan Persediaan Barang Menggunakan Metoe Weighted Moving Average Dan Metode Double Exponential Smoothing. Junral Pilar Nusa Mandiri. Vol. 13, No.2, 2017.

16. Suprayogi I, Trimaijon, Mahyudin. Model Prediksi Liku Kalibrasi Menggunakan Pendekatan Jaringan Saraf Tiruan (JST) Studi Kasus: Sub DAS Siak Hulu.

17. Nurmasani A, Utami E, Fatta A.H, 2017. Analisis Support Vector Machine Pada Prediksi Produksi Komoditi Padi. Jurnal Informasi Interaktif, Vol. 2, No.1.

18. NIA KUSUMA WARDHANI[1], REZKIANI[2], SIGIT KURNIAWAN3 , HENDRA SETIAWAN4 , GRACE GATA5 , SISWANTO TOHARI 6, WINDU GATA7 , MOCHAMAD WAHYUDI8 SENTIMENT ANALYSIS ARTICLE NEWS COORDINATOR MINISTER OF MARITIME AFFAIRS USING ALGORITHM NAIVE BAYES AND SUPPORT VECTOR MACHINE WITH PARTICLE SWARM OPTIMIZATION. 31st December 2018. Vol.96. No 24

19. M O Pratama, W Satyawan, R Jannati, B Pamungkas, Raspiani, M E Syahputra and I Neforawati The sentiment analysis of Indonesia commuter line using machine learning based on twitter data. Politeknik Negeri Jakarta, Depok, Indonesia

Diterima Redaksi: xx-xx-2022 | Selesai Revisi: xx-04-2022 | Diterbitkan Online: xx-04-2022

KERTAS KERJA

Ringkasan

Kertas kerja ini merupakan kelengkapan artikel jurnal dengan judul Implementasi Streamlit untuk Mengklasifikasikan Tingkat Pencemaran Udara di DKI Jakarta dengan Penggunaan Algoritma Regression yang berisi semua material hasil penelitian Tugas Akhir yang tidak dimuat atau disertakan di artikel jurnal. Dlam kertas kerja ini akan dijelaskan mengenai literature review, dataset yang digunakan, serta Langkah-langkah perancangan, tahapan implementasi dan hasil pengujian penelitian

Pendahuluan

Dalam proses pencemaran udara terdapat komponen atau zat lain yang masuk ke dalam udara, kualitas udara yang mengandung karbon monoksida (CO), nitrogen dioksida (NO_2), ozon (O_2) dan lainnya masuk kedalam tubuh masih dapat ditolerin bila nilai dibawah batas wajar sesaui rekomendasi organisasi kesehatan dunia (WHO). Akan tetapi, bila melampaui batas wajar dapat menyebabkan masalah serius pada kesehatan termasuk menyebabkan stroke, penyakit jantung, asma, infeksi pernapasan dan penyakit paru obstruktif kronis**.** Penyumbang emisi polusi udara akibat dari kegiatan manusia seperti pembangkit listrik tenaga uap berbahan bakar batu bara, minyak, gas dan kendaraan bermotor yang menghasilkan jumlah pencemar yang besar ke atmosfer bumi, hingga kualitas udara daerah sekitarnya.

Tingkat polusi udara yang semakin meningkat terutama di kota-kota besar sangat membahayakan bagi lingkungan dan kesehatan masyarakat. Salah satu penyumbang polusi adalah dari kendaraan bermotor. DKI Jakarta sebagai ibukota Republik Indonesia sekaligus kota metropolitan dan pusat perekonomian mengalami masalah yang sangat rumit dalam bidang transportasi. Jumlah penduduk yang banyak dengan daya beli yang meningkat menyebabkan kepemilikan kendaraan bermotor cukup tinggi. Dampak dari banyaknya kendaraan bermotor

yaitu semakin banyaknya sisa dari bahan bakar minyak yang dikeluarkan. Jakarta memiliki kualitas udara dengan kategori tidak sehat.

Berdasarkan pemantauan yang dilakukan oleh kedutaan besar AS di Jakarta menunjukkan selama periode 2017-2020 umumnya udara di ibukota DKI Jakarta dikategorikan tingkat sedang hingga tidak sehat**.** Dalam klasifikasi udara di Indonesia sendiri diatur oleh Keputusan Badan Pengendalian Dampak Lingkungan (Bapedal) Nomor KEP-107/Kabepedal/11/1997 Tentang Indeks Standar Pencemaran Udara (ISPU) guna mengkategorikan kualitas udara di suatu daerah berdasarkan parameter penyebabnya serta pengaruh terhadap kesehatan. Kategori ini dibagi menjadi 5 (lima) yaitu kategori baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya. sedangkan untuk parameter indeks zat yang mencemari kualitas udara ini sendiri dibagi menjadi partikel debu (PM_{10}), partikel halus ($PM_{2.5}$), sulfur dioksida (SO_2), karbon monoksida (CO), nitrogen dioksida (NO_2), ozon (O_3). Data Indeks Standar Pencemar Udara (ISPU) ini dapat mudah diklasifikasikan levelnya dengan menggunakan algoritma dari *machine learning. machine learning* sendiri salah satu kemampuan komputer dalam belajar mengenai data baru tanpa diprogram secara eksplisit. Kemudian *machine learning* dalam proses pembelajarannya melalui pengalaman *(experience)* terhadap tugas (*task)* dan mengukur peningkatan kinerja (*performance measure)***.*** *Machine learning* memiliki beberapa algoritma pembelajaran yang dibagi menjadi 3 (tiga) yaitu *supervised learning, unsupervised learning* dan *reinforcement learning*, dalam memudahkan proses klasifikasi umumnya menggunakan algoritma *supervised learning* seperti *K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes, Decission Tree, Random Forest Classifier* dan lainnya.

Penggunaan algortima *supervised learning* dalam mengklasifikasikan indeks standar pencemar udara telah diimplementasikan pada beberapa penelitian sebelumnya. Dalam penelitian terdahulu misalnya yang dilakukan oleh yaitu sistem prediksi kualitas udara berhasil dilakukan menggunakan metode *Support Vector Machine kernel RBF* dengan tingkat akurasi sebesar 96,03%. Tetapi kelemahan yang dimiliki pada penelitian ini hanya menggunakan satu metode serta hanya menggunakan data di 2 (dua) periode tahun 2017-2018 sehingga tingkat akurasi masih dapat berubah bila data yang digunakan lebih banyak atau dalam kurun waktu

5 tahun. Kemudian penelitian serupa dilakukan oleh yaitu Analisis dan Komparasi Algoritma Klasifikasi Dalam Indeks Pencemaran Udara di DKI Jakarta menggunakan algoritma *neural network backpropagation, support vector machine, k-nearest neighbors, naive bayes* dan *decission tree,* pada penelitian yang dilakukan didapat nilai akurasi tertinggi oleh algortima *decission tree* sebesar 99,80%, nilai kappa yakni 0.996, nilai RMSE terkecil dan di bawah 0.1 yakni 0.039, serta waktu yang dibutuhkan hanya 0.8 detik. Namun pada penelitian ini tidak hanya melakukan traning data saja tanpa menguji atau mengevaluasi dengan dataset yang baru. hal ini menjadi kelemahan dari penelitian yang dilakukan untuk menguji nilai akurasi berdasarkan training data yang dimiliki tidak menguji dan memvalidasi dataset yang terbaru diinputkan juga tidak terdapat *deploy* dari hasil *training* dan tidak adanya pembagian antara data *training* dan data *testing*. Oleh karena itu pada penelitian ini, peneliti memanfaatkan *streamlit* guna menguji dan memvalidasi data hasil *training* dan *testing* data baru untuk melihat hasil klasifikasi dari data yang diinputkan berdasarkan variabel *dependent* atau target level dari Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta.

Dalam mengklasifikasikan tingkat pencemaran udara menggunakan algoritma tipe regression yaitu Logistic Regression (Regresi Logistik) dan Random Forest. Evaluasi menggunakan nilai Confusion Matrix Accuracy dan evaluation regression mulai dari Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared. Melakukan data split data training dan data testing, terakhir untuk deployment sistem menjadi dashboard menggunakan library python Streamlit. menguji proses modelling dengan algortima dapat mengklasifikasikan data yang baru diinputkan. Data inputan tersebut terdiri dari nilai karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan partikel debu (PM_{10}) yang menjadi variabel independent. Tujuan penelitian sendiri guna mengetahui kaakuratan algortima regression dalam mengklasifikasikan data tingkat pencemaran udara di DKI Jakarta dengan algoritma regression yaitu Logistic Regression (Regresi Logistik) dan Random Forests.

Rumusan Masalah

Adapun rumusan permasalahan pada penelitian ini sebagai berikut:

1. Bagaimana membuktikan bahwa algoritma klasifikasi pada *supervised learning* dapat mengklasifikasikan Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta?
2. Bagaimana membandingkan kinerja dalam algoritma klasifikasi yaitu algoritma logistic regression dan algoritma random forest dalam mengklasfikasikan tingkat pencemaran udara di DKI Jakarta ?

Tujuan dan Manfaat

Tujuan

Dalam penelitian ini dilakukan memiliki beberapa tujuan sebagai berikut :

1. Mengetahui keakuratan algoritma klasifikasi dalam mengklasifikasikan data tingkat pencemaran udara di DKI Jakarta.
2. Menganalisis perbandingan kinerja algortima regression yaitu *logistic regression dan random forest*

Manfaat

Adapun manfaat yang dihaapkan dalam penelitian ini yaitu sebagai berikut :

1. Bagi Penulis

   Mendapatkan pengalaman atas ilmu yang diperoleh selama pendidikan yang diterapkan secara langsung pada kasus pencemaran udara
2. Bagi Pemerintah

   Pada hasil penelitian ini dapat dimanfaatkan oleh pemerintah dalam memberikan monitoring mengenai indeks pencemaran udara di suatu wilayah DKI Jakarta terkini, sehingga menjadi referensi dalam menyelesaikan dalam menghadapi permasalahan pencemaran udara diwilayah terkait
3. Bagi Masyarakat

Penelitian ini menjadikan masyarakat untuk sadar atas isu-isu akibat pencemaran udara dilingkungan sekitar serta mampu mempersiapkan penanggulangan akibat dari pencemaran udara.

Batasan Masalah

Berdasarkan rumusan masalah diatas, peneliti membatasi permasalahan dalam penelitian yang dilakukan yakni sebagai berikut :

1. Klasifikasi tingkat pencemaran udara sesuai indeks standar pencemar udara (ISPU) menggunakan algortima *supervised learning* tipe regression yaitu *logistic regression dan random forest*
2. Bahasa Pemrograman pada proses pengerjaan menggunakan *Python*
3. Untuk menguji validasi dari hasil data *training* dan data *testing* dideploy menggunakan *Streamlit library python.*
4. *Dataset* yang digunakan yaitu data Indeks Standar Pencemar Udara (ISPU) di DKI Jakarta yang diperoleh dari *website* resmi layanan Portal Data Tepadu Pemerintah Provinsi DKI Jakarta (data.jakarta.go.id) dari tahun 2017 - 2021
5. Pembagian data *training* dan data *testing* yang akan diterapkan pada model sebagai uji coba yaitu menggunakan perbandingan 80%:20%, 60%:40% secara sistematis
6. Dalam proses evaluasi menggunakan nilai *confusion matrix* dan metode *regression.* Confusion matrix dapat menghitung beberapa matrik lainnya seperti *accuracy*, *recall, specificity, precision, false positive rate, false negative rate, f-1 score*. Metode *regression* ini digunakan model evaluasi *matrix Mean Absolute Error (MAE)*, *Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared.*
7. Tampilan *output* dari proses *deploy* memiliki fitur input indeks karbon monoksida (CO), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), Ozon permukaan (O_3), dan partikel debu (PM_{10}). Kemudian jenis Algoritma Klasifikasi serta hasil labeling berupa kategori baik, sedang, tidak sehat, sangat tidak sehat, dan berbahaya.
8. Inputan yang dihasilkan hanya dapat menghasilkan satu kali *output.*