Final Fortex 5.0
National Data Science Competition

Model Prediksi Jumlah Limbah Berbahaya Pada Sepuluh Negara Bagian Terpadat Di Amerika Serikat

Presented By Statmat Team
Universitas Indonesia

Statmat Team

UNIVERSITAS INDONESIA

Siti Nur Salamah

Maryesta Apriliani Sihombing

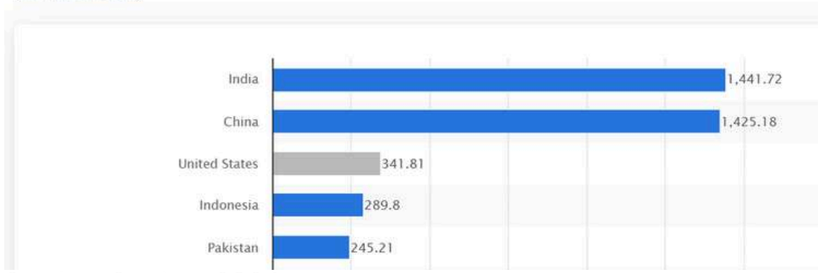
Raissa Anggia Maharani

Daftar Isi

PENDAHULUAN

1. Latar Belakang
2. Rumusan Masalah
3. Tujuan dan Manfaat

Latar Belakang



Sumber: Statista, 2024.

Amerika Serikat merupakan negara dengan jumlah penduduk ke-3 terbanyak di dunia.

Manufacturing Industry Outlook - United States

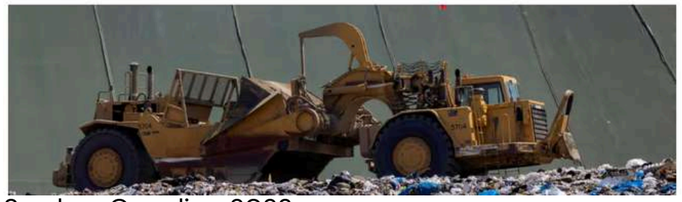
The United States of America holds the second-largest manufacturing industry in the world after China. The manufacturing value-added output of the United States of America was recorded to be at an all-time high of $2.89 trillion in Q4 of 2023. The industry output

Sumber: Yahoo Finance, 2024.

Amerika Serikat menduduki negara ke-2 dengan industri manufaktur terbesar di dunia setelah Cina.

US industry disposed of at least 60m pounds of PFAS waste in last five years

Estimate in new EPA analysis is probably 'dramatic' undercount because 'forever chemical' waste is unregulated in US



Sumber: Guardian, 2023.

Industri di Amerika Serikat menghasilkan limbah berbahaya dalam skala besar yang mencemari lingkungan.

Health and Ecological Hazards Caused by Hazardous Substances

Emergency response efforts must consider the health and ecological hazards of a hazardous substance release. These hazards impact emergency responders and effected communities. In some cases, hazardous substances may irritate the skin or eyes, make it difficult to breathe, cause headaches and nausea, or result in other types of illness. Some hazardous substances can cause far more severe health effects, including:

- behavioral abnormalities,
- cancer,
- genetic mutations,
- physiological malfunctions (e.g., reproductive impairment, kidney failure, etc.),
- physical deformations, and
- birth defects.

Sumber: EPA, 2024.

Penyakit yang bermunculan dapat disebabkan oleh limbah berbahaya.

Rumusan Masalah

1. Bagaimana model terbaik untuk memprediksi total limbah berbahaya berdasarkan data EPA pada sepuluh negara bagian terpadat di Amerika Serikat?

2. Apa faktor yang paling signifikan dalam mempengaruhi total limbah berbahaya pada sepuluh negara bagian terpadat di Amerika Serikat?

3. Bagaimana model prediksi total limbah berbahaya dapat digunakan untuk menyusun simulasi kebijakan lingkungan yang efektif pada sepuluh negara bagian terpadat di Amerika Serikat?

Tujuan Penelitian

1. Menganalisis model *machine learning* terbaik untuk memprediksi total limbah berbahaya berdasarkan data EPA pada sepuluh negara bagian terpadat di Amerika Serikat.
2. Mengidentifikasi faktor-faktor yang paling signifikan dalam mempengaruhi total limbah berbahaya pada sepuluh negara bagian terpadat di Amerika Serikat.
3. Menggunakan model prediksi total limbah berbahaya untuk menyusun simulasi kebijakan lingkungan yang efektif pada sepuluh negara bagian terpadat di Amerika Serikat.

Manfaat Penelitian

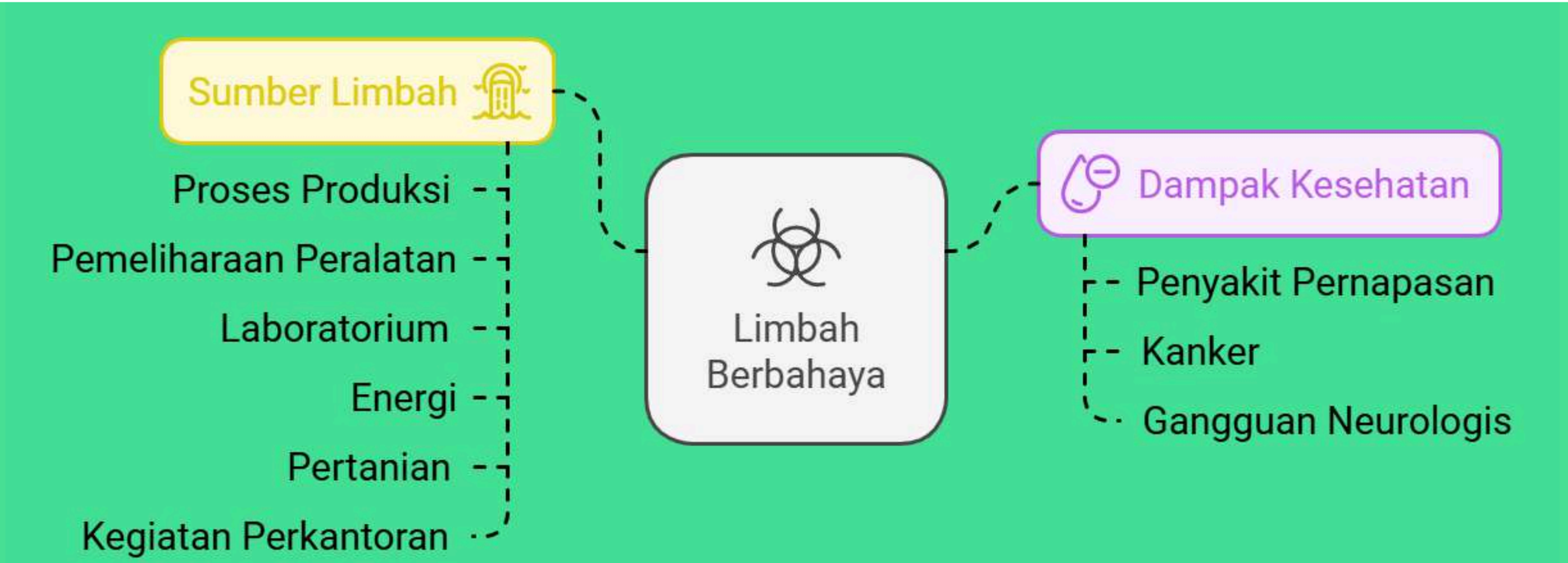
1. Menyediakan referensi mengenai model prediksi terbaik untuk memprediksi total limbah berbahaya berdasarkan data pada sepuluh negara bagian terpadat di Amerika Serikat yang dapat dimanfaatkan oleh peneliti maupun praktisi.
2. Mengidentifikasi faktor-faktor signifikan yang mempengaruhi total limbah berbahaya pada sepuluh negara bagian terpadat di Amerika Serikat yang dapat digunakan sebagai dasar dalam merancang kebijakan lingkungan dengan efektif.
3. Menghasilkan simulasi kebijakan lingkungan yang efektif berdasarkan hasil prediksi total limbah berbahaya yang dapat digunakan untuk mendukung pengambilan keputusan dalam upaya meningkatkan kesehatan masyarakat.

PEMBAHASAN

1. Landasan Teori
2. Metode Penelitian
3. Hasil dan Pembahasan

Limbah Berbahaya

Limbah berbahaya didefinisikan oleh Environmental Protection Agency (EPA) sebagai **limbah yang menimbulkan ancaman substansial atau potensial bagi kesehatan publik atau lingkungan**. Hal ini termasuk limbah yang bersifat toksik, korosif, mudah terbakar, atau reaktif.



Model *Machine Learning*

Pada penelitian ini, dipilih 4 model *machine learning* dengan bentuk regresi karena variabel prediktornya berupa numerik.

Regresi Linear

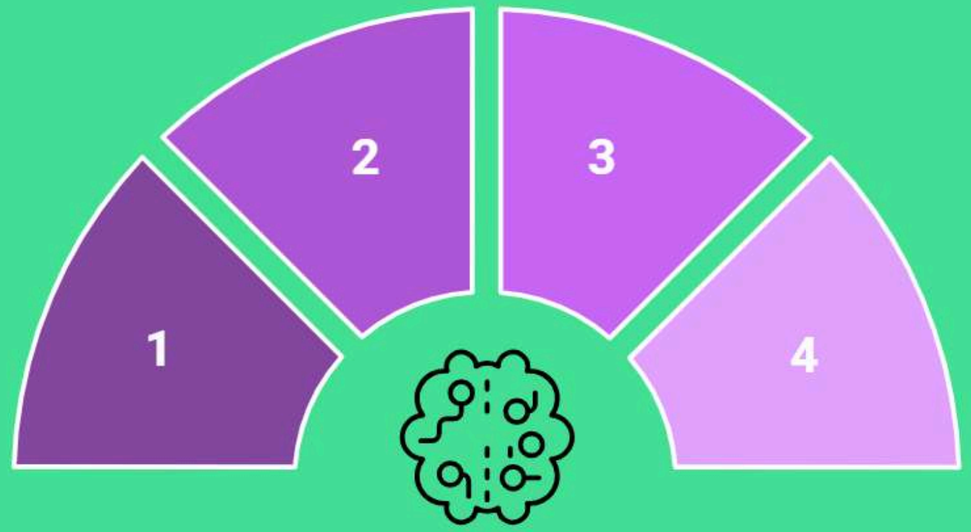
Metode dasar untuk memodelkan hubungan linier antara variabel (Kumari & Yadav, 2018).

Regresi Ridge

Mengatasi multikolinearitas dengan menerapkan penalti pada koefisien (Rajan, 2022).

Lasso

Mengurangi koefisien hingga nol untuk seleksi fitur dan menghindari overfitting (Melkumova & Shatskikh, 2017).



ElasticNet

Menggunakan regularisasi ganda untuk prediktor berkorelasi, cocok untuk data tinggi (Zou & Hastie, 2005).

Metrik *Error*

Digunakan 5 metrik error untuk mengevaluasi model dari berbagai sudut pandang, memastikan kinerja prediksi yang akurat, konsisten, dan andal secara keseluruhan.

Metrik	Definisi	Keunggulan
MAE	Rata-rata nilai absolut selisih antara hasil prediksi dan nilai sebenarnya.	Mudah diinterpretasikan dan tidak dipengaruhi dengan nilai outlier
MSE	Rata-rata kuadrat dari perbedaan antara prediksi dan aktual.	Menekankan kesalahan yang lebih besar terhadap nilai outlier.
RMSE	Akar kuadrat dari MSE untuk menyajikan error dalam satuan aslinya.	Memberikan gambaran yang lebih intuitif tentang besarnya kesalahan yang relatif terhadap skala data.
MAPE	Rata-rata error absolut dalam bentuk persentase relatif terhadap nilai sebenarnya.	Skala independen sehingga cocok untuk data dengan skala yang berbeda.
R-squared	Proporsi variasi data yang dijelaskan oleh model, dengan interval nilai 0-1.	Menjelaskan proporsi varians, mudah diinterpretasikan, tidak bergantung terhadap skala data

Metode Penelitian CRISP-DM

01 Business Understanding

02 Data Understanding

03 Data Preparation

04 Modeling

05 Evaluation

06 Deployment

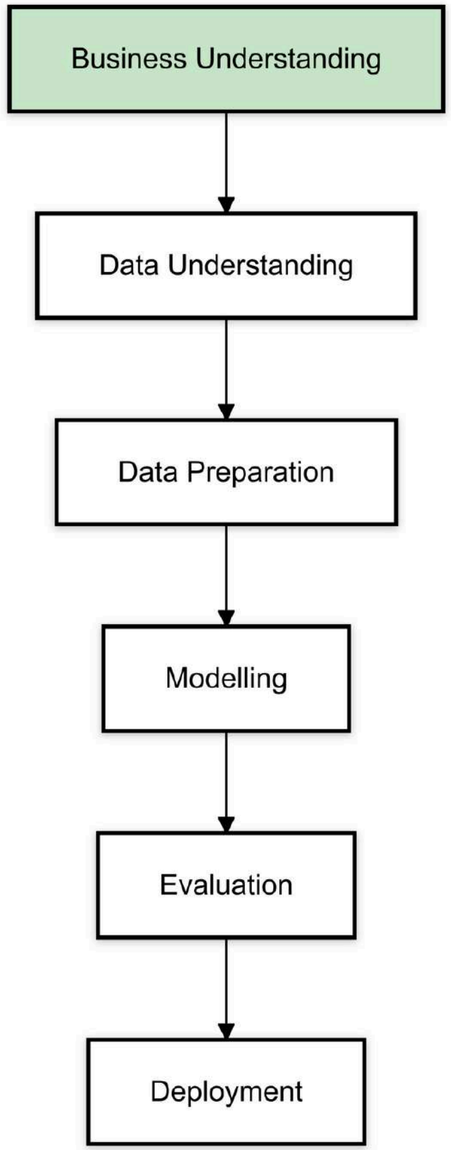
Business Understanding

Penelitian ini berfokus pada **prediksi jumlah limbah berbahaya** berdasarkan data 10 negara bagian terpadat di Amerika Serikat.

Variabel Target

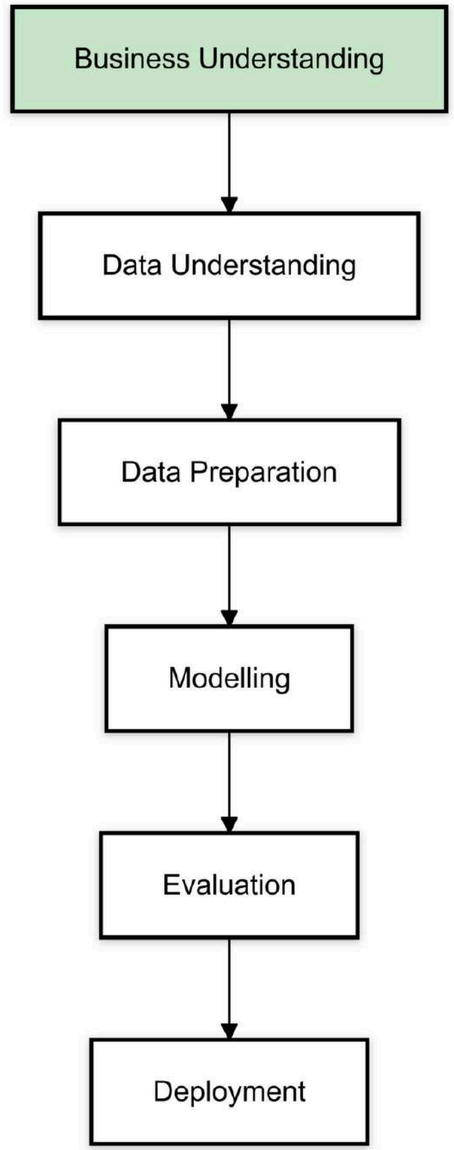
Total Releases

Jumlah total bahan kimia beracun yang dilepaskan baik secara on-site (di lokasi fasilitas) maupun off-site (dipindahkan ke lokasi lain untuk pembuangan)



Business Understanding

Urgensi Prediksi Total Releases



Data Understanding

Sumber Data

EPA pada tahun 2023

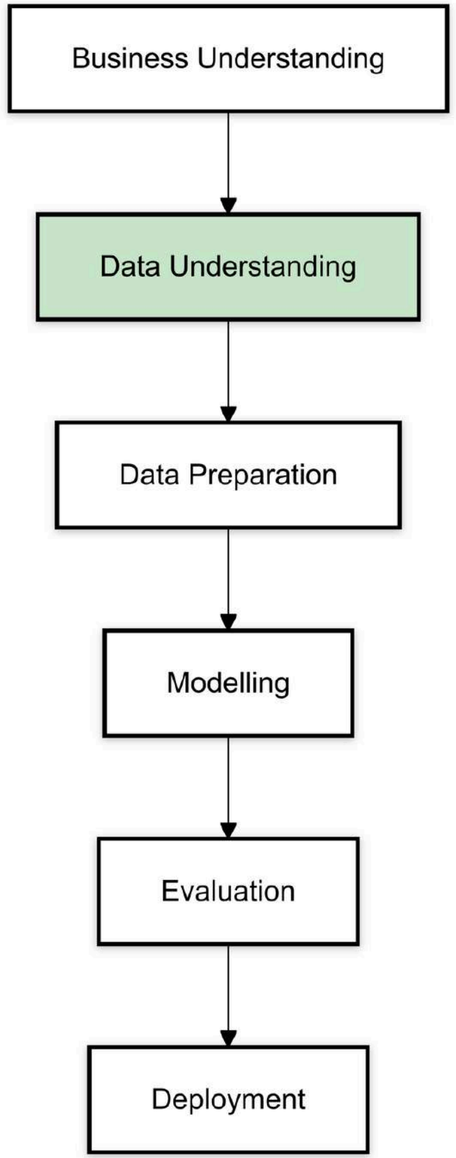
Informasi Data

- 77964 baris
- 122 kolom

Penelitian ini berfokus pada **10 negara bagian terpadat di Amerika Serikat berdasarkan data sensus Amerika Serikat**, yaitu:

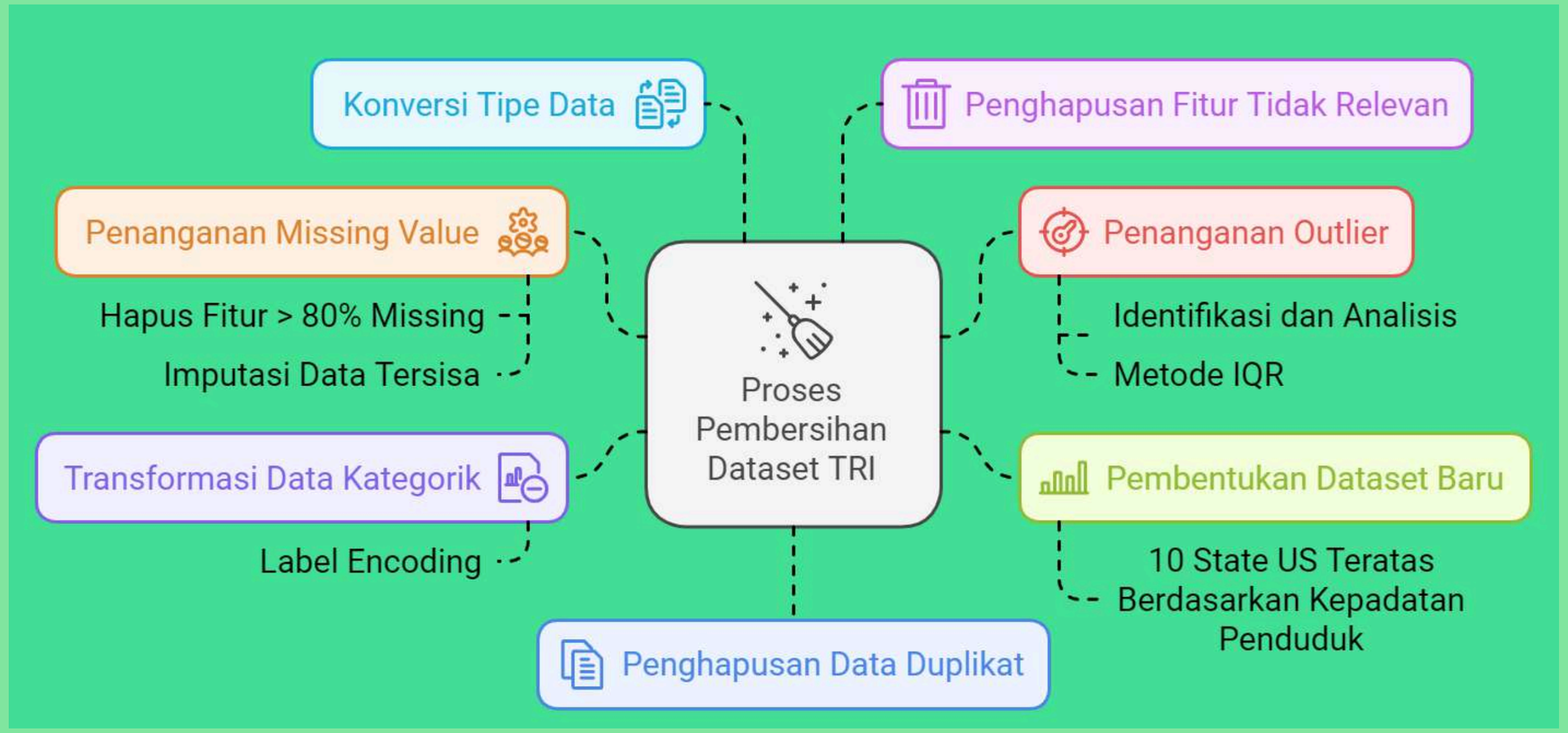
1. District of Columbia
2. New Jersey
3. Rhode Island
4. Massachusetts
5. Connecticut
6. Maryland
7. Delaware
8. Florida
9. New York
10. Pennsylvania

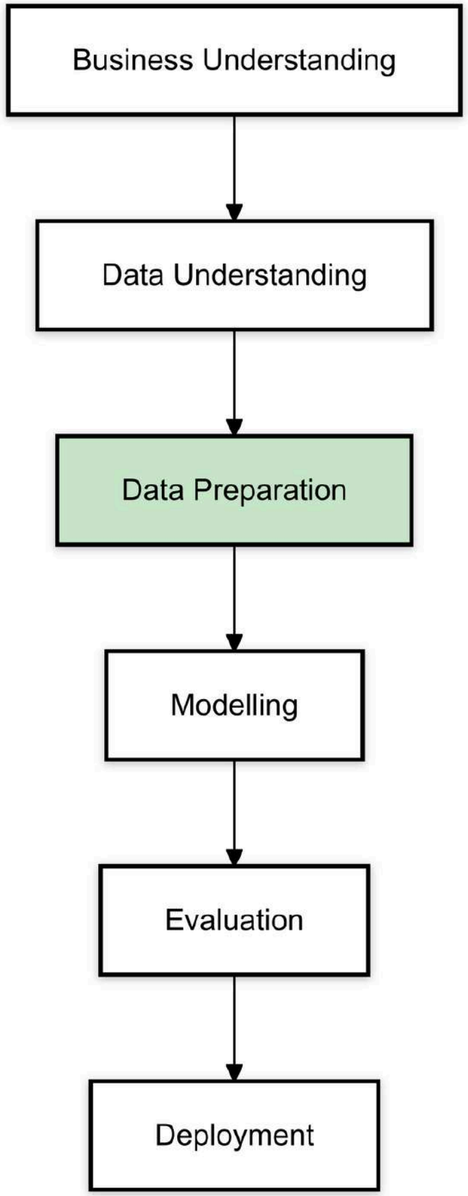
Dengan data **10 negara bagian terpadat** di Amerika Serikat, didapatkan **datanya mencakup 9943 baris dan 122 kolom.**



Data Preparation

Diagram Alur Proses *Preprocessing*





Data Preparation

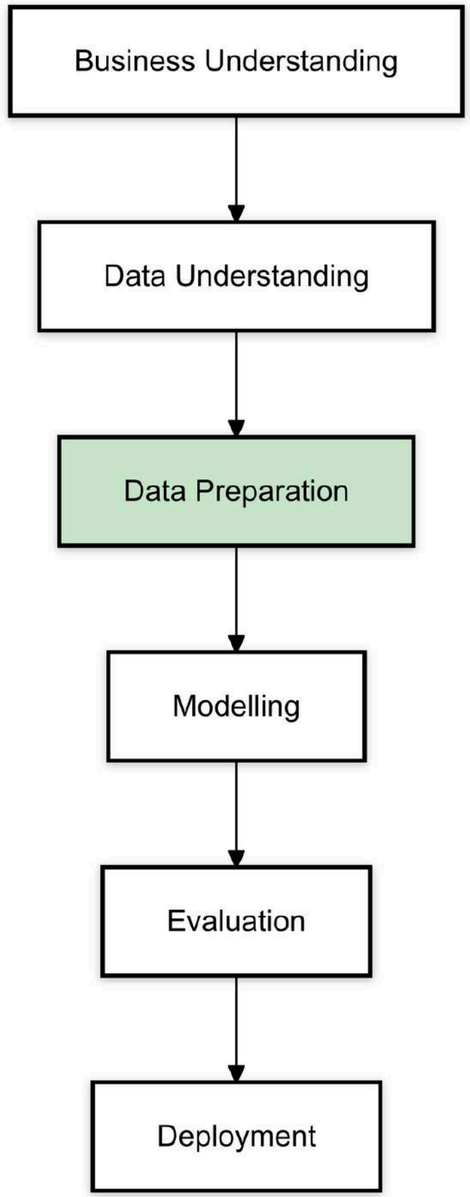
Konversi Tipe Data

```
1. YEAR: int64
2. TRIFD: object
3. FRS ID: float64
4. FACILITY NAME: object
5. STREET ADDRESS: object
6. CITY: object
7. COUNTY: object
8. ST: object
9. ZIP: int64
10. BIA: float64
11. TRIBE: object
12. LATITUDE: float64
13. LONGITUDE: float64
14. HORIZONTAL DATUM: object
15. PARENT CO NAME: object
16. PARENT CO DB NUM: object
17. STANDARD PARENT CO NAME: object
18. FOREIGN PARENT CO NAME: object
19. FOREIGN PARENT CO DB NUM: object
20. STANDARD FOREIGN PARENT CO NAME: object
21. FEDERAL FACILITY: object
22. INDUSTRY SECTOR CODE: int64
23. INDUSTRY SECTOR: object
24. PRIMARY SIC: float64
25. SIC 2: float64
26. SIC 3: float64
27. SIC 4: float64
28. SIC 5: float64
29. SIC 6: float64
30. PRIMARY NAICS: int64
31. NAICS 2: float64
32. NAICS 3: float64
33. NAICS 4: float64
34. NAICS 5: float64
35. NAICS 6: float64
36. DOC_CTRL_NUM: int64
37. CHEMICAL: object
38. ELEMENTAL METAL INCLUDED: object
39. TRI CHEMICAL/COMPOUND ID: object
40. CAS#: object
41. SRS ID: float64
```

Kolom Konversi

Ada **19 fitur** yang diubah menjadi tipe data object

1. YEAR, 3. FRS ID, 9. ZIP, 10. BIA, 22. INDUSTRY SECTOR CODE, 24. PRIMARY SIC, 25. SIC 2, 26. SIC 3, 27. SIC 4, 28. SIC 5, 29. SIC 6, 30. PRIMARY NAICS, 31. NAICS 2, 32. NAICS 3, 33. NAICS 4, 34. NAICS 5, 35. NAICS 6, 36. DOC_CTRL_NUM, dan 41. SRS ID.



Data Preparation

Drop Fitur

1 Fitur nilai unik = 1

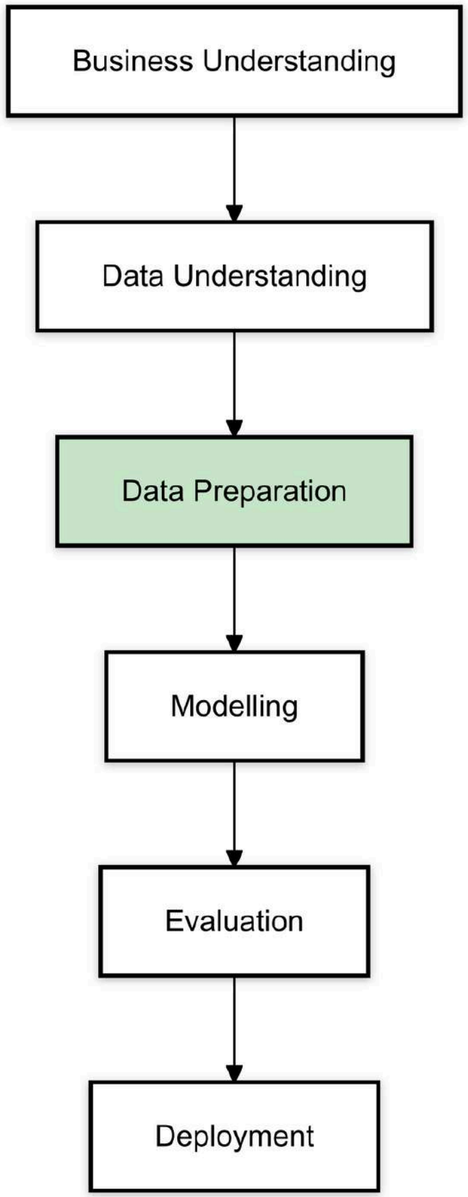
Ada 11 fitur dengan nilai unik = 1

	Kolom	Nilai Unik
0	1. YEAR	2023
1	54. 5.4 - UNDERGROUND	0
2	57. 5.5.1 - LANDFILLS	0
3	61. 5.5.3 - SURFACE IMPNDMNT	0
4	72. 6.2 - M40 METAL	0
5	73. 6.2 - M61 METAL	0
6	74. 6.2 - M71	0
7	77. 6.2 - M72	0
8	78. 6.2 - M63	0
9	105. 6.2 - UNCLASSIFIED	0
10	108. 8.1 - RELEASES	0

2 Fitur Tidak Relevan

Ada 9 fitur tidak relevan dengan target

No	Kolom	Alasan Penghapusan
1	3. FRS ID	ID kurang relevan
2	4. FACILITY NAME	Nama fasilitas (deskriptif, tidak relevan untuk analisis).
3	5. STREET ADDRESS	Alamat fasilitas (deskriptif).
4	36. DOC_CTRL_NUM	ID dokumen (unik, tidak relevan).
5	39. TRI CHEMICAL/COMPOUND ID, 40. CAS#, 41. SRS ID	Ketiganya mirip, salah satu bisa dihapus.
6	15. PARENT CO NAME, 16. PARENT CO DB NUM, 17. STANDARD PARENT CO NAME	Nama perusahaan induk (deskriptif, tidak relevan untuk analisis).
7	22. INDUSTRY SECTOR CODE	Rendundan dengan 23. INDUSTRY SECTOR
8	65. ON SITE RELEASE TOTAL	Fitur yang menghasilkan variabel prediktor '107. TOTAL RELEASES'
9	88. OFF SITE RELEASE TOTAL	Fitur yang menghasilkan variabel prediktor '107. TOTAL RELEASES'



Data Preparation

Handling Missing Value

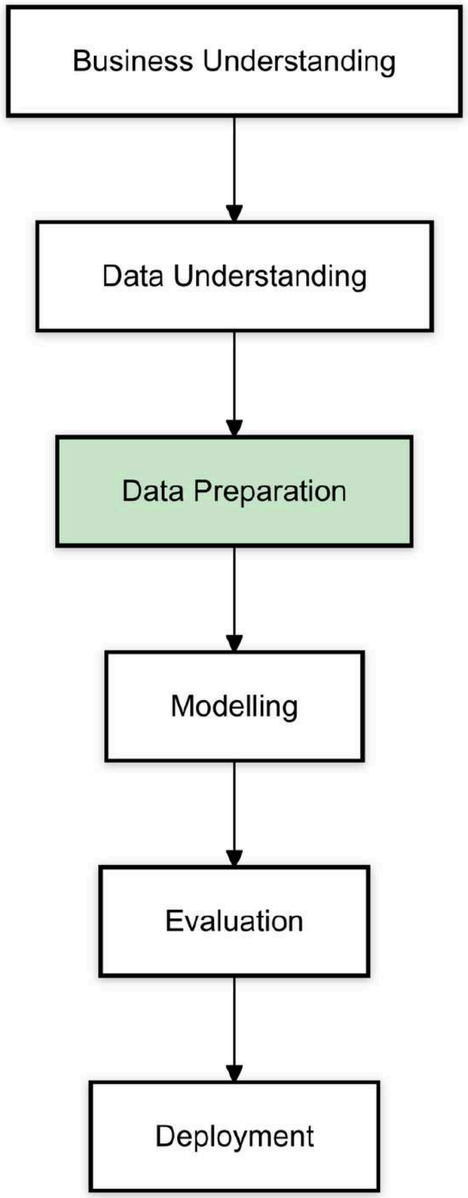
	Kolom	Jumlah Missing Values	Persentase Missing Values (%)
1	3. FRS ID	19	0.024370
8	10. BIA	77462	99.356113
9	11. TRIBE	77462	99.356113
12	14. HORIZONTAL DATUM	34	0.043610
13	15. PARENT CO NAME	15848	20.327331
14	16. PARENT CO DB NUM	24704	31.686419
15	17. STANDARD PARENT CO NAME	9997	12.822585
16	18. FOREIGN PARENT CO NAME	67764	86.917039
17	19. FOREIGN PARENT CO DB NUM	71575	91.805192
18	20. STANDARD FOREIGN PARENT CO NAME	63321	81.218255
22	24. PRIMARY SIC	77964	100.000000
23	25. SIC 2	77964	100.000000
24	26. SIC 3	77964	100.000000
25	27. SIC 4	77964	100.000000
26	28. SIC 5	77964	100.000000
27	29. SIC 6	77964	100.000000
29	31. NAICS 2	68853	88.313837
30	32. NAICS 3	75417	96.733108
31	33. NAICS 4	77129	98.928993
32	34. NAICS 5	77710	99.674209
33	35. NAICS 6	77902	99.920476
39	41. SRS ID	69	0.088502
108	120. 8.8 - ONE-TIME RELEASE	67171	86.156431
109	121. PROD_RATIO_OR_ACTIVITY	10358	13.285619
110	122. 8.9 - PRODUCTION RATIO	1594	2.044533

Hapus fitur dengan missing value di atas 80%

Hapus baris yang memiliki missing value kecil (< 5%)

Imputasi modus untuk fitur kategorik “PROD_RATIO_ACTIVITY„

Tersisa 76336 baris



Data Preparation

Duplikat Data, Filter Data, Outlier

Duplikat Data

```
Jumlah data duplikat: 13
```

Pada data ini, terdapat 13 duplikat yang akan dihapus karena dapat menyebabkan bias dalam analisis serta mempengaruhi performa model machine learning.

Filter Data

```
Jumlah data setelah filter: 9943
```

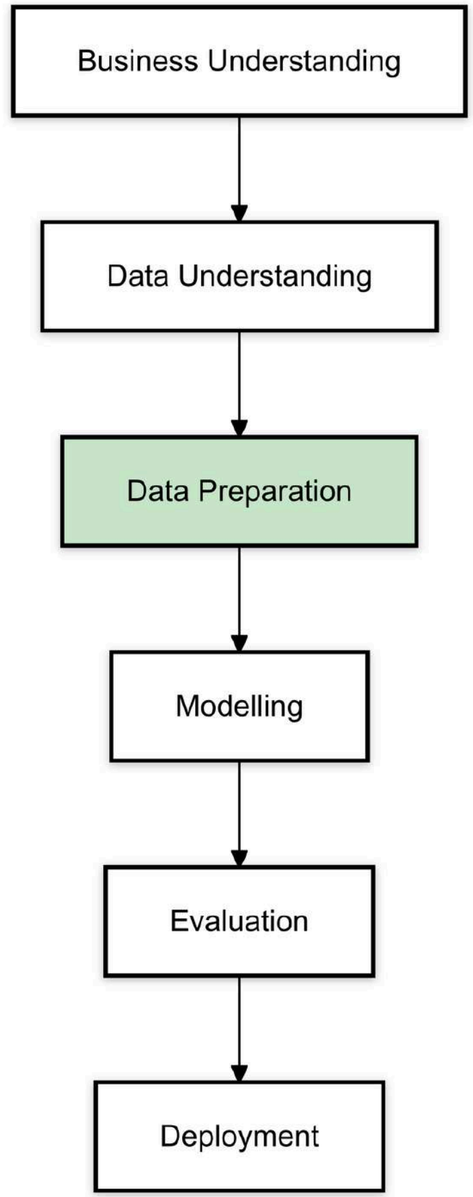
Fokus penelitian ini pada 10 negara bagian terpadat di AS, sehingga dataset difilter menjadi 9943 baris.

Outlier

Dipilih metode IQR (Interquartile Range) pada data ini untuk mendeteksi *outlier*.

	Kolom	Jumlah Outlier
0	12. LATITUDE	926
1	13. LONGITUDE	6536
2	66. 6.1 - POTW - TRNS RLSE	8925
3	67. 6.1 - POTW - TRNS TRT	3362
4	68. POTW - TOTAL TRANSFERS	9766
5	89. 6.2 - M20	1910
6	90. 6.2 - M24	8357
7	91. 6.2 - M26	4640
8	92. 6.2 - M28	15
9	93. 6.2 - M93	5073
10	94. OFF-SITE RECYCLED TOTAL	18316
11	95. 6.2 - M56	5739
12	96. 6.2 - M92	1992
13	97. OFF-SITE ENERGY RECOVERY T	7164
14	98. 6.2 - M40 NON-METAL	637
15	99. 6.2 - M50	5698
16	100. 6.2 - M54	1281
17	101. 6.2 - M61 NON-METAL	1317
18	102. 6.2 - M69	1022
19	103. 6.2 - M95	616
20	104. OFF-SITE TREATED TOTAL	11543
21	106. 6.2 - TOTAL TRANSFER	14578
22	107. TOTAL RELEASES	13844
23	109. 8.1A - ON-SITE CONTAINED	3689
24	110. 8.1B - ON-SITE OTHER	14505
25	111. 8.1C - OFF-SITE CONTAIN	16291
26	112. 8.1D - OFF-SITE OTHER R	17326
27	113. 8.2 - ENERGY RECOVER ON	2144
28	114. 8.3 - ENERGY RECOVER OF	7164
29	115. 8.4 - RECYCLING ON SITE	3804
30	116. 8.5 - RECYCLING OFF SIT	18301
31	117. 8.6 - TREATMENT ON SITE	16500
32	118. 8.7 - TREATMENT OFF SITE	11507
33	119. PRODUCTION WSTE (8.1-8.7)	11788
34	122. 8.9 - PRODUCTION RATIO	14762

Outlier di setiap fitur dibiarkan karena dianggap bagian dari variasi alami yang berguna untuk analisis dan pemodelan.



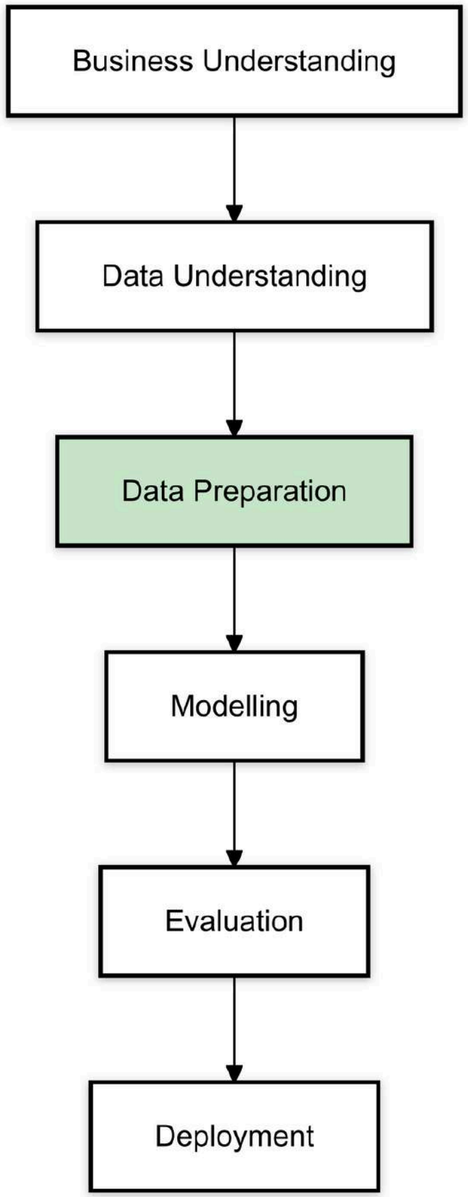
Data Preparation

Encoding Data Kategorik

```
Jumlah Nilai Unik dari '2. TRIFD': 3525
Jumlah Nilai Unik dari '6. CITY': 1321
Jumlah Nilai Unik dari '7. COUNTY': 208
Jumlah Nilai Unik dari '8. ST': 10
Jumlah Nilai Unik dari '9. ZIP': 1713
Jumlah Nilai Unik dari '14. HORIZONTAL DATUM': 1
Jumlah Nilai Unik dari '21. FEDERAL FACILITY': 2
Jumlah Nilai Unik dari '23. INDUSTRY SECTOR': 29
Jumlah Nilai Unik dari '30. PRIMARY NAICS': 312
Jumlah Nilai Unik dari '37. CHEMICAL': 270
Jumlah Nilai Unik dari '38. ELEMENTAL METAL INCLUDED': 2
Jumlah Nilai Unik dari '39. TRI CHEMICAL/COMPOUND ID': 256
Jumlah Nilai Unik dari '42. CLEAN AIR ACT CHEMICAL': 2
Jumlah Nilai Unik dari '43. CLASSIFICATION': 3
Jumlah Nilai Unik dari '44. METAL': 2
Jumlah Nilai Unik dari '45. METAL CATEGORY': 6
Jumlah Nilai Unik dari '46. CARCINOGEN': 2
Jumlah Nilai Unik dari '47. PBT': 2
Jumlah Nilai Unik dari '48. PFAS': 2
Jumlah Nilai Unik dari '49. FORM TYPE': 2
Jumlah Nilai Unik dari '50. UNIT OF MEASURE': 2
Jumlah Nilai Unik dari '121. PROD_RATIO_OR_ ACTIVITY': 2
```

22 kolom kategorik pada data ini akan di encoding menggunakan Label Encoding.

Label Encoding dipilih pada kasus ini karena **jumlah kategori pada setiap fitur sangat banyak**, sehingga metode ini lebih efisien dan praktis dibandingkan teknik lain seperti One-Hot Encoding.



Modelling

Tujuan

Prediksi total limbah (*Total Releases*) di 10 negara bagian terpadat AS.

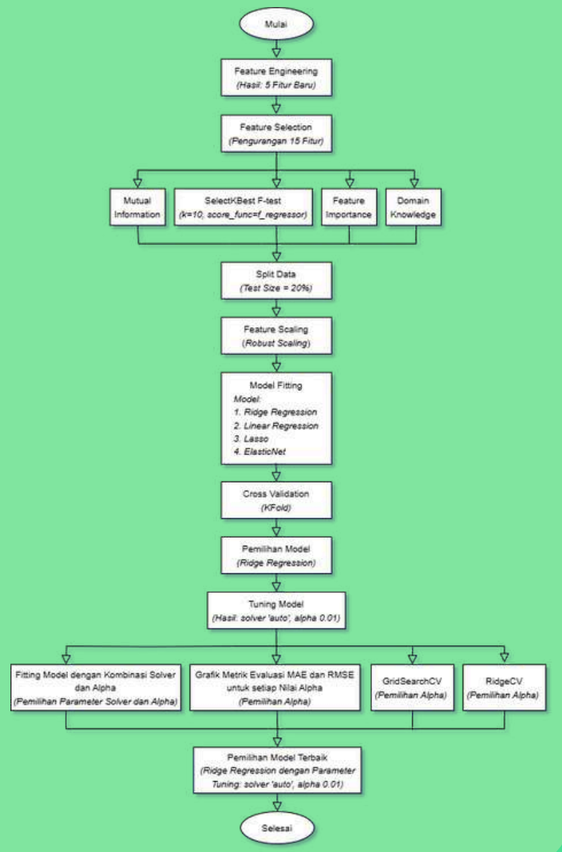
Jumlah Fitur Awal

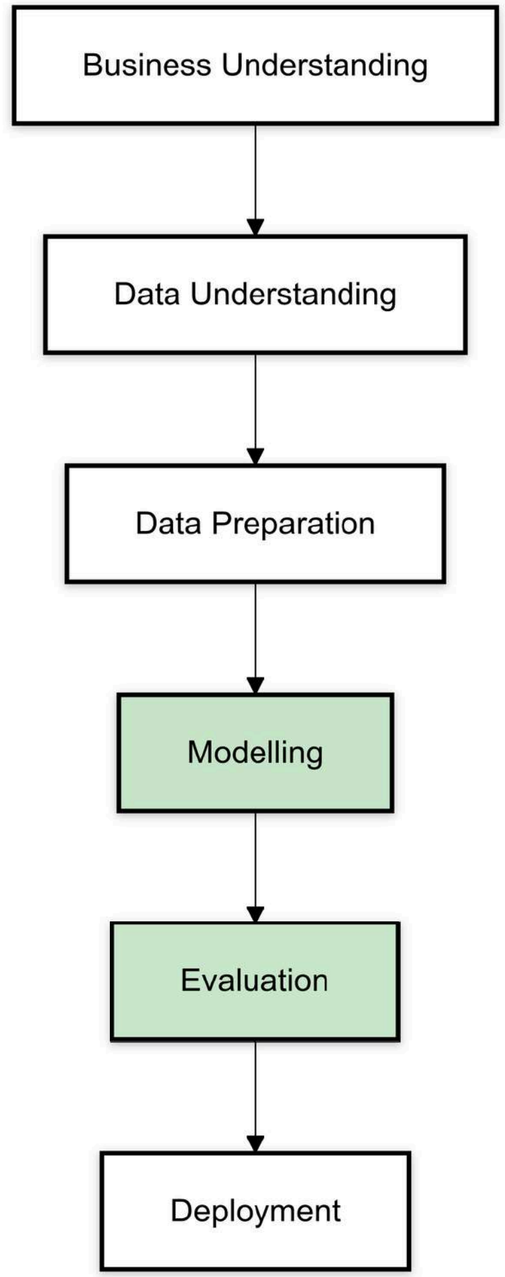
81 buah

Metrik Evaluasi

MAE, RMSE, MSE, MAPE, dan R^2

Diagram Alur Proses
Modelling & Evaluation





Modeling & Evaluation:

Feature Engineering

1 Geographical Features

2 Fitur Baru:

1. Region
2. Distance from Center

2 Indikator Risiko

1 Fitur Baru

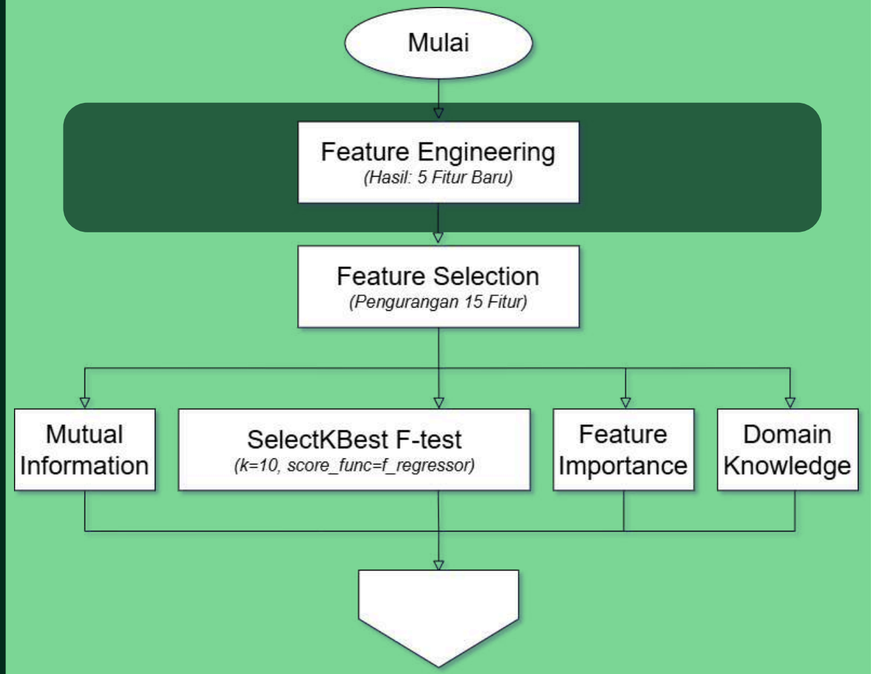
3 Efisiensi Daur Ulang (Recycling) dan Pengolahan (Treatment)

2 Fitur Baru:

1. Treatment Efficiency
2. Recycling Efficiency

Diagram Alur Proses

Modeling & Evaluation



Modeling & Evaluation:

Feature Engineering

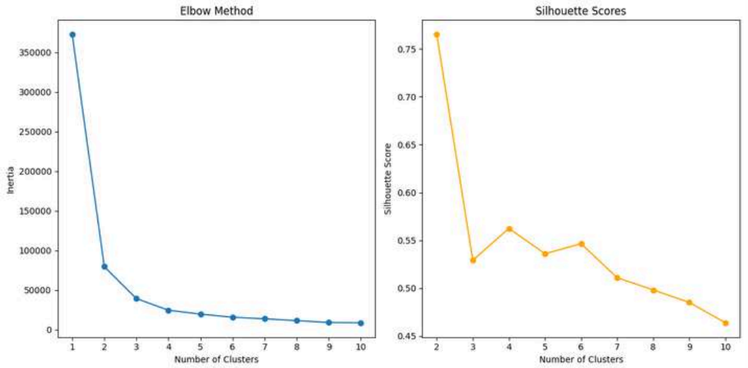
1 Geographical Features

1. Region

Tujuan: Mengidentifikasi pola geografis.

Ide: Mengelompokan data dengan K-Means Clustering berdasarkan fitur Latitude dan Longitude

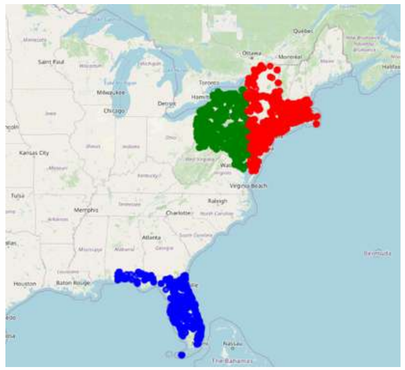
Step:



Identifikasi jumlah klaster dengan Elbow Method dan Silhoutte Scores. Dipilih k=3.

KMEANS CLUSTE-RING

Clustering dengan K-Means

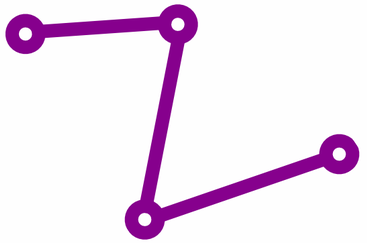


Hasil: Fitur Region yang terdiri dari 3 kategori

2. Distance from Center

Tujuan: Menangkap hubungan jarak

Ide: Menghitung jarak ke pusat geografis AS menggunakan Geodesic Distance dalam kilometer.



Modeling & Evaluation:

Feature Engineering

2 Indikator Risiko

Environmental Risk Score

Tujuan: Mengukur risiko lingkungan dari bahan kimia berbahaya.
Ide: Menggabungkan variabel CARCINOGEN, PBT, dan PFAS
Rumus:
Environmental_Risk_Score = CARCINOGEN + PBT + PFAS

3 Efisiensi Daur Ulang (Recycling) dan Pengolahan (Treatment)

1. Recycling Efficiency

Tujuan: Mengukur proporsi limbah yang berhasil didaur ulang.
Rumus:
Recycling Efficiency = (8.4 – RECYCLING ON SITE + 8.5 – RECYCLING OFF SITE) / PRODUCTION WSTE (8.1–8.7)

2. Treatment Efficiency

Tujuan: Mengukur proporsi limbah yang berhasil diolah sebelum dilepaskan.
Rumus:
Treatment Efficiency = (8.6 – TREATMENT ON SITE + 8.7 – TREATMENT OFF SITE) / PRODUCTION WSTE (8.1–8.7)

Modeling & Evaluation:

Feature Selection

3 Metode Mutual Information, Select K-Best: f_regressor, dan Feature Importance.

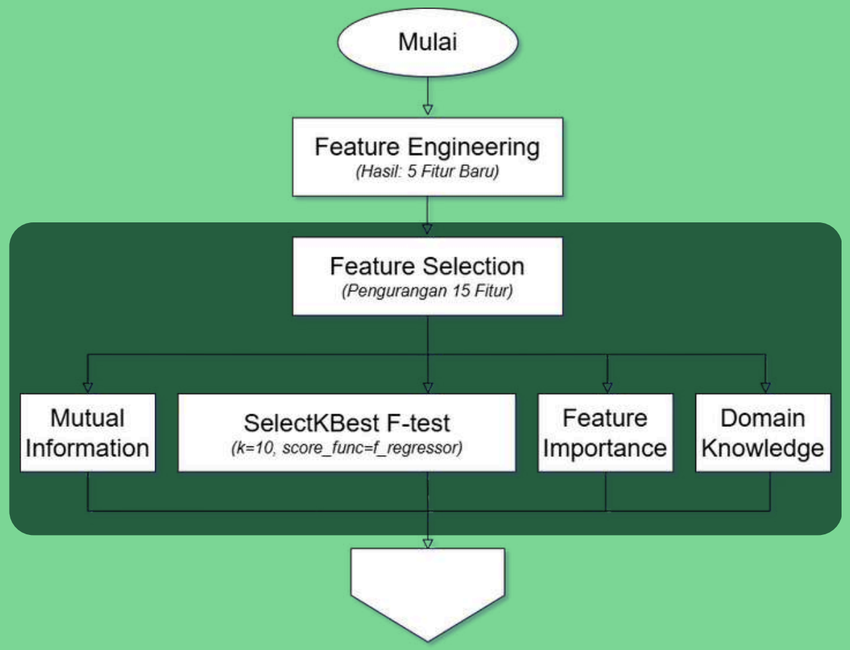
Pertimbangan

"Region", "Environmental_Risk_Score", "Recycling_Efficiency", "INDUSTRY SECTOR", dan "Treatment_Efficiency"	"PFAS", "LONGITUDE", dan "LATITUDE"	"POTW - TRNS TRT", "OFF-SITE TREATED TOTAL", "ENERGY RECOVER OF", "RECYCLING OFF SIT", "TREATMENT ON SITE", "RECYCLING ON SITE", dan "ENERGY RECOVER ON"
Dipertahankan. Memberikan informasi penting terkait lokasi, risiko lingkungan, efisiensi pengelolaan, dan konteks industri.	**Dipertahankan** Dipakai untuk feature engineering.	**Dipertahankan.** Memiliki pasangan terkait dalam fitur lainnya.

Penghapusan 15 Feature

Berdasarkan Domain Knowledge dan didukung nilai Feature Importance-nya yang negatif

Diagram Alur Proses *Modeling & Evaluation*



Modeling & Evaluation:

Model Selection

Sebelum Model Selection:

Variabel

- **Target:** Total Releases
- **Prediktor**: 71 (setelah feature selection)

Split Data

- **Split Data**: train (80%) dan test (20%)
- **Reproducibility**: random state = 42

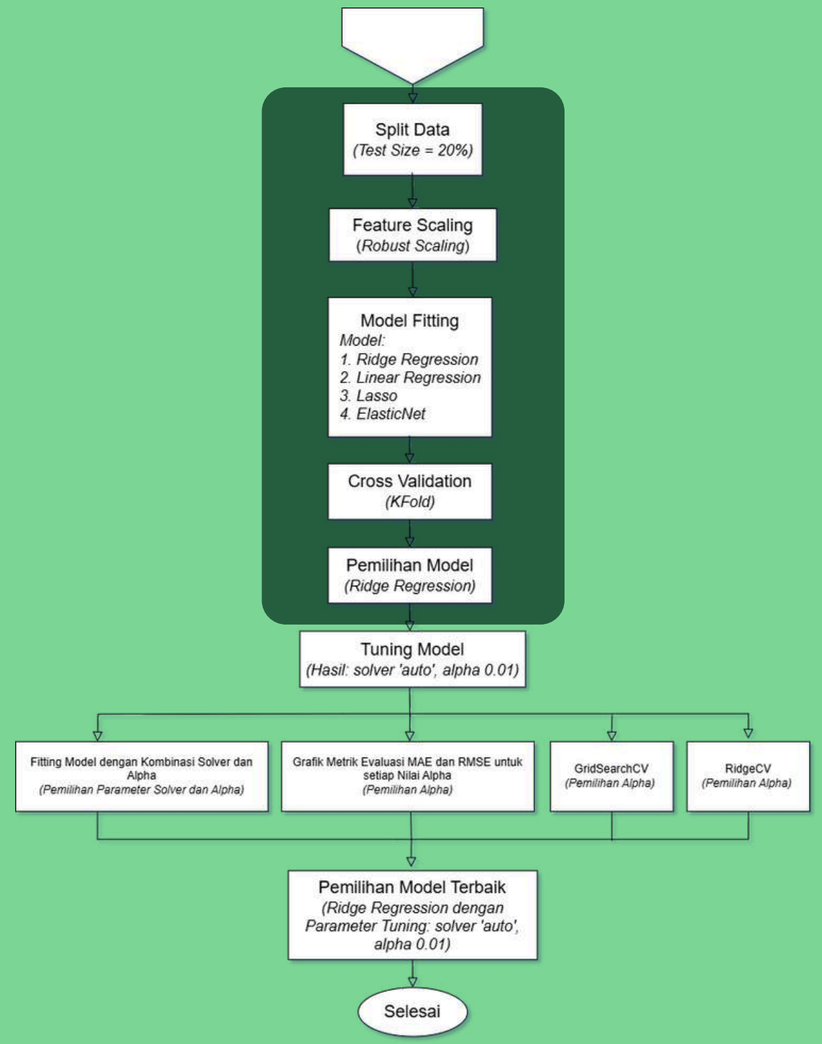
Scaling — Robust Scaler untuk mengurangi dampak pencilan.

Model Fitting: 4 Model

Ridge Regression, Linear Regression, Lasso, ElasticNet

Cross-Validation — **KFold (10 folds)** untuk memastikan stabilitas dan generalisasi model terhadap data baru.

Diagram Alur Proses *Modeling & Evaluation*



Modeling & Evaluation:

Model Selection

Cross-Validation **KFold (10 folds)** untuk memastikan stabilitas dan generalisasi model terhadap data baru.

Hasil CV:

	MSE	MSE_std	RMSE	RMSE_std	MAE	MAE_std	MAPE	MAPE_std	R^2	R^2_std
Ridge Regression	3025.699304	8958.376424	55.006357	51.257992	0.735157	1.898918	8.621096e+10	4.355550e+10	1.000000	9.585253e-07
LinearRegression	2983.870470	8951.611329	54.624816	51.820993	0.612733	1.837933	5.331431e+09	3.407804e+08	1.000000	9.570260e-07
Lasso	67763.056413	198078.910664	260.313381	235.957681	8.145157	10.529416	6.421832e+15	4.300754e+15	0.999992	2.108485e-05
ElasticNet	69686.274246	206975.919676	263.981579	245.165926	5.088067	10.584096	7.451824e+14	2.641659e+14	0.999992	2.205541e-05

Linear Regression unggul dengan MSE, RMSE, dan MAE terendah

Ridge Regression memiliki performa yang mendekati Linear Regression

Model Selection: Berdasarkan Hasil CV

Ridge Regression dipilih karena:

- Performanya sangat stabil dengan metrik yang sangat baik.
- Lebih tahan terhadap multicollinearity dibanding Linear Regression.
- Kinerja jauh lebih baik dibanding Lasso dan ElasticNet dalam hal akurasi dan stabilitas.

Ridge Regression menjadi **model pilihan** untuk prediksi.

Modeling & Evaluation:

Tuning Selected Model

Concern Parameter **Solver** dan **Alpha** pada Ridge Regression

4 Pendekatan

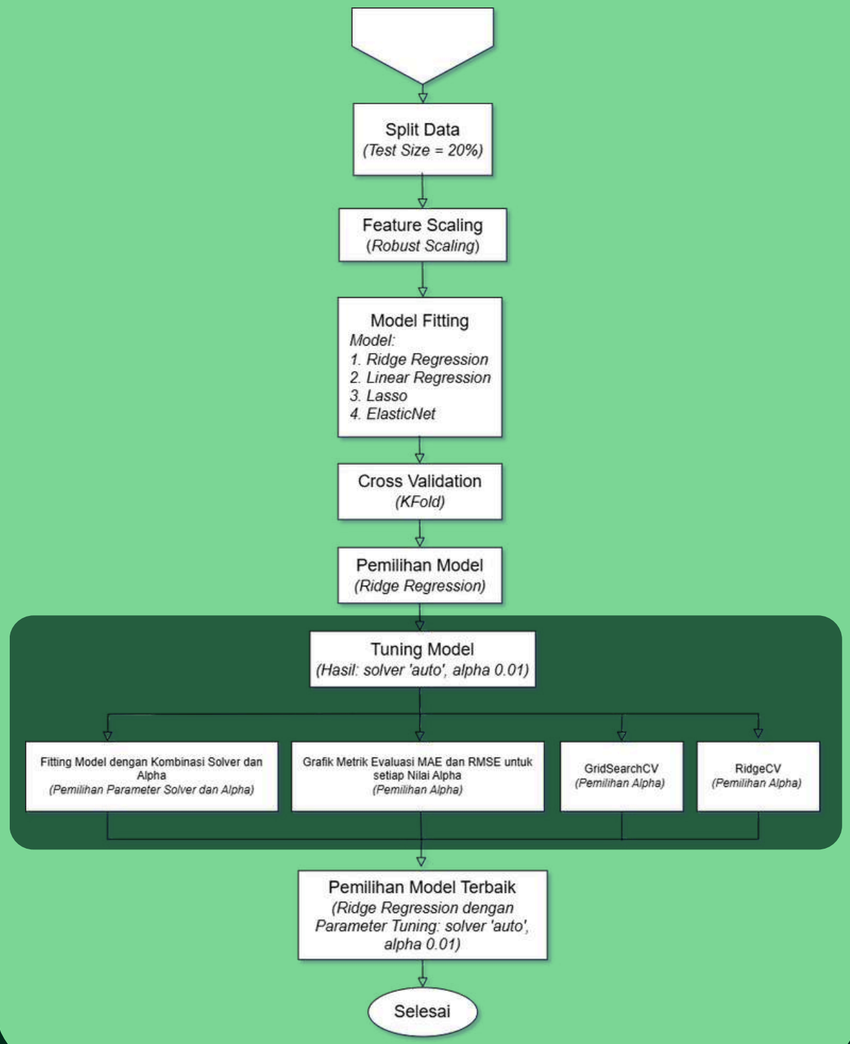
Pemilihan Parameter Solver

1. K-Fold Cross Validation dengan Kombinasi Solver dan Alpha

Pemilihan Parameter Alpha

1. Fitting Model dengan Kombinasi Solver dan Alpha
2. GridSearchCV
3. RidgeCV
4. Grafik Metrik Evaluasi MAE dan RMSE untuk setiap nilai Alpha

Diagram Alur Proses *Modeling & Evaluation*



Modeling & Evaluation:

Tuning Selected Model

Hasil 4 Pendekatan

1. K-Fold Cross Validation dengan Kombinasi Solver dan Alpha

Solver: ‘auto’
Performa konsisten dan optimal di seluruh metrik, termasuk MSE, RMSE, MAE, MAPE, dan R2.

Alpha
Semakin kecil nilai alpha, semakin baik performa model dalam hal akurasi prediksi

	Solver	Alpha	MSE	RMSE	MAE	MAPE	R^2
0	auto	0.01	2.999687e+03	18.726306	0.664412	1.399599e+10	1.000000
1	svd	0.01	2.999687e+03	18.723584	0.664310	7.192405e+09	1.000000
2	cholesky	0.01	2.999687e+03	18.726306	0.664412	1.399599e+10	1.000000
3	lsqr	0.01	2.884056e+08	15573.547035	3666.830194	2.252182e+18	0.946324
4	sparse_cg	0.01	1.305203e+08	5650.843846	1879.866000	1.973900e+18	0.975994
5	sag	0.01	7.950174e+09	56249.150151	8084.315781	4.121760e+18	0.769020
6	saga	0.01	1.487122e+10	72991.593603	9944.079162	5.056738e+18	0.685431
7	auto	0.10	3.008062e+03	19.452119	0.692677	2.660039e+10	1.000000
8	svd	0.10	3.008062e+03	19.449837	0.692594	1.761399e+10	1.000000
9	cholesky	0.10	3.008062e+03	19.452119	0.692677	2.660039e+10	1.000000
10	lsqr	0.10	2.884056e+08	15573.547035	3666.830194	2.252182e+18	0.946324
11	sparse_cg	0.10	1.304897e+08	5649.473845	1879.396155	1.973398e+18	0.975999
12	sag	0.10	7.967349e+09	56283.137766	8086.016180	4.122288e+18	0.768934
13	saga	0.10	1.486834e+10	72980.642042	9943.695142	5.057148e+18	0.685487
14	auto	1.00	3.025699e+03	19.957896	0.735157	8.621096e+10	1.000000
15	svd	1.00	3.025702e+03	19.960167	0.735240	8.199816e+10	1.000000
16	cholesky	1.00	3.025699e+03	19.957896	0.735157	8.621096e+10	1.000000

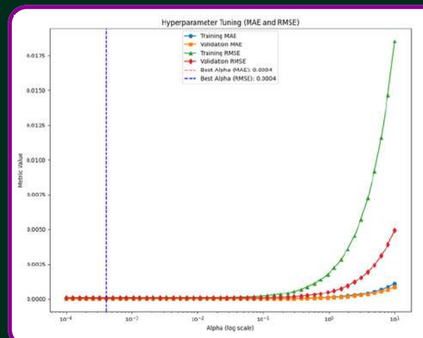
Dipilih Parameter **Solver = ‘auto’**

2. GridSearchCV

Alpha (based on MAE): 0.01
Alpha (based on RMSE): 0.01

3. RidgeCV

Alpha (based on MAE): 0.01
Alpha (based on RMSE): 0.01



4. Grafik Metrik Evaluasi MAE dan RMSE untuk setiap nilai Alpha

Alpha (based on MAE dan RMSE)
Semakin kecil nilai alpha, semakin baik performa model dalam hal akurasi prediksi

Dipilih Parameter **Alpha= 0.01**

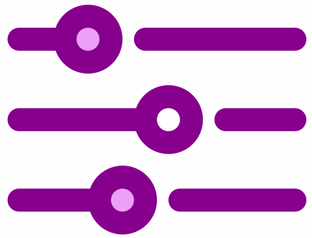
Jadi, dipilih model akhir **Ridge Regression** dengan Parameter **Solver = ‘auto’** dan **Alpha= 0.01**

Deployment

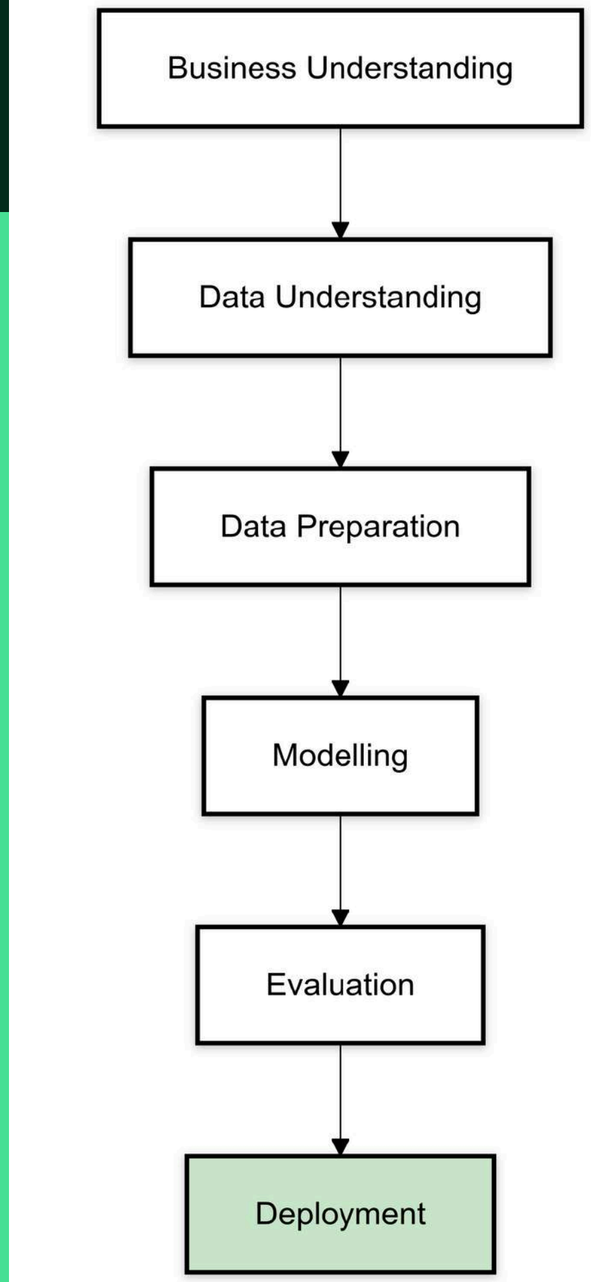
Model terbaik untuk 10 negara bagian terpadat di Amerika Serikat diterapkan dalam simulasi kebijakan menggunakan platform Dash.

Fitur Untuk Parameter Kebijakan

Pemilihan fitur didasarkan pada analisis feature importance, dengan 3 fitur paling berpengaruh yang dipilih.



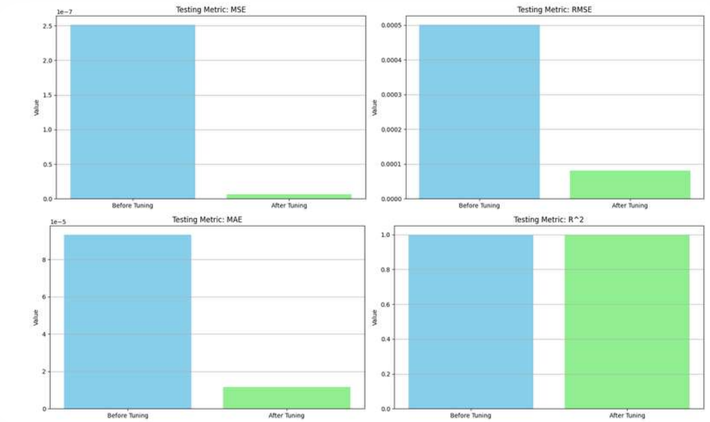
Parameter kebijakan dapat diatur melalui kontrol interaktif pada dashboard, memungkinkan pengguna untuk menguji berbagai skenario secara fleksibel.



Hasil dari Best Model Ridge Regression

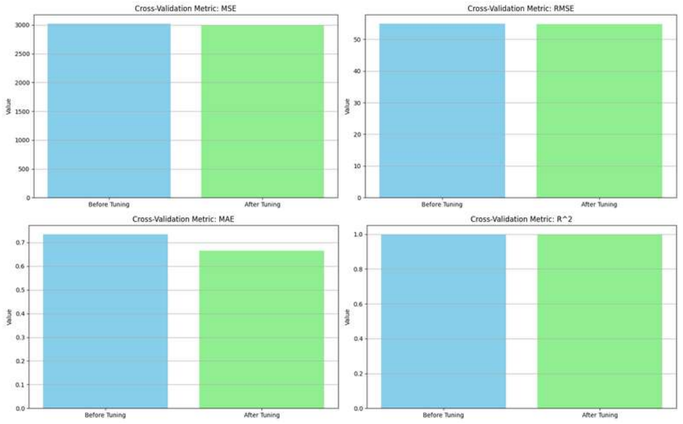
Berikut merupakan perbandingan Kinerja Ridge Regression sebelum dan sesudah Hyperparameter Tuning.

Testing Metrics



Metrik	Before Tuning	After Tuning
MSE	2.516371e-07	6.638939e-09
RMSE	5.016345e-04	8.147968e-05
MAE	9.338976e-05	1.162472e-05
R-Squared	1.000000e+00	1.000000e+00

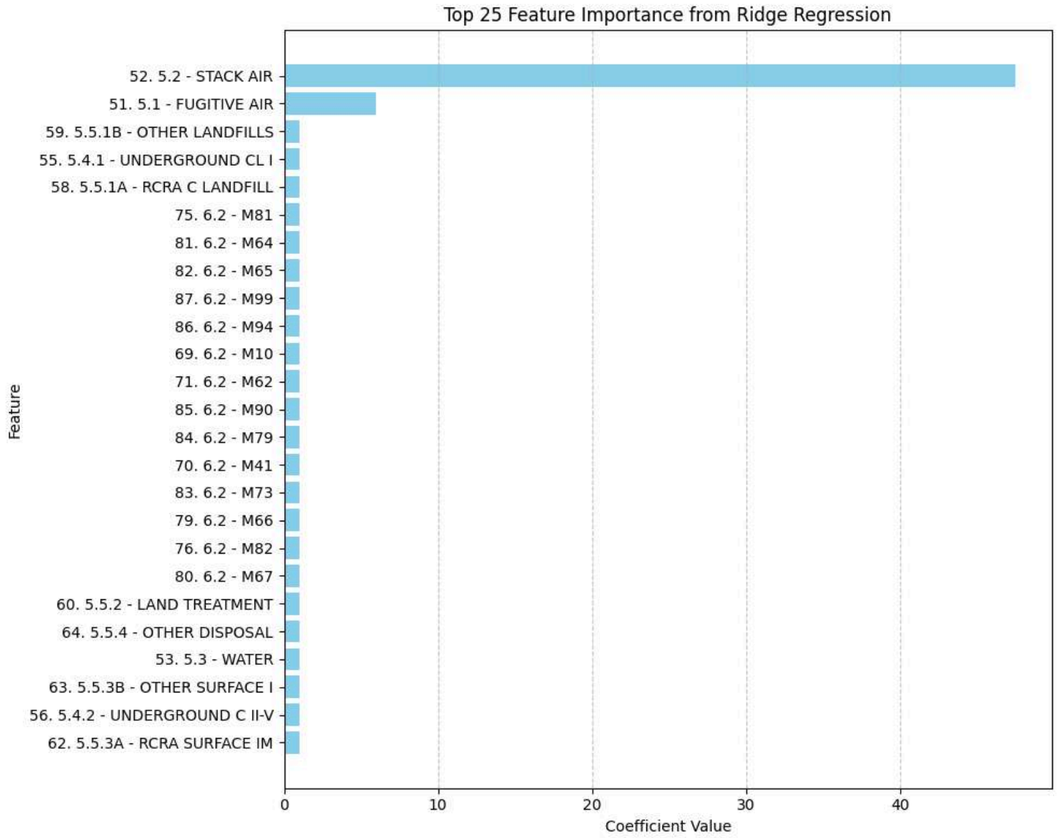
Cross-Validation Metrics



Metrik	Before Tuning	After Tuning
MSE	3025.699304	2999.687473
RMSE	55.006357	54.769403
MAE	0.735157	0.664412
R-Squared	1.000000e+00	1.000000e+00

Analisis Model Terbaik

Feature Importance



STACK AIR: Perkiraan total bahan kimia yang dilepas sebagai emisi udara melalui cerobong (sumber titik) di fasilitas terkait.
FUGITIVE AIR: Perkiraan total bahan kimia beracun yang dilepas sebagai emisi udara tak terkendali di fasilitas terkait.

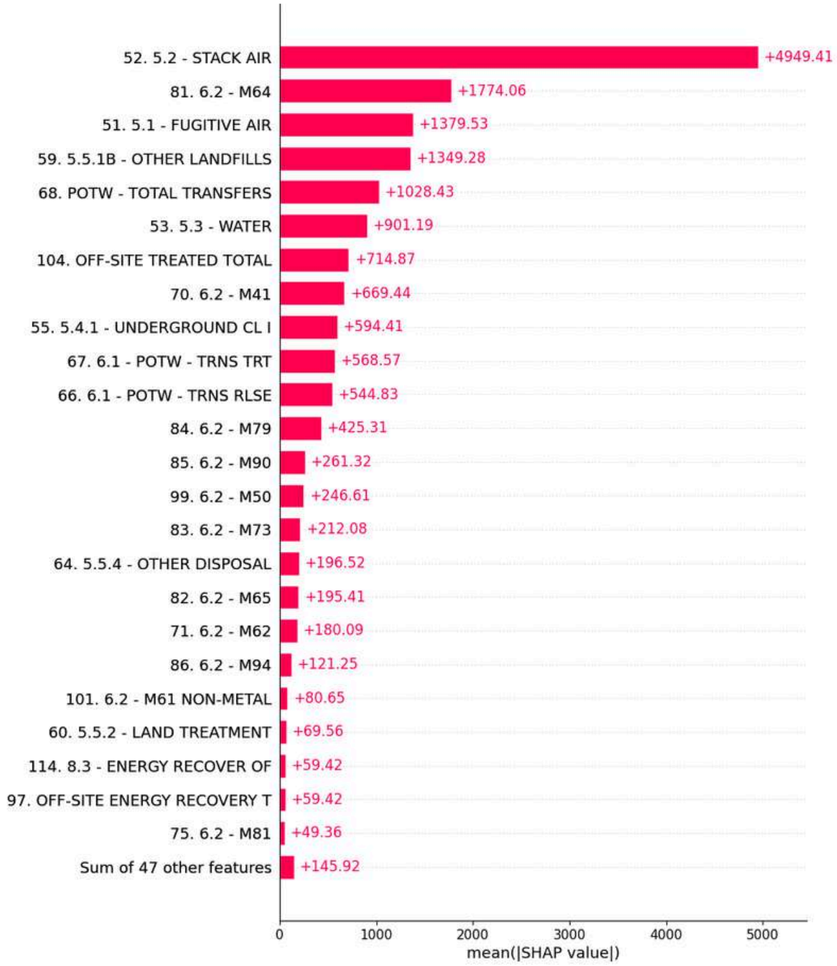
Emisi udara melalui STACK AIR dan FUGITIVE AIR memiliki kontribusi dominan.

Interpretasi

Di 10 negara bagian terpadat di AS, polusi udara menjadi limbah utama yang dihasilkan. Kondisi ini menunjukkan perlunya fokus lebih dalam pada sumber polusi udara untuk mengendalikan dampak lingkungan.

Analisis Model Terbaik

SHAP



M64: Jumlah total bahan kimia yang dipindahkan untuk dibuang ke landfill lain

Emisi udara melalui STACK AIR, M64, dan FUGITIVE AIR memiliki kontribusi dominan.

Interpretasi

Di 10 negara bagian terpadat di AS, STACK AIR dan FUGITIVE AIR menjadi sumber utama polusi udara, sementara M64 berkontribusi pada polusi tanah. Kondisi ini menekankan perlunya perhatian khusus untuk mengendalikan emisi udara dari STACK AIR dan FUGITIVE AIR, serta pembuangan limbah ke M64 agar dampak lingkungan dapat diminimalkan.

Simulasi Prediksi

Model terbaik :
Ridge Regression

Simulasi Prediksi menggunakan Dash

Fitur Untuk Simulasi Prediksi

3 fitur paling berpengaruh yang dipilih berdasarkan *feature importance,* yaitu FUGITVE AIR, SLACK AIR, dan OTHER LANDFILLS.

Pengguna bisa menggeser 3 parameter tersebut dan fitur lainnya digunakan nilai *baseline*.

Perubahan fitur dihitung dalam bentuk persentase (%), memungkinkan pengguna untuk melihat dampaknya secara langsung. Dengan mengklik atau menggeser slider, pengguna dapat melihat bagaimana perubahan pada fitur tersebut mempengaruhi prediksi Total Releases.

Simulasi Prediksi Total Limbah Berbahaya

Dashboard Prediksi Total Limbah Berbahaya

Petunjuk Penggunaan Dashboard:

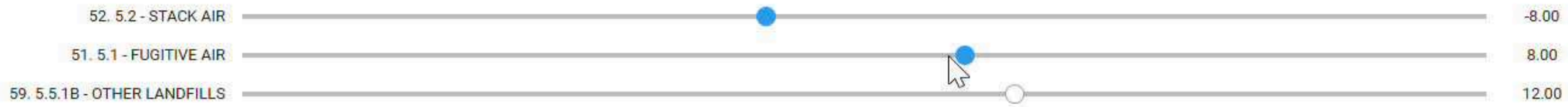
- Atur nilai perubahan fitur menggunakan slider untuk fitur **dampak tinggi**.
- Perubahan fitur dihitung dalam bentuk persentase (%).
- Klik atau geser slider untuk melihat perubahan prediksi **Total Releases**.
- Fitur yang tidak diubah akan menggunakan nilai baseline.

Contoh:

- **52. 5.2 - STACK AIR**: Tingkatkan nilai sebanyak 10%, maka prediksi akan menyesuaikan.
- **51. 5.1 - FUGITIVE AIR**: Kurangi nilai sebanyak 20%, maka hasil prediksi juga berubah.

Dashboard ini dapat dikembangkan dengan menambahkan lebih banyak fitur atau grafik visualisasi.

Fitur yang Dapat Diubah:



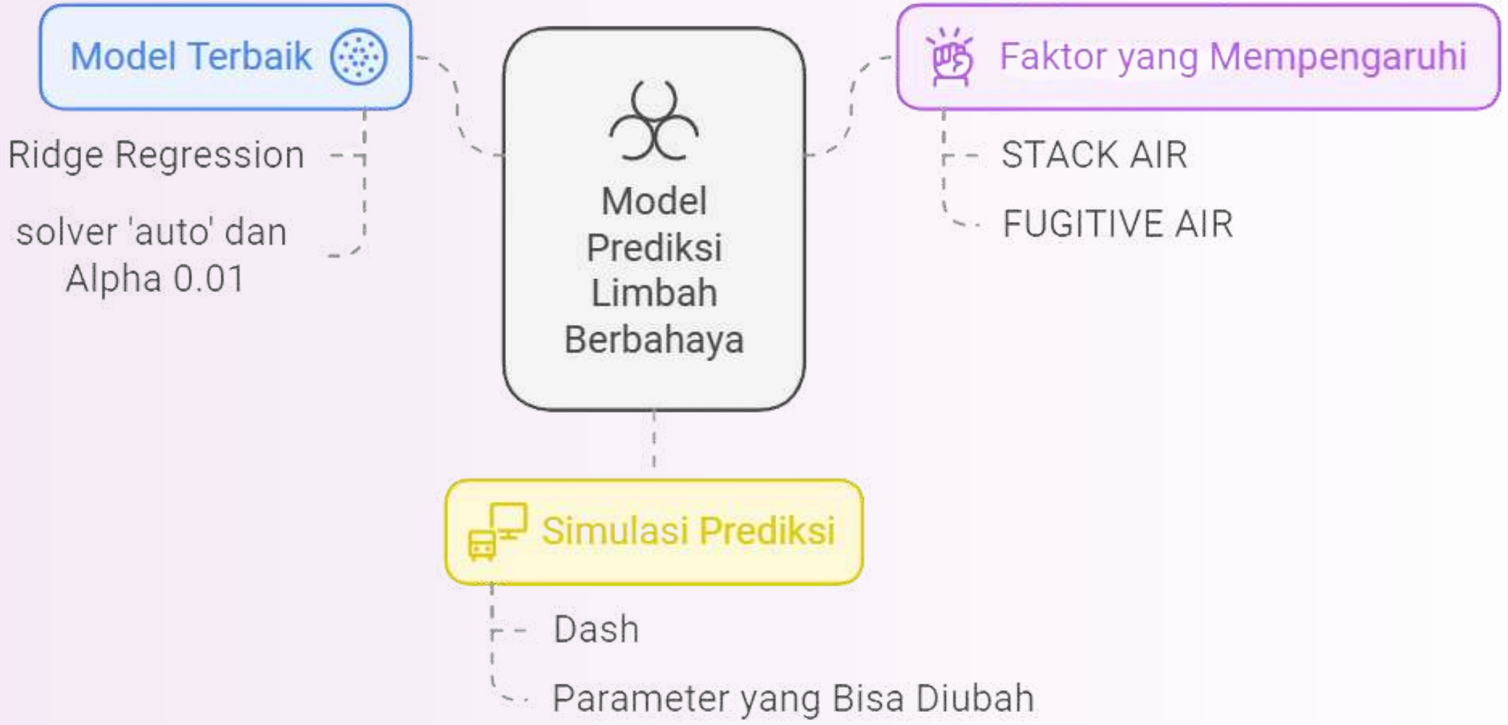
Hasil Prediksi:

```
Prediksi Total Releases: 4325.56
```

Kesimpulan

1. Kesimpulan
2. Rekomendasi

Kesimpulan


Model Terbaik
Ridge Regression
solver 'auto' dan Alpha 0.01
Model Prediksi Limbah Berbahaya
Faktor yang Mempengaruhi
STACK AIR
FUGITIVE AIR
Simulasi Prediksi
Dash
Parameter yang Bisa Diubah

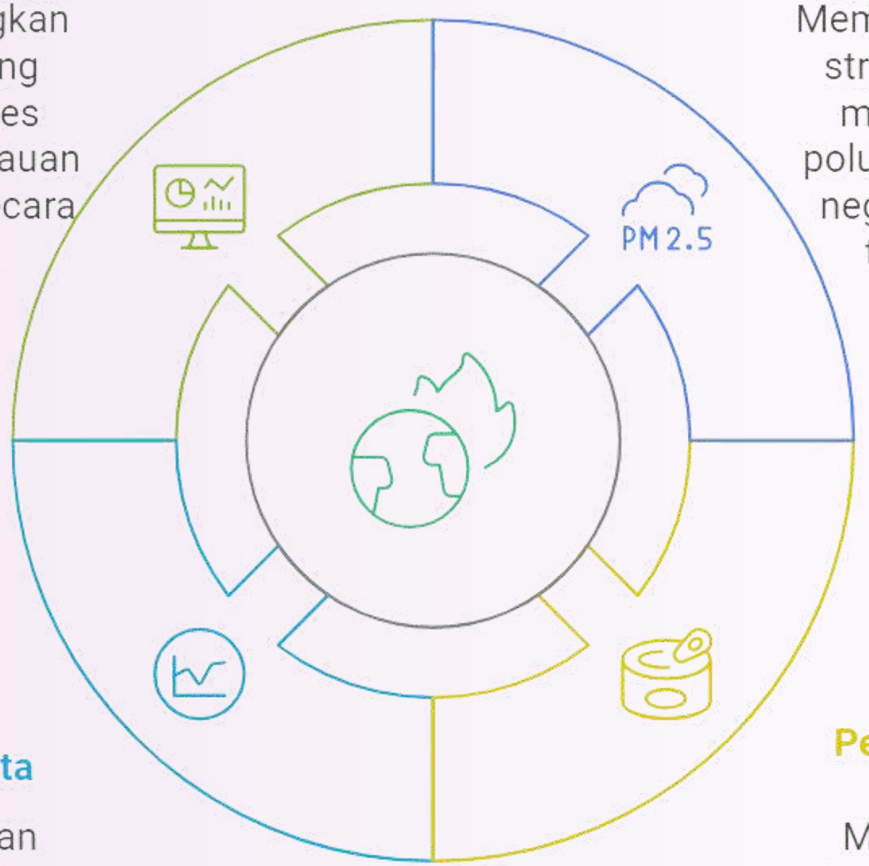
Rekomendasi

Dashboard Publik

Mengembangkan platform yang dapat diakses untuk pemantauan lingkungan secara real-time.

Polusi Udara

Memprioritaskan strategi untuk mengurangi polutan udara di negara bagian terpadat.



Analisis Data

Menggunakan data historis untuk mengidentifikasi tren dan pola lingkungan.

Pengelolaan Limbah

Mengurangi jumlah limbah yang dibuang ke tempat pembuangan sampah di negara bagian terpadat.

Daftar Pustaka

[1] Statista. (2024). Countries with the largest population worldwide in 2023. Diakses pada 6 Januari 2025 dari https://www.statista.com/statistics/262879/countries-with-the-largest-population/

[2] Yahoo Finance. (2024). Top manufacturing country in the world in 2024. Diakses pada 6 Januari 2025 dari https://finance.yahoo.com/news/top-manufacturing-country-world-2024-231816018.html

[3] The Guardian. (2023). EPA faces calls to regulate PFAS 'forever chemicals' in waste. Diakses pada 6 Januari 2025 dari https://www.theguardian.com/environment/2023/nov/17/epa-pfas-forever-chemicals-waste-pollution-unregulated

[4] U.S. Environmental Protection Agency. (2024.). Health and ecological hazards caused by hazardous substances. Diakses pada 6 Januari 2025 dari https://www.epa.gov/emergency-response/health-and-ecological-hazards-caused-hazardous-substances

[5] U.S. Environmental Protection Agency. (2024). Hazardous waste. Diakses pada 6 Januari 2025 dari https://archive.epa.gov/epawaste/hazard/web/html/index.html

[6] Sadat, L. A., MKK, S. O., Aminuddin, S. K., Putri Rizki Amalia Badri, M. K. M., Luh Suranadi, S. K. M., Sujaya, I. N., ... & Ok, S. (2024). PENGANTAR KESEHATAN MASYARAKAT. CV Rey Media Grafika.

[7] Indonesia Safety Center. (2024). 6 sumber utama munculnya limbah B3 di perusahaan: Identifikasi dan pengelolaan yang tepat. Indonesia Safety Center. https://indonesiasafetycenter.org/6-sumber-utama-munculnya-limbah-b3-di-perusahaan-identifikasi-dan-pengelolaan-yang-tepat/

[8] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

[9] Rajan, M. P. (2022). An efficient Ridge regression algorithm with parameter estimation for data analysis in machine learning. SN Computer Science, 3(2), 171.

[10] Kumari, K., & Yadav, S. (2018). Linear regression analysis study. Journal of the practice of Cardiovascular Sciences, 4(1), 33-36.

[11] Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. Procedia engineering, 201, 746- 755.

[12] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301-320

[13] Amansyah, I., Indra, J., Nurlaelasari, E., & Juwita, A. R. (2024). Prediksi Penjualan Kendaraan Menggunakan Regresi Linear: Studi Kasus pada Industri Otomotif di Indonesia. Innovative: Journal Of Social Science Research, 4(4), 1199-1216.

[14] Murukonda, V. S. N. M., & Gogineni, A. C. (2022). Prediction of air quality index using supervised machine learning.

Terima Kasih

Statmat Team
Universitas Indonesia

Siti Nur Salamah
Maryesta Apriliani Sihombing
Raissa Anggia Maharani