

Klasifikasi Kondisi Udara dengan Sensor MQ Series dan MG Menggunakan K-NN, Decision Tree dan Random Forest

Agus Dwi Churniawan¹, Titik Lusiani²

Universitas Dinamika, agusdwi@dinamika.ac.id,

Universitas Dinamika, lusiani@dinamika.ac.id

Abstrak

Polusi udara merupakan salah satu masalah lingkungan yang sangat serius di perkotaan. Polusi udara dapat menyebabkan berbagai penyakit pernapasan, kardiovaskuler, dan bahkan kanker. Kota Jakarta merupakan salah satu kota yang terkena dampak polusi udara secara signifikan. Tujuan Penelitian ini adalah Mengembangkan sistem pemantauan polusi udara menggunakan sensor MQ. dan menganalisis akurasi pengukuran sensor MQ. Manfaat Penelitian mengembangkan teknologi pemantauan polusi udara yang efektif dan efisien. Hasil penelitian ini adalah perbandingan prediksi untuk menentukan hasil Normal situation, Preparing meals, Presence of smoke, dan Cleaning dengan metode K Nearest Neighbors, Decision Tree dan Random Forest menunjukkan nilai F1-Score K Nearest Neighbors lebih baik dari Decision Tree dan Random Forest yaitu 0.941 sampai 0.981.

Kata Kunci: K-NN, Decision Tree, Random Forest, sensor MQ.

PENDAHULUAN

Polusi udara merupakan masalah lingkungan yang serius di perkotaan, menyebabkan berbagai penyakit pernapasan dan kardiovaskular. Sensor MQ-135, MQ-7 dan MQ-9 dapat mendeteksi konsentrasi gas-gas berbahaya seperti CO, NO₂, dan O₃. Penelitian ini bertujuan untuk mengembangkan sistem pemantauan polusi udara menggunakan sensor MQ.

Pengembangan Sistem Pemantauan Polusi Udara dengan Sensor MQ-135 dan Klasifikasi menggunakan Algoritma K-NN untuk meningkatkan akurasi pengukuran.

Polusi udara memerlukan pemantauan akurat. Penelitian ini menganalisis klasifikasi polusi udara menggunakan sensor MQ-7 dan algoritma Decision Tree untuk meningkatkan efektivitas pemantauan.

Perbandingan Algoritma Klasifikasi Polusi Udara dengan Sensor MQ-9 menggunakan K-NN, Decision Tree dan Random Forest.

Pemilihan algoritma klasifikasi yang tepat sangat penting dalam pemantauan polusi udara. Penelitian ini membandingkan akurasi algoritma K-NN, Decision Tree dan Random Forest dalam klasifikasi polusi udara menggunakan sensor MQ-9.

Pengaruh Faktor Lingkungan terhadap Akurasi Pengukuran Polusi Udara dengan Sensor MQ-135.

Faktor lingkungan seperti suhu dan kelembaban mempengaruhi akurasi pengukuran polusi udara. Penelitian ini menganalisis pengaruh faktor

lingkungan terhadap akurasi pengukuran polusi udara menggunakan sensor MQ-135.

Pendahuluan: Integrasi IoT dalam pemantauan polusi udara meningkatkan efisiensi dan akurasi. Penelitian ini mengembangkan sistem pemantauan polusi udara menggunakan sensor MQ-7 dan integrasi IoT untuk meningkatkan kemampuan pemantauan.

Berdasarkan penelitian diatas bagaimana cara mengukur konsentrasi polusi udara dengan sensor MQ dan Seberapa akurat sensor MQ dalam mendeteksi polusi udara

Tujuan Penelitian ini adalah Mengembangkan sistem pemantauan polusi udara menggunakan sensor MQ. dan menganalisis akurasi pengukuran sensor MQ.

Manfaat Penelitian mengembangkan teknologi pemantauan polusi udara yang efektif dan efisien.

METODE

Skema Penelitian

Pada penelitian ini dilakukan tahapan pengambilan data dari input 6 sensor dengan fungsi dan range konversi ke volt

- MG811 untuk deteksi kadar CO₂ range 350~10000 ppm
- MQ2 untuk deteksi kadar CO atau smoke range 300~10000 ppm
- MQ9 untuk deteksi dual gas CO range 10~500 ppm dan range 300~10000 ppm CH₄
- MQ135 untuk deteksi ammonia gas, toluene, Hydrogen, smoke range: 10~1000ppm,

- MQ137 untuk deteksi Ammonia, NH₃ range: 5~500ppm NH₃
- MQ138 untuk deteksi toluene, acetone, alcohol, hydrogen range: 5~500ppm

Output dengan kondisi tertentu ditetapkan keluaran ada 4 kondisi

1. Normal situation = clean air (sample 595)
2. Preparing meals = cooking (sample 515)
3. Presence of smoke = burn (sample 195)
4. Cleaning = spray dan ammonia dan alcohol (sample 540)

Data yang didapatkan adalah

	MQ2	MQ9	MQ135	MQ137	MQ138	MG811	TARGET
0	641	674	1156	1652	1410	2433	4
1	642	646	1159	1643	1455	2361	4
2	640	590	1105	1608	1459	2427	4
3	616	627	1192	1637	1466	2447	4
4	780	896	1438	1813	1489	2434	4
	MQ2	MQ9	MQ135	MQ137	MQ138	MG811	TARGET
1839	862	826	1564	1768	1540	2037	4
1840	917	821	1571	1779	1543	2008	4
1841	925	832	1582	1776	1545	1989	4
1842	928	840	1587	1787	1538	1986	4
1843	926	840	1606	1785	1555	1996	4

Proses selanjutnya adalah proses kalibrasi sensor dengan output analog to digital pada raspberry pi, dengan formula berikut:

1. Tekanan (Pressure Sensor):

$$P = (V_{out} - V_0) / (V_{fs} - V_0) * P_{fs}$$
 - P : tekanan sebenarnya
 - V_{out} : tegangan keluaran sensor
 - V₀ : tegangan offset
 - V_{fs} : tegangan penuh skala
 - P_{fs} : tekanan penuh skala
2. Suhu (Thermocouple):

$$T = (V_{out} - V_0) / (V_{fs} - V_0) * T_{fs}$$
 - T : suhu sebenarnya
 - V_{out} : tegangan keluaran sensor
 - V₀ : tegangan offset
 - V_{fs} : tegangan penuh skala
 - T_{fs} : suhu penuh skala
3. Kualitas Udara (MQ Sensor):

$$ppm = (R_s / R_0)^{(-b)}$$
 - ppm: konsentrasi gas
 - R_s: resistansi sensor
 - R₀: resistansi awal
 - b: konstanta kalibrasi

Proses selanjutnya adalah perbandingan skala antara fitur pada sensor input.

Proses selanjutnya adalah proses klasifikasi dari input sebagai fitur dan target output dengan menerapkan metode k means. Kemudian dilanjutkan proses prediksi output dengan random input.

Proses evaluasi dengan Accuracy merujuk pada tingkat keakuratan model machine learning dalam memprediksi hasil. Nilai accuracy dihitung dengan rumus:

Rumus Accuracy

$$\text{Accuracy} = \frac{(\text{Jumlah prediksi benar})}{(\text{Jumlah total data})}$$

Classification Report adalah sebuah laporan yang menampilkan kinerja model klasifikasi dalam memprediksi kelas atau kategori dari dataset. Berikut adalah penjelasan komponen-komponen yang terdapat dalam Classification Report:

Kinerja Model

1. Precision (Presisi): Jumlah prediksi benar positif dibagi jumlah prediksi positif.
 Contoh: 90% berarti 90% dari prediksi positif benar.
2. Recall (Panggilan Kembali): Jumlah prediksi benar positif dibagi jumlah sampel positif sebenarnya.
 Contoh: 80% berarti 80% dari sampel positif sebenarnya diprediksi benar.
3. F1-Score: Rata-rata harmonik dari precision dan recall.
 Contoh: 0,85 berarti keseimbangan antara precision dan recall.
4. Support: Jumlah sampel dalam setiap kelas.

Metriks Evaluasi

1. Accuracy (Akurasi): Jumlah prediksi benar dibagi jumlah total sampel.
2. Error Rate (Tingkat Kesalahan): Jumlah prediksi salah dibagi jumlah total sampel.
3. Mean Absolute Error (MAE): Rata-rata selisih absolut antara prediksi dan nilai sebenarnya.
4. Mean Squared Error (MSE): Rata-rata kuadrat selisih antara prediksi dan nilai sebenarnya.

Klasifikasi

1. True Positive (TP): Prediksi benar positif.
2. True Negative (TN): Prediksi benar negatif.
3. False Positive (FP): Prediksi salah positif.
4. False Negative (FN): Prediksi salah negatif.

Algoritma K-Nearest Neighbors (KNN), Formula Dasar

1. Jarak antara sampel baru (x) dan sampel pelatihan (x_i):
$$d(x, x_i) = \sqrt{(x_1 - x_{i1})^2 + (x_2 - x_{i2})^2 + \dots + (x_n - x_{in})^2}$$
2. K-Nearest Neighbors (KNN) memilih k-sampel terdekat berdasarkan jarak.
3. Voting untuk menentukan kelas:

Jenis Jarak

1. Euclidean Distance (jarak Euklides): $d(x, x_i) = \sqrt{\sum (x_i - x)^2}$
2. Manhattan Distance (jarak Manhattan):
 $H = -\sum (p * \log_2(p))$
 $d(x, x_i) = \sum |x_i - x|$
3. Minkowski Distance (jarak Minkowski):
 $d(x, x_i) = (\sum |x_i - x|^p)^{1/p}$
4. Cosine Similarity (kesamaan kosinus):
 $d(x, x_i) = (x \cdot x_i) / (||x|| ||x_i||)$

Algoritma KNN

1. Masukkan data pelatihan dan data uji.
2. Tentukan nilai k.
3. Hitung jarak antara data uji dan data pelatihan.
4. Pilih k-sampel terdekat.
5. Tentukan kelas berdasarkan voting.

Kelebihan:

1. Mudah dipahami dan diimplementasikan.
2. Tidak memerlukan asumsi distribusi data.
3. Dapat menangani data tidak linier.

Kekurangan:

1. Sensitif terhadap noise dan outlier.
2. Memerlukan pemilihan nilai k yang tepat.
3. Komputasi mahal untuk dataset besar.

Pemilihan Nilai K

1. Cross-validation.
2. Grid search.
3. Random search.
4. Metode heuristik (misalnya, $k = \sqrt{n}$, dengan n sebagai jumlah sampel).

Decision Tree Classification formula dan konsep dasar, Formula Dasar

1. Entropy (H): Ukuran ketidakpastian dalam dataset. dengan p sebagai proporsi kelas.
$$H = -\sum (p * \log_2(p))$$
2. Information Gain (IG): Perubahan entropi setelah membagi dataset.
$$IG = H(\text{sebelum}) - H(\text{setelah})$$
3. Gini Index (GI): Ukuran ketidakmurnian dataset.
$$GI = 1 - \sum (p^2)$$

Algoritma Decision Tree

1. Pilih atribut terbaik berdasarkan Information Gain atau Gini Index.
2. Bagi dataset berdasarkan atribut terpilih.
3. Ulangi langkah 1-2 hingga mencapai kriteria berhenti (misalnya, jumlah sampel minimal).
4. Tentukan kelas berdasarkan mayoritas kelas dalam node.

Random Forest formula dan konsep dasar untuk:

1. Entropy (H): Ukuran ketidakpastian dalam dataset.
$$H = -\sum (p * \log_2(p))$$

dengan p sebagai proporsi kelas.
2. Gini Index (GI): Ukuran ketidakmurnian dataset.
$$GI = 1 - \sum (p^2)$$
3. Information Gain (IG): Perubahan entropi setelah membagi dataset.
$$IG = H(\text{sebelum}) - H(\text{setelah})$$

Algoritma Random Forest

- 1) Buat ensemble dari banyak Decision Tree.
- 2) Setiap Decision Tree dibuat dengan:
 - Mengambil sampel acak dari dataset (Bootstrap).
 - Menggunakan atribut acak untuk membagi node.
 - Membuat keputusan berdasarkan mayoritas kelas.
- 3) Gabungkan hasil dari semua Decision Tree untuk memprediksi kelas.

Parameter Utama

1. n_estimators: Jumlah Decision Tree dalam ensemble.
2. max_depth: Kedalaman maksimum Decision Tree.
3. min_samples_split: Jumlah sampel minimal untuk membagi node.
4. min_samples_leaf: Jumlah sampel minimal dalam leaf node.
5. random_state: Pengaturan biji acak.

Kelebihan

1. Mengurangi overfitting.
2. Meningkatkan akurasi.
3. Tahan terhadap noise dan outlier.
4. Dapat menangani data tidak linier.

Kekurangan

1. Komputasi mahal.
2. Sulit untuk diinterpretasikan.
3. Memerlukan pemilihan parameter yang tepat.

HASIL DAN PEMBAHASAN

Pada penelitian dengan input enam sensor pendukung faktor polusi udara dengan beberapa metode pada machine learning.

Metode K Nearest Neighbors didapatkan

Accuracy: 0.9621

Classification Report:

	precision	recall	f1-score	support
1	0.973	0.988	0.981	179.00
2	0.949	0.961	0.955	155.00
3	0.918	0.965	0.941	58.00
4	0.981	0.932	0.955	162.00
accuracy	0.96	0.96	0.96	0.96
macro avg	0.955	0.962	0.958	554.00
weighted avg	0.963	0.962	0.962	554.00

Confusion Matrix:

```
[[177 1 0 1] [ 4 149 1 1]
 [ 0 1 56 1] [ 1 6 4 151]]
```

Metode Decision Tree di dapatkan

Accuracy: 0.9368

Classification Report:

	precision	recall	f1-score	support
1	0.972	0.977	0.974	179.00
2	0.928	0.916	0.922	155.00
3	0.932	0.948	0.940	58.00
4	0.907	0.907	0.907	162.00
accuracy	0.936	0.936	0.936	0.936
macro avg	0.934	0.937	0.936	554.00
weighted avg	0.936	0.936	0.936	554.00

Confusion Matrix:

```
[[175 2 0 2] [ 3 142 0 10]
 [ 0 0 55 3] [ 2 9 4 147]]
```

Metode Random Forest di dapatkan

Accuracy: 0.9566

Classification Report:

	precision	recall	f1-score	support
1	0.994	0.977	0.985	179.00
2	0.948	0.954	0.951	155.00
3	0.931	0.931	0.931	58.00
4	0.932	0.944	0.938	162.00
accuracy	0.956	0.956	0.956	0.956
macro avg	0.951	0.951	0.951	554.00
weighted avg	0.956	0.956	0.956	554.00

Confusion Matrix:

```
[[175 2 0 2] [ 0 148 1 6]
 [ 0 1 54 3] [ 1 5 3 153]]
```

KESIMPULAN

Pada penelitian ini perbandingan prediksi untuk menentukan hasil Normal situation, Preparing meals, Presence of smoke, dan Cleaning dengan metode K Nearest Neighbors, Decision Tree dan Random Forest menunjukkan nilai F1-Score K Nearest Neighbors lebih baik dari Decision Tree dan Random Forest yaitu 0.941 sampai 0.981.

DAFTAR PUSTAKA

- Aufar, M., Andreswari, R., & Pramesti, D. (2020). Analisis Sentimen Media Sosial Youtube Menggunakan Algoritma Decision Tree dan Random Forest: Studi Kasus. Konferensi Internasional Data Science dan Aplikasinya (ICoDSA) 2020.
- Chen, L., et al. (2020). A Comparative Study of Machine Learning Algorithms for Air Quality Classification. *IEEE Access*, 8, 187331-187342. DOI: 10.1109/ACCESS.2020.3033438
- Febriantono, M. A., Herasmara, R., & Pangestu, G. (2020). Cost Sensitive Tree dan Naïve Bayes pada Klasifikasi Multiclass. *Repositori ITERA*.
- Kumar, P., & Gupta, R. K. (2020). Air Quality Index Prediction Using Machine Learning Algorithms. *Journal of Intelligent Information Systems*, 61(2), 257-273. DOI: 10.1007/s10844-020-00614-1.
- Li, J., & Fan, J. (2020). Air Quality Classification Using Machine Learning Techniques. *Journal of Environmental Science and Health, Part C*, 34(1), 53-65. DOI: 10.1080/10590501.2020.1715394
- Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier dengan Metode Decision Tree (C4.5) untuk Menganalisis Kelancaran Pembiayaan. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 3(2), 196-201.
- Siregar, F. A., & Widyanto, M. R. (2020). Pemantauan Kualitas Udara dengan Sensor MQ-135 dan Algoritma K-NN. *Jurnal Ilmu Komputer dan Informasi*, 13(2), 123-130.[8] Siregar, F. A., & Widyanto, M. R. (2020). Klasifikasi Polusi Udara dengan Sensor MQ-7 dan Algoritma Decision Tree. *Jurnal Ilmu Komputer dan Informasi*, 13(2), 141-148