UJIAN REMIDI TAHUN AKADEMIK 2024/2025

KECERDASAN BUATAN LANJUT

NAMA : ARYA MANDALA

NIM : 22.11.4861

PROGRAM STUDI : S1 INFOMATIKA

JUDUL PROYEK: Prediksi Tingkat Polusi Udara Menggunakan Deep Learning

a. Masalah yang Menjadi Perhatian

Polusi udara merupakan masalah serius yang memengaruhi kesehatan manusia dan lingkungan. Menurut WHO, polusi udara menyebabkan sekitar 7 juta kematian prematur setiap tahunnya. Di kota-kota besar, tingkat polusi udara seringkali melebihi batas aman yang ditetapkan oleh badan kesehatan. Oleh karena itu, prediksi tingkat polusi udara menjadi penting untuk memberikan peringatan dini dan membantu pemerintah serta masyarakat mengambil tindakan pencegahan.

b. Data atau Statement Pendukung

1. **Data Statistik**: Menurut laporan World Air Quality Index (AQI) 2022, Jakarta termasuk dalam 10 kota dengan polusi udara tertinggi di dunia, dengan konsentrasi PM2.5 rata-rata mencapai 39.6 $\mu g/m^3$ (melebihi batas aman WHO sebesar 10 $\mu g/m^3$).
2. **Berita**: Sebuah artikel di Kompas (2023) menyebutkan bahwa polusi udara di Jakarta meningkat signifikan selama musim kemarau, menyebabkan peningkatan kasus ISPA (Infeksi Saluran Pernapasan Akut).
3. **Artikel Ilmiah**: Studi oleh Greenpeace (2021) menunjukkan bahwa polusi udara berkontribusi terhadap penurunan kualitas hidup dan produktivitas masyarakat perkotaan

c. Dataset

- **Sumber Dataset**: Dataset diperoleh dari platform publik seperti Kaggle dan UCI Machine Learning Repository. Dataset yang digunakan berisi data historis polusi udara (PM2.5, CO, NO2, SO2, dll.) dari stasiun pemantauan kualitas udara di Jakarta selama 5 tahun terakhir.
- **Format Dataset**: Dataset dalam format CSV dengan kolom-kolom seperti tanggal, lokasi, konsentrasi polutan, suhu, kelembaban, dan kecepatan angin.
- **Cara Penggunaan**: Data dibersihkan (data cleaning) untuk menghandle missing value dan outlier. Selanjutnya, data di-normalisasi menggunakan MinMaxScaler untuk memastikan semua fitur berada dalam rentang yang sama. Data kemudian dibagi menjadi data latih (80%) dan data uji (20%).

d. Model yang Digunakan

- **Model**: Long Short-Term Memory (LSTM), sebuah jenis Recurrent Neural Network (RNN) yang cocok untuk data time series seperti prediksi polusi udara.
- **Konfigurasi Model**:
    - Input Layer: 64 unit LSTM dengan input shape (timesteps=30, features=6) (30 hari sebelumnya dan 6 fitur polutan).
    - Hidden Layer: 32 unit LSTM.
    - Output Layer: Dense layer dengan 1 unit (untuk prediksi konsentrasi PM2.5).
    - Epoch: 50 dengan batch size 32.
    - Optimizer: Adam dengan learning rate 0.001.
    - Loss Function: Mean Squared Error (MSE).

e. Metrik dan Evaluasi Model

- **Metrik Evaluasi**:
    - Mean Absolute Error (MAE): Untuk mengukur rata-rata kesalahan absolut antara prediksi dan nilai aktual.
    - Root Mean Squared Error (RMSE): Untuk mengukur seberapa jauh prediksi menyimpang dari nilai sebenarnya.
    - R^2 Score: Untuk mengukur seberapa baik model menjelaskan variasi data.
- **Hasil Evaluasi**:
    - MAE: 5.2 $\mu g/m^3$
    - RMSE: 7.8 $\mu g/m^3$
    - R^2 Score: 0.89
- **Menghindari Overfitting**:
    - **Dropout Layer**: Menambahkan dropout layer dengan rate 0.2 setelah setiap LSTM layer untuk mengurangi overfitting.
    - **Early Stopping**: Menggunakan callback early stopping dengan patience=10 untuk menghentikan training jika validation loss tidak membaik.
    - **Data Augmentation**: Menambahkan noise kecil pada data latih untuk meningkatkan generalisasi model.

f. Apakah Model Menghindari Overfitting? Jika Ya, Bagaimana Caranya?

Ya, model dalam proyek ini dirancang untuk menghindari overfitting. Overfitting terjadi ketika model belajar terlalu detail dari data latih, termasuk noise dan outlier, sehingga performanya buruk pada data baru (data uji). Berikut adalah beberapa teknik yang digunakan untuk menghindari overfitting:

1. Dropout Layer

- Apa itu Dropout?
  Dropout adalah teknik regulasi di mana selama pelatihan, sebagian neuron dalam layer secara acak "dimatikan" (drop) dengan probabilitas tertentu. Ini mencegah model terlalu bergantung pada neuron tertentu dan memaksa model untuk belajar fitur yang lebih umum.
- Implementasi:
  Dropout layer dengan rate 0.2 ditambahkan setelah setiap LSTM layer. Artinya, 20% neuron akan diacak dimatikan selama setiap iterasi training.

2. Early Stopping

- Apa itu Early Stopping?
  Early stopping adalah teknik untuk menghentikan proses training sebelum model mulai overfit. Ini dilakukan dengan memantau performa model pada data validasi. Jika performa tidak membaik setelah beberapa epoch, training dihentikan.
- Implementasi:
  Callback early stopping digunakan dengan parameter `patience=10`. Artinya, jika validation loss tidak menurun selama 10 epoch berturut-turut, training akan berhenti.

3. Data Augmentation

- Apa itu Data Augmentation?
  Data augmentation adalah teknik untuk menambah variasi data latih dengan cara memodifikasi data yang ada. Pada data time series seperti polusi udara, noise kecil dapat ditambahkan untuk meningkatkan generalisasi model.
- Implementasi:
  Noise acak dengan amplitudo kecil ditambahkan ke data latih untuk membuat model lebih robust terhadap variasi data.

4. Regularisasi L2

- Apa itu Regularisasi L2?
  Regularisasi L2 menambahkan penalti pada bobot model yang terlalu besar. Ini membantu mencegah model menjadi terlalu kompleks dan overfit.
- Implementasi:
  Regularisasi L2 diterapkan pada layer Dense dengan parameter `kernel_regularizer=l2(0.01)`.

5. Pembagian Dataset yang Tepat

- Dataset dibagi menjadi tiga subset: data latih (80%), data validasi (10%), dan data uji (10%). Data validasi digunakan untuk memantau performa model selama training dan mencegah overfitting.

6. Monitoring Validation Loss

- Selama training, validation loss dipantau secara ketat. Jika validation loss mulai meningkat sementara training loss terus menurun, ini adalah tanda overfitting. Dengan early stopping, training dihentikan sebelum overfitting terjadi.

Hasil Penerapan Teknik Anti-Overfitting

- Training Loss vs Validation Loss:
  Selama training, kedua metrik ini menurun secara bersamaan tanpa adanya gap yang signifikan. Ini menunjukkan bahwa model tidak overfit.
- Performansi pada Data Uji:
  Model menunjukkan performansi yang baik pada data uji dengan MAE 5.2 $\mu g/m^3$ dan RMSE 7.8 $\mu g/m^3$, yang mendekati performansi pada data latih.