Bukti Serving Model

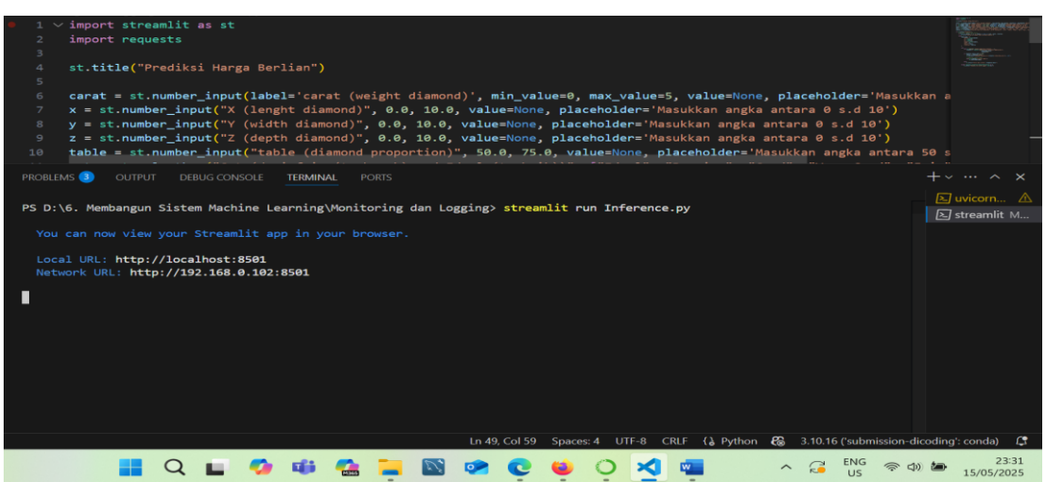
1. **Serving melalui Streamlit dan FastAPI**

Serving model dilakukan menggunakan artefak model yang sudah dibuat dan disimpan melalui MLflow. Model ini dimuat kembali di aplikasi FastAPI pada port 8000 dengan perintah:

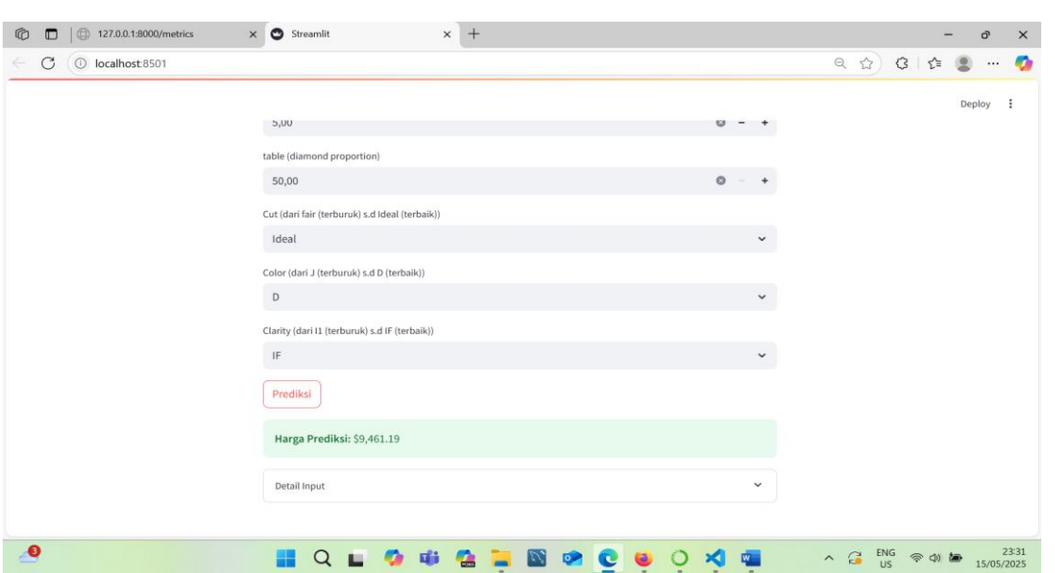
```python
model = mlflow.pyfunc.load_model(MODEL_PATH)
```

Endpoint POST /predict menerima data input dalam format JSON, memproses data tersebut dan menghasilkan prediksi harga berlian.

Untuk input data, dibuat antarmuka berbasis Streamlit. Pengguna mengisi fitur seperti carat, x, y, z, table, serta kategori cut, color, dan clarity. Setelah menekan tombol "Prediksi", Streamlit mengirimkan data ke endpoint FastAPI di http://localhost:8000/predict. Hasil prediksi kemudian langsung ditampilkan di halaman Streamlit secara real-time.



Hasil :

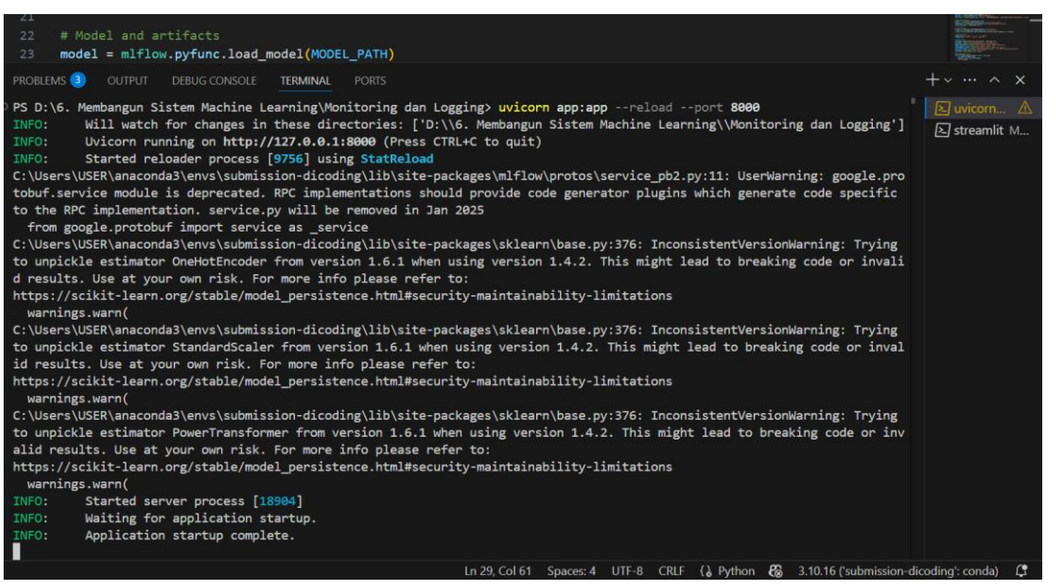


2. **Monitoring melalui Prometheus**

Kedua, dilakukan integrasi monitoring menggunakan **Prometheus**. Endpoint /metrics disediakan oleh aplikasi FastAPI, dan berisi berbagai metrik performa sistem dan model, seperti:

- Metrik sistem: penggunaan CPU, RAM, dan disk.
- Metrik model: total prediksi, error, latency prediksi, dan skor evaluasi seperti R^2 dan RMSE.

Prometheus dikonfigurasi untuk *scrape* endpoint /metrics ini, dan menyimpan metrik tersebut secara periodik untuk divisualisasikan lebih lanjut di Grafana.



Hasil :