

Bukti Serving Model

1. Serving melalui Streamlit dan FastAPI

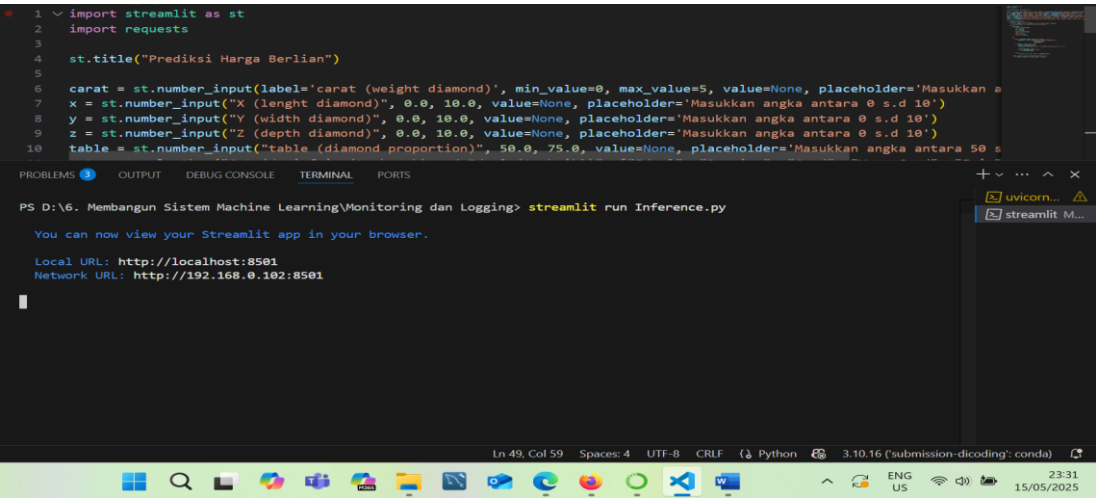
Serving model dilakukan menggunakan artefak model yang sudah dibuat dan disimpan melalui MLflow. Model ini dimuat kembali di aplikasi FastAPI pada port 8000 dengan perintah:

```
python

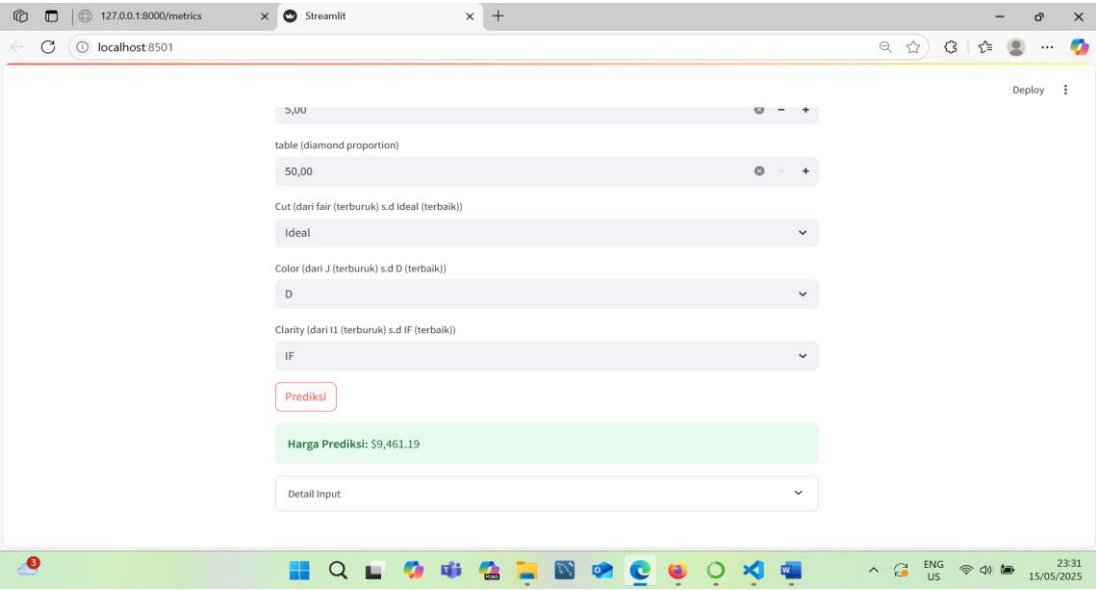
model = mlflow.pyfunc.load_model(MODEL_PATH)
```

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.

```

21
22 # Model and artifacts
23 model = mflow.pyfunc.load_model(MODEL_PATH)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

PS D:\6. Membangun Sistem Machine Learning\Monitoring dan Logging> uvicorn app:app --reload --port 8000

INFO: Will watch for changes in these directories: ['D:\6. Membangun Sistem Machine Learning\Monitoring dan Logging']

INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)

INFO: Started reload process [9756] using StatReload

C:\Users\USER\anaconda3\envs\submission-dicoding\lib\site-packages\mflow\protos\service_pb2.py:11: UserWarning: google.pro
tobuf.service module is deprecated. RPC implementations should provide code generator plugins which generate code specific
to the RPC implementation. service.py will be removed in Jan 2025

from google.protobuf import service as _service

C:\Users\USER\anaconda3\envs\submission-dicoding\lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying
to unpickle estimator OneHotEncoder from version 1.6.1 when using version 1.4.2. This might lead to breaking code or invali
d results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(

C:\Users\USER\anaconda3\envs\submission-dicoding\lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying
to unpickle estimator StandardScaler from version 1.6.1 when using version 1.4.2. This might lead to breaking code or inval
id results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(

C:\Users\USER\anaconda3\envs\submission-dicoding\lib\site-packages\sklearn\base.py:376: InconsistentVersionWarning: Trying
to unpickle estimator PowerTransformer from version 1.6.1 when using version 1.4.2. This might lead to breaking code or inv
alid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(

INFO: Started server process [18904]

INFO: Waiting for application startup.

INFO: Application startup complete.

Ln 29, Col 61 Spaces: 4 UTF-8 CRLF Python 3.10.16 (submission-dicoding: conda)

Hasil :

