

Analisis Prediktif Multiplikasi Jet Partikel pada Tabrakan Proton-Proton menggunakan Algoritma Random Forest, XGBoost, dan KNN

Penulis : Robert Gunanta Bukit

Tanggal : 2025-04-29

Abstrak

Penelitian ini bertujuan untuk memprediksi jumlah jet partikel yang terbentuk dalam peristiwa tumbukan proton-proton menggunakan dataset simulasi dari Kaggle. Metode yang digunakan mencakup Random Forest, XGBoost, dan K-Nearest Neighbors (KNN), dengan pendekatan seleksi fitur otomatis dan oversampling menggunakan SMOTE. Evaluasi dilakukan berdasarkan akurasi, f1-score, serta ketahanan model terhadap kelas minoritas. Hasil menunjukkan bahwa XGBoost dengan tuning parameter menghasilkan performa terbaik dengan akurasi 96% dan peningkatan recall pada kelas minoritas. Kajian lebih lanjut dilakukan pada korelasi fitur terhadap variabel fisika dan uji ketahanan terhadap noise data, serta analisis explainability menggunakan SHAP.

Pendahuluan

Prediksi jumlah jet partikel (n_{Jets}) dalam tabrakan proton-proton merupakan tantangan penting dalam fisika partikel karena berkaitan erat dengan identifikasi partikel berat dan pengujian teori beyond the Standard Model. Dalam konteks ini, teknik Machine Learning memungkinkan analisis data kompleks secara efisien dan presisi tinggi. Penelitian ini mengangkat masalah tersebut dengan pendekatan model ensemble dan pembelajaran berbasis jarak, serta evaluasi mendalam terhadap performa dan interpretabilitasnya.

Metodologi

Dataset terdiri dari 21.726 peristiwa tabrakan dengan 16 fitur numerik termasuk MR, Rsq, E1, Px1, ..., n_{BJets} , dan target n_{Jets} . Pra-pemrosesan melibatkan penghapusan fitur ID, standardisasi, dan balancing menggunakan SMOTE. Tiga model dilatih: Random Forest, XGBoost, dan KNN. Feature selection otomatis dilakukan menggunakan SelectKBest dan SelectFromModel. Evaluasi performa menggunakan classification report untuk interpretasi model.

Hasil dan Diskusi

Model XGBoost dengan tuning ($\text{maxdepth}=5$, $\text{nestimators}=300$, $\text{learning_rate}=0.1$) menghasilkan akurasi 96% dan recall tinggi pada kelas $n_{\text{Jets}} = 2$ dan 3. Random Forest juga menunjukkan performa kuat, terutama saat bootstrap dinonaktifkan. KNN tertinggal secara akurasi. Visualisasi feature importance dan SHAP menunjukkan bahwa fitur HT, MR, dan E1 merupakan prediktor utama. Eksperimen robustness menunjukkan bahwa XGBoost lebih stabil terhadap noise dibanding Random Forest. Selain itu, evaluasi data

asli vs hasil SMOTE menunjukkan adanya trade-off antara generalisasi dan sensitivitas model.

Kesimpulan

XGBoost terbukti sebagai model paling andal dalam prediksi jumlah jet pada data tabrakan proton. Fitur seleksi dan interpretasi memperkuat validitas model. Penelitian ini menyarankan bahwa penggabungan fisika partikel dan explainable AI membuka potensi baru dalam klasifikasi event di eksperimen LHC. Ke depan, perluasan studi dapat mencakup integrasi dengan data CMS nyata dan eksplorasi model hybrid atau graph neural networks untuk menangkap interaksi partikel secara topologis.