

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Prediksi hasil pertandingan olahraga telah berkembang menjadi bidang aplikasi yang signifikan bagi pembelajaran mesin (machine learning) dan kecerdasan buatan. Riset mutakhir menunjukkan bahwa model machine learning dapat memprediksi hasil pertandingan sepak bola dengan akurasi yang kompetitif (55-92% tergantung kompleksitas fitur dan metode), memberikan keunggulan strategis bagi analisis olahraga, pelatih, dan organisasi manajemen tim.

Liga Inggris (English Premier League / EPL) adalah salah satu kompetisi sepak bola paling kompetitif di dunia dengan catatan data yang lengkap dan berkualitas tinggi. Data historis 5 musim terakhir (2019/20 hingga 2023/24) mencakup lebih dari 1.873 pertandingan dengan ratusan metrik performa yang dapat dimanfaatkan untuk membangun model prediktif yang robust.

Dari perspektif metodologi penelitian, pengembangan model prediksi bukan hanya memerlukan algoritma yang akurat, tetapi juga kemampuan untuk menjelaskan keputusan model kepada stakeholder. Transparansi dan interpretabilitas model menjadi kritis dalam konteks pengambilan keputusan strategis, analisis risiko, dan penelitian pengembangan.

Ada beberapa celah dalam literatur yang kami coba isi, 1). Interpretabilitas sering terlupakan. Di bidang NLP dan computer vision, alat interpretabilitas seperti SHAP dan LIME sudah merevolusi cara kita memahami model. Tapi di dunia prediksi olahraga, teknik-teknik ini belum banyak digunakan. 2). Perbandingan algoritma masih chaotis. Banyak penelitian yang membandingkan algoritma, tapi masing-masing menggunakan dataset atau fitur yang berbeda, sehingga sulit untuk menarik kesimpulan yang fair. 3). Interaksi antar fitur jarang dianalisis. Model berbasis pohon seperti XGBoost secara otomatis menangkap hubungan kompleks antar fitur, tapi penjelasan spesifik tentang interaksi ini jarang diuraikan untuk praktisi. 4). Banyak penelitian tidak "berbicara" dengan dunia nyata. Temuan akademis sering kali tetap di paper dan tidak berubah menjadi rekomendasi praktis yang bisa langsung digunakan.

Kesenjangan Penelitian Utama

Perlunya pendekatan sistematis yang menggabungkan evaluasi performa algoritma machine learning dengan analisis interpretabilitas berbasis SHAP untuk memahami faktor-faktor kunci dalam prediksi hasil pertandingan sepak bola.

Signifikansi Penelitian

Penelitian ini dirancang untuk mengisi kesenjangan dengan kontribusi:

1. **Untuk Pengambilan Keputusan Praktis:** Sistem prediksi yang dapat dijelaskan (explainable) yang membantu stakeholder memahami driver utama dari hasil prediksi
2. **Untuk Pengembangan Ilmu Pengetahuan:** Metodologi inovatif dalam mengintegrasikan feature engineering olahraga dengan interpretability analysis berbasis game theory
3. **Untuk Komunitas Research:** Empirical evidence mengenai trade-off antara akurasi model dan kemampuan interpretasi

4. **Untuk Industri Analitik:** Benchmark dan best practices dalam pengembangan sistem prediksi yang dapat dijelaskan

1.2 Rumusan Masalah

Penelitian ini merumuskan masalah penelitian dalam pertanyaan-pertanyaan berikut:

1. Bagaimana performa tiga algoritma machine learning (XGBoost, Random Forest, Logistic Regression) dalam memprediksi hasil pertandingan Liga Inggris berdasarkan metrik Accuracy, F1-Score, dan ROC-AUC?
2. Algoritma mana yang memberikan akurasi dan interpretabilitas terbaik, dan bagaimana trade-off antara kedua dimensi tersebut?
3. Fitur-fitur apa yang paling berpengaruh terhadap prediksi menurut analisis SHAP?
4. Apakah model machine learning dapat mengidentifikasi fenomena "home advantage" secara empiris melalui SHAP analysis?
5. Bagaimana feature interactions dan non-linear relationships ditangkap oleh gradient boosting?

1.3 Batasan Masalah

Untuk memastikan fokus penelitian dan validitas metodologi, penelitian ini dibatasi sebagai berikut:

1. **Geografis & Kompetisi:** Analisis terbatas pada Liga Inggris (2019/20 – 2023/24)
2. **Sumber Data:** Data dari database publik (bukan proprietary klub)
3. **Fitur & Variabel:** Maksimal 22 fitur engineered (excludes: injury data, transfer values)
4. **Algoritma:** Fokus pada XGBoost, Random Forest, Logistic Regression
5. **Evaluasi:** Single execution run dengan 70:30 stratified split
6. **Generalisasi:** Model optimal untuk EPL; aplikasi ke liga lain memerlukan domain adaptation

1.4 Tujuan Penelitian

Tujuan Utama

Mengembangkan dan membandingkan tiga algoritma machine learning dalam memprediksi hasil pertandingan sepak bola Liga Inggris dengan fokus pada akurasi prediktif dan interpretabilitas model melalui analisis SHAP.

Tujuan Khusus

1. Mengumpulkan dan memproses data EPL dari 5 musim dengan validasi kualitas
2. Mengekstrak maksimal 22 fitur yang meaningful dari data historis
3. Mengimplementasikan tiga algoritma dengan hyperparameter tuning
4. Mengevaluasi performa menggunakan Accuracy, F1-Score, dan ROC-AUC
5. Menganalisis feature importance dan interactions menggunakan SHAP
6. Membandingkan accuracy vs interpretability trade-off

1.5 Manfaat Penelitian

Manfaat Teoritis

1. Pengayaan literatur ML & Sports Analytics dengan bukti empiris integrasi domain knowledge
2. Inovasi metodologi dalam comparing algorithms dengan dual focus pada accuracy dan interpretability
3. Understanding feature interactions dalam tree-based models untuk prediksi olahraga

Manfaat Praktis

1. Sistem prediksi yang dapat dijelaskan untuk pengambilan keputusan strategis
2. Benchmark praktis untuk platform analitik olahraga komersial
3. Pemahaman mendalam tentang driver prediksi untuk better risk assessment

Manfaat Akademik

1. Referensi penelitian lanjutan dan reproducible science
2. Foundation untuk transfer learning ke domain olahraga lainnya