

ABSTRAK

Tujuan: Penelitian ini mengatasi masalah kurangnya analisis interpretabilitas dalam prediksi olahraga. Kami membandingkan tiga algoritma machine learning (XGBoost, Random Forest, Logistic Regression) dan menerapkan SHAP untuk memahami fitur apa saja yang penting dalam memprediksi hasil pertandingan Liga Inggris (EPL). Metode: Kami melatih ketiga algoritma menggunakan 1.873 data pertandingan EPL dari 5 musim (2019/20–2023/24). Dataset diperkaya dengan 22 fitur khusus yang menggambarkan performa tim. Data dibagi dengan proporsi 70% untuk training dan 30% untuk testing, dengan tetap mempertahankan distribusi kelas di kedua bagian. Performa diukur menggunakan akurasi, F1-score, dan ROC-AUC. Untuk model terbaik (XGBoost), kami menerapkan SHAP TreeExplainer untuk menganalisis kontribusi setiap fitur. Hasil: XGBoost mencapai akurasi 54,45% ($F1=0,4336$, $AUC=0,5365$), lebih baik dibanding Random Forest (54,27%) dan Logistic Regression (51,42%, $AUC=0,4759$). Analisis SHAP menunjukkan bahwa penyelesaian passing tim tandang dan pertahanan tim tuan rumah adalah prediktor terkuat. Kegagalan Logistic Regression (AUC kurang dari 0,50) membuktikan bahwa model linear fundamental tidak cocok untuk masalah ini. Kesimpulan: Dengan menggabungkan XGBoost dan SHAP, kami berhasil menciptakan model yang sekaligus akurat dan dapat dijelaskan. Pendekatan kami dapat direplikasi untuk prediksi olahraga lain di berbagai liga dan cabang olahraga.

Kata Kunci: machine learning, SHAP, interpretabilitas, prediksi olahraga, feature importance, ensemble methods

ABSTRACT

Objective: This study addresses the lack of interpretability analysis in sports prediction. We compare three machine learning algorithms (XGBoost, Random Forest, Logistic Regression) and apply SHAP to understand which features are important in predicting English Premier League (EPL) match outcomes. Methods: We trained all three algorithms using 1,873 EPL match data from 5 seasons (2019/20–2023/24). The dataset was enriched with 22 specialized features describing team performance. Data was split 70% for training and 30% for testing while maintaining class distribution. Performance was measured using accuracy, F1-score, and ROC-AUC. For the best model (XGBoost), we applied SHAP TreeExplainer to analyze each feature's contribution. Results: XGBoost achieved 54.45% accuracy ($F1=0.4336$, $AUC=0.5365$), outperforming Random Forest (54.27%) and Logistic Regression (51.42%, $AUC=0.4759$). SHAP analysis showed that away team passing completion and home team defense are the strongest predictors. Logistic Regression failure (AUC below 0.50) proves that fundamental linear models are unsuitable for this problem. Conclusion: By combining XGBoost and SHAP, we successfully created a model that is both accurate and explainable. Our approach can be replicated for other sports predictions across various leagues and sports.

Keywords: machine learning, SHAP, interpretability, sports prediction, feature importance, ensemble methods