Prediction of EPL Match Outcomes Based on Match Statistics

Rafie Zaidan Umara
Departemen Teknik Informatika
Institut Teknologi Sepuluh Nopember
Surabay, Indonesia
rafieumara@gmail.com

Prediksi hasil pertandingan sepak bola telah menjadi topik menarik dalam ranah analisis data olahraga, khususnya pada English Premier League (EPL) yang merupakan liga paling kompetitif di dunia. Penelitian ini bertujuan untuk membangun model prediktif berbasis statistik pertandingan guna memperkirakan hasil akhir pertandingan (menang, seri, atau kalah). Data yang digunakan mencakup statistik historis seperti penguasaan bola, jumlah tembakan, jumlah pelanggaran, dan statistik lainnya. Berbagai algoritma pembelajaran mesin seperti Random Forest dan Logistic Regression diuji untuk mengevaluasi performa model prediksi. Hasil eksperimen menunjukkan bahwa pemodelan dengan algoritma tertentu mampu memberikan akurasi prediksi yang cukup tinggi, sehingga berpotensi untuk digunakan sebagai alat bantu pengambilan keputusan dalam konteks analisis olahraga. **(***Abstract***)**

Keywords— English Premier League (EPL), prediksi hasil pertandingan, pembelajaran mesin, statistik pertandingan, klasifikasi.Introduction (Heading 1)

I. INTRODUCTION

Prediksi hasil pertandingan sepak bola, khususnya English Premier League (EPL), merupakan tantangan menarik dalam analisis olahraga. Hasil pertandingan dipengaruhi oleh berbagai faktor statistik seperti jumlah tembakan, penguasaan bola, dan akurasi umpan. Studi ini bertujuan membangun model klasifikasi multikelas untuk memprediksi hasil pertandingan EPL (menang, seri, kalah) dengan menggunakan data statistik pertandingan dan algoritma machine learning. Analisis dilakukan pada data musim 2023/24 dan sebagian musim 2024/25 untuk mengevaluasi pengaruh fitur statistik terhadap akurasi prediksi..

II. METODOLOGI

A. Pra-pemorsesan Data

Dataset yang digunakan sudah bersih. Fitur dibuat dengan menghitung rata-rata statistik per tim seperti tembakan, penguasaan bola, dan akurasi umpan. Performa kandang dan tandang juga dihitung secara terpisah. Label target dikodekan menjadi tiga kelas: menang kandang, seri, dan menang tandang untuk keperluan klasifikasi multikelas.

B. Pemilihan Fitur

Fitur akhir yang digunakan dalam model mencakup atribut waktu, statistik rata-rata dan deviasi performa tim, serta fitur selisih kekuatan antar tim. Semua fitur numerik yang hilang diisi dengan nol untuk menjaga konsistensi input model..

C. Feature Engineering

Untuk meningkatkan kemampuan model dalam memprediksi jumlah gol kandang dan tandang dalam pertandingan sepak bola, kami melakukan proses feature engineering yang terdiri atas beberapa langkah berikut:

- Ekstraksi Fitur Tanggal: Tanggal pertandingan dikonversi ke format datetime
- Statistik Tim Historis: Rata-rata, deviasi jumlah gol kandang dan tandang yang dicetak dan kebobolan, serta tembakan dan tembakan tepat sasaran
- Selisih Kekuatan Serangan dan Pertahanan: AttackDiff dan DefenseDiff

D. Fitur Agregat Gol Tim Kandang

Ditambahkan juga fitur HomeTeam_HighScoring, yaitu persentase pertandingan kandang suatu tim di mana mereka mencetak lebih dari dua gol. Hal ini bertujuan untuk mengidentifikasi kecenderungan tim mencetak banyak gol saat bermain kendang

E. Pembentukan Dataset

Dataset dibentuk dengan menyatukan data statistik tim kandang dan tim tandang ke dalam satu baris per pertandingan. Setiap baris mencerminkan satu pertandingan dengan fitur gabungan, seperti rata-rata tembakan tim kandang vs tim tandang, serta rasio atau selisih statistik antar kedua tim. Fitur tambahan seperti performa lima pertandingan terakhir, poin klasemen sementara, dan status kandang atau tandang juga ditambahkan untuk memperkaya konteks.

Hasil pertandingan digunakan sebagai label target, dikodekan menjadi tiga kelas: 0 untuk menang tandang, 1 untuk hasil seri, dan 2 untuk menang kandang. Dataset akhir kemudian dibagi menjadi fitur (X) dan label (y) untuk proses pelatihan model..

III. MODELLING

Setelah melakukan ekstraksi fitur, kami melakukan pelatihan model regresi untuk memprediksi jumlah gol yang dicetak oleh tim tuan rumah (FullTimeHomeGoals) dan tim tamu (FullTimeAwayGoals). Selanjutnya adalah proses pelatihan dan evaluasi dilakukan dalam beberapa tahap

A. Fitur yang Digunakan

Fitur-fitur ini telah dipilih untuk mencerminkan waktu pertandingan, statistik historis performa tim, serta kekuatan ofensif dan defensif relatif antar tim. Seluruh fitur numerik ini kemudian digunakan sebagai input pada model regresi untuk memprediksi nilai kontinu jumlah gol

- Home_FullTimeHomeGoals_mean/std
- Home_FullTimeAwayGoals_mean/std
- Home_HomeShots_mean/std
- Home_HomeShotsOnTarget_mean
- Away_FullTimeAwayGoals_mean/std
- Away_FullTimeHomeGoals_mean/std

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- Away_AwayShots_mean/std
- Away_AwayShotsOnTarget_mean
- AttackDiff
- DefenseDiff
- HomeTeam_HighScoring

B. Hasil Pemodelan

Hasil pemodelan dan evaluasi untuk prediksi Home Goals

Model	***Tipe Algoritma***	***RMSE***
Ridge Regression	Linier Regulerisasi	1.277554
Linear Regression	Linier	1.277588
Gradient Boosting	Boosting (Gradien)	1.287697
Neural Network	Non-linier (Neural Network / MLP)	1.296628
Lasso Regression	Linier Regulerisasi (L1)	1.305619
SVR	Non-linier (Kernel-based)	1.310591
LightGBM	Boosting (Light GBDT)	1.335212
XGBoost	Boosting (Ekstensi GBDT)	1.415307
Random Forest	Tree Ensemble	1.437867

C. Tune Hyperparameters

Untuk meningkatkan akurasi model, dilakukan pencarian kombinasi hyperparameter optimal menggunakan dua pendekatan berbeda, tergantung kompleksitas model

Grid Search, Metode **GridSearchCV** dari scikit-learn digunakan untuk mengevaluasi seluruh kombinasi hyperparameter dalam ruang parameter yang telah ditentukan secara eksplisit.

- **Model:** Ridge Regresssion
- **Paramater Space:**
    - alpha: [0.001, 0.01, 0.1, 1, 10, 100]
    - solver: ['auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga']
- **Evaluasi:** 5-fold Cross-Validation
- **Metrik Evaluasi**: Negative Root Mean Squared Error (neg-RMSE)

Ridge Regression memberikan RMSE terbaik sebesar ±1.475 setelah tuning dengan kombinasi alpha=1.0 dan solver='lsqr'.

D. Kesimpulan Analisis

Penelitian ini bertujuan untuk memprediksi hasil pertandingan sepak bola (khususnya jumlah gol kandang dan tandang) menggunakan pendekatan machine learning berbasis data historis pertandingan Liga Inggris. Berdasarkan eksperimen dan evaluasi yang telah dilakukan, diperoleh beberapa temuan penting sebagai

Efektivitas Feature Engineering, Proses *feature engineering* yang memanfaatkan statistik historis tim (rata-rata gol, tembakan, dan selisih performa antara tuan rumah dan tamu) terbukti krusial dalam meningkatkan akurasi prediksi. Fitur-fitur seperti AttackDiff, DefenseDiff, dan IsWeekend berkontribusi signifikan terhadap performa model.

Model yang telah dilatih digunakan untuk memprediksi pertandingan-pertandingan Liga Inggris musim 2024/2025. Output utama mencakup prediksi jumlah gol kandang dan tandang, hasil pertandingan (H, D, A), serta probabilitas kemenangan dari masing-masing tim berdasarkan distribusi Poisson

- Kemenangan Tuan Rumah (H): 40.3%
- Kemenangan Tim Tamu: 31.1%
- Seri (D): 28.7%
- Prediksi model menunjukkan rata-rata skor mendekati kenyataan historis EPL (sekitar 2.6–2.8 total gol per pertandingan), menandakan bahwa model berhasil menangkap dinamika skor yang realistis.
- Dengan menyusun hasil pertandingan berdasarkan skor prediksi, diperoleh klasemen sementara musim 2024/2025 berdasarkan output model

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] M. Hui, *English Premier League (EPL) Match Data 2000–2025*, Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/marcohuiii/english-premier-league-epl-match-data-2000-2025

[2] G. Baio and M. Blangiardo, “Bayesian hierarchical model for the prediction of football results,” *Journal of Applied Statistics*, vol. 37, no. 2, pp. 253–264, 2010.

[3] A. C. Constantinou, N. E. Fenton, and M. Neil, “Pi-football: A Bayesian network model for forecasting Association Football match outcomes,” *Knowledge-Based Systems*, vol. 36, pp. 322–339, 2012.

[4] N. Tax and Y. Joustra, “Predicting the Dutch football competition using public data: A machine learning approach,” in *Proc. SIGKDD Workshop on Large-Scale Sports Analytics*, 2015.

[5] M. J. Maher, “Modelling association football scores,” *Statistica Neerlandica*, vol. 36, no. 3, pp. 109–118, 1982.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.