

# BAB 1

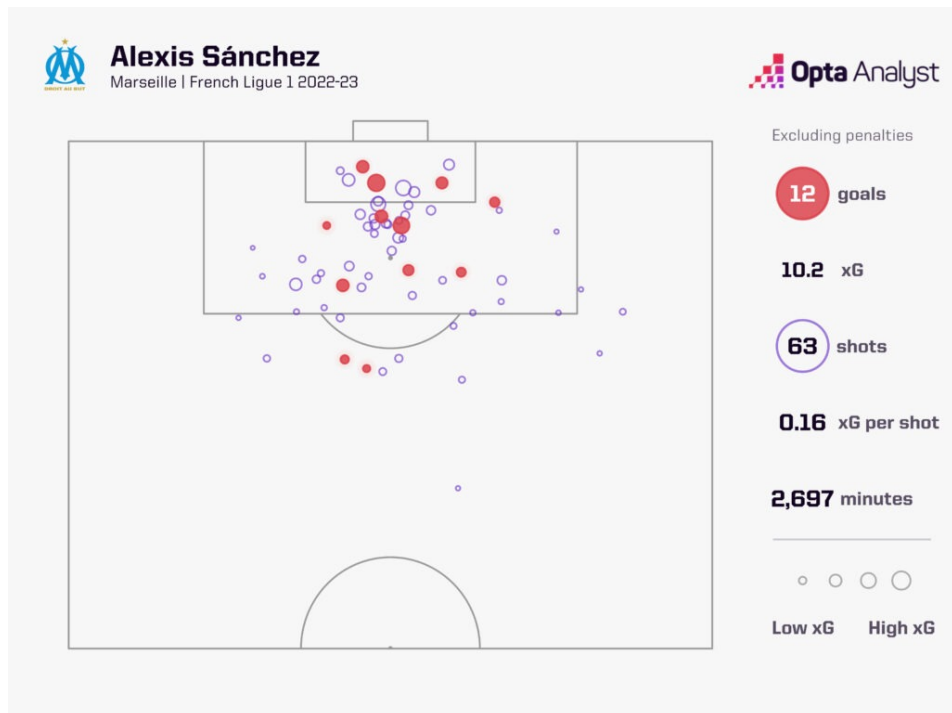
## PENDAHULUAN

### 1.1 Latar Belakang

Sepak bola modern tidak lagi hanya dipandang sebagai olahraga semata, tetapi telah berevolusi menjadi industri global yang kompleks dan kompetitif. Dalam transformasi ini, peran data *science* dan teknologi *machine learning* menjadi sangat sentral. Klub profesional kini memanfaatkan sistem informasi canggih dan perangkat *wearable* untuk mengumpulkan serta menganalisis data dalam jumlah besar. Melalui pemanfaatan data ini, pengambilan keputusan dalam aspek-aspek penting seperti taktik pertandingan, *scouting* pemain, hingga pencegahan cedera dapat dilakukan secara lebih presisi dan berbasis bukti (Chatziparaskevas *et al.*, 2024). Fenomena ini menandai pergeseran paradigma dalam pengelolaan sepak bola yang kini semakin didukung oleh pendekatan ilmiah dan teknologi prediktif.

Pada dunia analisis sepak bola modern, salah satu metrik yang paling sering digunakan untuk mengukur kualitas peluang mencetak gol adalah *Expected Goals* (xG). Metrik ini menggambarkan probabilitas suatu tembakan akan menghasilkan gol berdasarkan sejumlah variabel kontekstual. Menurut Eggels (2016), *Expected Goals* tidak hanya merepresentasikan kualitas peluang dengan cukup akurat, tetapi juga mampu memberikan wawasan penting terhadap hasil pertandingan secara keseluruhan. Agregasi nilai xG dari setiap pertandingan bahkan dapat digunakan untuk memperkirakan hasil yang seharusnya terjadi, menjadikannya alat evaluasi performa tim yang sangat berguna.

Dalam praktiknya, metrik xG dapat divisualisasikan melalui representasi spasial seperti *shot map* yang menggambarkan lokasi dan kualitas tembakan tiap pemain. Sebagai contoh, Gambar 1.2 memperlihatkan distribusi tembakan Alexis Sánchez dalam satu musim, lengkap dengan nilai xG dari masing-masing tembakan. Visualisasi ini sangat membantu pelatih dan analis dalam mengevaluasi efektivitas penyelesaian akhir dan pengambilan keputusan di area sepertiga akhir lapangan.

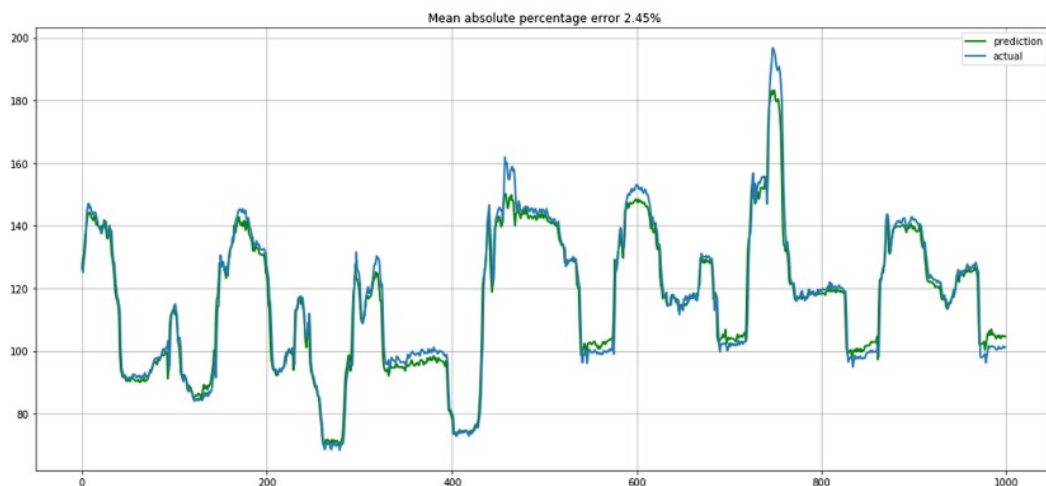


Gambar 1.1 Visualisasi *Shot-map* xG

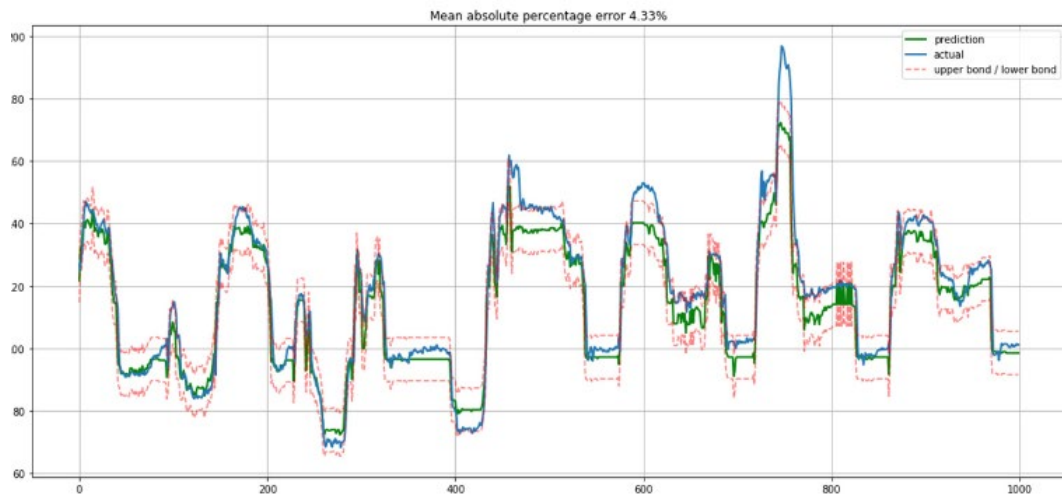
Metrik xG bertujuan untuk memprediksi probabilitas terjadinya gol dari suatu tembakan berdasarkan berbagai variabel seperti jarak, sudut, dan konteks permainan. Untuk menangkap kompleksitas spasial dan non-linear dari data pertandingan sepak bola, dibutuhkan algoritma yang tidak hanya akurat tetapi juga efisien secara komputasi. Dalam penelitian ini, *Light Gradient Boosting Machine* (LightGBM) menjadi algoritma yang sangat potensial karena kemampuannya dalam menangani *dataset* besar, memproses fitur dalam jumlah banyak, serta membangun model prediktif non-linear dengan waktu pelatihan yang jauh lebih cepat dibandingkan metode *boosting* konvensional tanpa mengorbankan akurasi (Hartanto *et al.*, 2023).

LightGBM sangat relevan diterapkan dalam konteks data sepak bola yang bersifat spasial-temporal, kompleks, dan sering kali tidak linier. Selain itu, algoritma ini memiliki kelebihan dalam kemampuan interpretasi dan efisiensi waktu komputasi yang menjadikannya ideal untuk kebutuhan praktis seperti pembuatan model prediksi xG yang presisi dan dapat diterapkan secara *real-time* (Shalom Artzi *et al.*, 2020).

Perbandingan performa antara LightGBM dengan model lain seperti XGBoost menunjukkan hasil yang kompetitif. Dalam penelitian oleh Nemeth *et al.* (2019), LightGBM menunjukkan tingkat kesalahan prediksi (MAPE) yang lebih rendah dibandingkan XGBoost, yaitu sebesar 2,45%. Ini mengindikasikan bahwa LightGBM mampu memberikan prediksi yang lebih akurat dalam konteks pemodelan numerik, dan memiliki potensi besar untuk diterapkan pada berbagai domain prediktif, termasuk analisis sepak bola. Gambar 1.2 menunjukkan grafik yang membandingkan MAPE dari berbagai model, di mana LightGBM menonjol dengan nilai MAPE terbaik. Sementara itu, Gambar 1.3 menampilkan performa model lain, yang meskipun kompetitif, tidak mencapai tingkat akurasi yang sama seperti yang dicapai oleh LightGBM. Visualisasi ini memperkuat argumen tentang keunggulan LightGBM dalam analisis prediktif.



Gambar 1.2 Grafik Performa Model LGBM (Nemeth et al., 2019)



Gambar 1.3 Grafik Performa Model Lain (Nemeth et al., 2019)

Lebih jauh, LightGBM dikenal sebagai *framework gradient boosting* berperforma tinggi yang berbasis algoritma *decision tree*. Algoritma ini menawarkan kecepatan pelatihan yang tinggi dan efisiensi dalam menangani *dataset* besar tanpa mengorbankan akurasi. Keunggulan ini telah dibuktikan dalam berbagai domain, bahkan ke sektor kesehatan seperti diagnosis penyakit dan prediksi klinis, di mana kebutuhan akan klasifikasi cepat dan akurat sangat penting (Artzi et al., 2020).

Pada pembangunan model xG, LightGBM menawarkan kemampuan untuk belajar dari data historis dengan efisiensi tinggi. Menurut Ke *et al.* (2017), LightGBM dikembangkan untuk mengatasi keterbatasan *Gradient Boosting Decision Tree* (GBDT) dalam menangani *big data*, dengan waktu pelatihan yang hingga 20 kali lebih cepat namun tetap mempertahankan tingkat akurasi yang sebanding (Hartanto *et al.*, 2023). Kemampuan ini menjadikannya kandidat kuat untuk membangun model prediksi dalam skenario dengan data tembakan yang kompleks dan besar seperti dalam sepak bola profesional.

Meskipun demikian, eksplorasi mendalam mengenai penerapan LightGBM pada model xG belum banyak dilakukan. Penelitian yang ada lebih banyak menggunakan pendekatan lain dengan berbagai kelebihan dan kekurangannya. Sebagai contoh, penelitian awal oleh Lucey *et al.* (2015) mengembangkan model *Expected Goal Value* (EGV) menggunakan algoritma *Conditional Random Fields*

(CRF) dan data *spasio-temporal* dari lebih dari 9.000 tembakan. Model ini memberikan konteks yang kuat terhadap peluang gol dengan mempertimbangkan sepuluh detik *gameplay* sebelum tembakan. Namun, CRF memiliki kekurangan dalam hal kompleksitas pelatihan dan sensitivitas terhadap kualitas fitur *input* (Sutton & McCallum, 2012), yang menghambat penerapannya dalam skala besar atau sistem *real-time*.

Penelitian oleh Fairchild *et al.* (2018) mengembangkan model *Expected Goals* menggunakan algoritma regresi logistik, berdasarkan 1.115 tembakan non-penalti dari 99 pertandingan dalam kompetisi *Major League Soccer* (MLS). Mereka menyusun model berdasarkan koordinat tembakan serta variabel spasial lain, dan menggabungkannya dengan pendekatan analisis *fraktal* untuk mengukur kompleksitas area tembakan. Regresi logistik sebagai model linier umum, memang mudah diinterpretasi dan memiliki kompleksitas komputasi rendah. Namun, pendekatan ini memiliki keterbatasan dalam menangani hubungan non-linear yang kompleks antar fitur (Bache-Mathiesen *et al.*, 2021), serta kurang adaptif terhadap *outlier* (Idris *et al.*, 2024) dan interaksi variabel spasial yang sering kali bersifat dinamis dalam sepak bola (Mishra *et al.*, 2021).

Penelitian oleh Tureen dan Olthoff (2022) menjadi salah satu pendekatan paling modern dalam kuantifikasi kontribusi pemain melalui model *Estimated Player Impact* (EPI). Mereka menggunakan algoritma *Generalised Linear Mixed Models* (GLMM) pada lebih dari 900 pertandingan dari Liga Inggris dan *Women's Super League*, dengan data dari penyedia yang sama dengan penelitian ini, yaitu StatsBomb. GLMM menawarkan fleksibilitas dalam menangani hierarki data, namun cenderung memerlukan asumsi distribusi yang ketat (McCulloch & Neuhaus, 2011) dan proses komputasi yang berat saat diterapkan pada *dataset* besar (Guan & Haran, 2016). Hal ini menyulitkan penerapannya untuk kasus prediksi *granular* seperti estimasi xG tembakan per tembakan yang membutuhkan efisiensi prediksi tinggi dan ketepatan dalam menangkap pola non-linear yang kompleks (Bolker *et al.*, 2009).

Sementara itu, Cavus dan Biecek (2022) mengevaluasi berbagai model dalam kerangka AutoML menggunakan lebih dari 315.000 data tembakan dari lima

liga top Eropa. Mereka menguji algoritma seperti XGBoost, CatBoost, LightGBM, dan *Random Forest* dalam membangun model *Explainable Expected Goals*. Meskipun LightGBM disertakan dalam eksperimen, model terbaik justru ditemukan pada *Random Forest*. Namun, penelitian tersebut tidak melakukan eksplorasi mendalam terhadap optimasi LightGBM atau bagaimana algoritma ini dapat disesuaikan lebih lanjut dalam konteks prediksi xG. Selain itu, pendekatan AutoML cenderung menyamakan konfigurasi antar model tanpa mempertimbangkan kekuatan spesifik dari setiap algoritma, yang menyebabkan potensi LightGBM dalam hal efisiensi, kemampuan interpretasi, dan performa klasifikasi tingkat lanjut belum tergali sepenuhnya. Hal ini menandakan adanya peluang terbuka untuk mengkaji secara lebih fokus kemampuan LightGBM dalam membangun model xG yang efisien, akurat, dan mudah diinterpretasikan.

Berdasarkan tinjauan terhadap berbagai pendekatan model prediksi xG di atas, dapat disimpulkan bahwa meskipun sejumlah algoritma seperti CRF, regresi logistik, GLMM, hingga *Random Forest* telah menunjukkan potensi dalam membangun model prediktif, masing-masing memiliki keterbatasan dalam hal efisiensi, *scalability*, dan fleksibilitas dalam menangani kompleksitas spasial maupun temporal data sepak bola. Algoritma LightGBM muncul sebagai alternatif yang menjanjikan karena kemampuannya dalam mengelola *dataset* besar dengan waktu komputasi yang relatif singkat (Ke *et al.*, 2017), serta keunggulannya dalam menangkap hubungan non-linear dan interaksi fitur tanpa memerlukan asumsi distribusi yang kaku (Sheridan *et al.*, 2021).

Namun, sejauh ini belum terdapat penelitian yang secara spesifik memfokuskan penerapan dan optimasi LightGBM dalam membangun model xG secara komprehensif. Oleh karena itu, diperlukan eksplorasi lebih lanjut yang mengangkat kekuatan LightGBM baik dari sisi presisi prediksi, efisiensi komputasi, maupun kemampuan interpretasi hasil agar dapat berkontribusi pada pengembangan model xG yang tidak hanya akurat, tetapi juga praktis untuk diaplikasikan dalam lingkungan analitik sepak bola modern.

Penelitian ini akan menggunakan data yang bersumber dari StatsBomb Open Data, sebuah *dataset* terbuka yang secara resmi dirilis oleh perusahaan

StatsBomb untuk mendorong kegiatan penelitian akademik dan pengembangan analisis dalam dunia sepak bola. *Dataset* ini tersedia untuk publik dan mencakup berbagai liga serta kompetisi ternama, termasuk Liga Inggris, La Liga, Liga *Champions*, dan Piala Dunia. Ketersediaan data *granular* seperti lokasi tembakan, posisi pemain, jenis aksi sebelum tembakan, hingga *freeze frame* menjadikan StatsBomb *Open Data* sebagai salah satu sumber yang sangat relevan dalam membangun model prediktif seperti *Expected Goals*. Penggunaan *dataset* ini selaras dengan misi StatsBomb dalam "*encouraging academic research and analysis through open access to high-quality football data*" (StatsBomb, 2022).

StatsBomb sendiri merupakan perusahaan penyedia data olahraga yang berbasis pada analisis dan riset, didirikan oleh para analis sepak bola profesional untuk memenuhi kebutuhan para analis pula. Mereka memiliki visi untuk menyajikan data sepak bola paling komprehensif di dunia, baik dalam aspek kuantitas maupun relevansi, yang dikumpulkan secara presisi dan dapat disesuaikan dengan kebutuhan riset lanjutan. Dalam pernyataan resminya, StatsBomb menyatakan bahwa platform mereka dibangun dari nol untuk menjamin fleksibilitas dalam menghadapi tantangan dan peluang baru di dunia olahraga yang terus berkembang (StatsBomb, 2024). Dengan pendekatan berbasis teknologi dan kedalaman data yang tidak dimiliki penyedia lain, StatsBomb menjadi rujukan utama dalam banyak riset akademik dan industri. Gambar 1.4 berikut menampilkan logo resmi dari perusahaan StatsBomb yang menjadi sumber data utama dalam penelitian ini.



Gambar 1.4 Logo Statsbomb

Berdasarkan latar belakang serta pedoman dari penelitian-penelitian sebelumnya, penulis menyimpulkan bahwa terdapat kebutuhan untuk mengembangkan model xG dengan algoritma yang lebih efisien dan akurat. LightGBM, dengan kemampuan dan keunggulannya dalam menangani *big data*, menawarkan peluang untuk menghasilkan model yang lebih baik dibandingkan

model tradisional atau algoritma lain yang telah diterapkan. Oleh karena itu, penelitian ini dilakukan sebagai upaya inovatif dalam analisis sepak bola dengan mengimplementasikan LightGBM untuk xG. Dengan demikian, skripsi ini disusun dengan judul: **"PENERAPAN *LIGHT GRADIENT BOOSTING MACHINE (LGBM)* UNTUK PERHITUNGAN METRIK *EXPECTED GOALS (xG)* DALAM ANALISIS SEPAK BOLA."**

## **1.2 Identifikasi Masalah**

Berdasarkan latar belakang yang telah dipaparkan, berikut merupakan identifikasi masalah pada penelitian ini:

- a. Meskipun model xG pada penelitian sebelumnya telah menunjukkan hasil yang baik, pendekatan yang lebih efisien dan adaptif seperti LightGBM berpotensi menawarkan peningkatan dalam akurasi, kecepatan, dan kemampuan interpretasi model.
- b. Penggunaan LightGBM dalam pengembangan model xG belum banyak diteliti padahal algoritma ini menawarkan efisiensi komputasi tinggi, kemampuan menangani data besar dengan fitur kompleks, dan performa prediksi non-linear yang sesuai untuk dinamika data sepak bola.

## **1.3 Rumusan Masalah**

Berdasarkan latar belakang yang telah dipaparkan, berikut merupakan rumusan masalah pada penelitian ini:

- a. Bagaimana penerapan algoritma LightGBM untuk meningkatkan akurasi dan efisiensi dalam perhitungan xG dalam analisis sepak bola?
- b. Bagaimana performa dari algoritma LightGBM dalam perhitungan xG dalam analisis sepak bola pada penilaian evaluasi nilai *Area Under Curve (AUC)* dan *Brier Score*?



#### 1.4 Batasan Masalah

Batasan masalah yang terdapat pada penelitian ini yaitu:

- a. Penelitian ini hanya akan berfokus pada implementasi LGBM untuk perhitungan xG dalam analisis sepak bola.
- b. Data yang digunakan diambil dari Hudl StatsBomb *open-data* yang berlisensi resmi oleh StatsBomb Services Ltd yang berkantor pusat di University of Bath Innovation Centre, Carpenter House, Broad Quay, Bath, BA1 1UD.
- c. Data terbatas pada *event* data statistik pertandingan, termasuk posisi, jarak, teknik, sudut tembakan dan lainnya.
- d. Penelitian ini fokus pada perhitungan xG menggunakan LightGBM tanpa membandingkan dengan model lain.
- e. Model probabilitas dibangun menggunakan LightGBM, tanpa membahas algoritma lain.
- f. *Preprocessing* dilakukan menggunakan *Python*, fokus pada pembersihan dan transformasi data.
- g. Data dibagi untuk *training* dan *testing* tanpa validasi silang.
- h. Metrik evaluasi terbatas pada *Area Under Curve* (AUC) dan *Brier Score*.

#### 1.5 Tujuan Penelitian

Tujuan dilakukannya penelitian ini adalah sebagai berikut:

- a. Penerapan algoritma LightGBM dalam upaya meningkatkan akurasi dan efisiensi perhitungan metrik xG pada analisis sepak bola.

- b. Evaluasi performa algoritma LightGBM dalam perhitungan metrik xG dengan menggunakan penilaian *Area Under Curve* (AUC) dan *Brier Score*.

## 1.6 Manfaat Penelitian

Manfaat dari penelitian ini yaitu sebagai berikut:

- a. Bagi peneliti, penelitian ini merupakan implementasi dari teori yang telah dipelajari dalam bidang analisis data dan *machine learning*, sehingga dapat lebih memahami penerapan algoritma LightGBM dalam perhitungan metrik xG. Selain itu, penelitian ini juga merupakan salah satu syarat kelulusan Strata Satu (S1) Sistem Informasi UIN Syarif Hidayatullah Jakarta.
- b. Bagi Universitas, penelitian ini dapat dijadikan sebagai tolak ukur pengetahuan mahasiswa terkait penerapan algoritma *machine learning* dalam analisis sepak bola, serta sebagai kontribusi dalam pengembangan penelitian di bidang ilmu komputer dan sistem informasi.
- c. Bagi pembaca, penelitian ini dapat memberikan informasi yang komprehensif mengenai algoritma LightGBM dan aplikasinya dalam perhitungan xG, serta dapat dijadikan sebagai referensi tambahan terkait penelitian dalam program studi Sistem Informasi UIN Syarif Hidayatullah Jakarta, khususnya dalam konteks analisis data olahraga. Penelitian ini juga dapat memberikan pemahaman tentang pentingnya analisis data dalam pengambilan keputusan dalam sepak bola.

- d. Bagi klub sepak bola, media sepak bola dan analis sepak bola, hasil dari penelitian ini dapat berfungsi sebagai referensi dalam mengadopsi metode analisis berbasis *machine learning*, serta dalam pengambilan keputusan strategis yang berkaitan dengan taktik permainan, rekrutmen pemain, dan evaluasi kinerja tim.

## 1.7 Metode Penelitian

Metode penelitian ini dibagi menjadi dua bagian, yaitu:

### a. Metode Pengumpulan Data

#### 1) Studi Literatur

Metode studi literatur dilakukan dengan mengumpulkan dan menganalisis berbagai sumber tertulis, seperti buku, artikel ilmiah, dan laporan penelitian yang relevan dengan topik penelitian.

#### 2) Data *Extraction*

Data *extraction* adalah proses pengambilan data dari berbagai sumber untuk dianalisis lebih lanjut. Dalam penelitian ini, data yang digunakan diambil dari Hudl StatsBomb *open-data* yang tersedia di GitHub dengan lisensi resmi.

### b. Metode Analisis Data

Penelitian ini menggunakan metode data mining yang dikenal sebagai *Knowledge Discovery in Databases* (KDD). Metode KDD terdiri dari beberapa tahap yang saling berhubungan, sebagai berikut:

#### 1) Data Selection

Data *selection* adalah proses pemilihan sub set data yang relevan dari kumpulan data yang lebih besar untuk analisis lebih lanjut. Dalam penelitian ini, pemilihan data difokuskan pada informasi

yang terkait dengan tembakan dan peluang gol, sehingga dapat digunakan dalam perhitungan metrik xG.

## 2) Preprocessing

*Preprocessing* adalah langkah yang dilakukan untuk menyiapkan dan membersihkan data sebelum analisis. Ini melibatkan penghapusan data yang tidak relevan, pengisian nilai yang hilang, dan pengubahan format data agar sesuai dengan kebutuhan analisis. Tahap ini penting untuk memastikan bahwa data yang digunakan dalam penelitian akurat dan dapat diandalkan.

## 3) Data Transformation

*Data transformation* adalah proses mengubah data ke dalam format yang lebih sesuai untuk analisis. Ini termasuk teknik seperti normalisasi, pengkodean variabel kategorial, dan agregasi data. Proses ini memungkinkan model *machine learning* untuk memproses data dengan lebih efisien dan efektif.

## 4) Data Mining

Pada tahap data mining, penelitian ini menggunakan algoritma LGBM untuk membangun model prediktif berdasarkan data yang telah diproses. LGBM dipilih karena kemampuannya dalam menangani data besar dengan efisiensi tinggi, serta akurasi yang dihasilkannya dalam perhitungan xG.

## 5) Evaluation

Setelah model dibangun, evaluasi dilakukan untuk mengukur performa model menggunakan metrik evaluasi seperti AUC dan *Brier Score*.

# 1.8 Sistematika Penulisan

Laporan pada penelitian ini terdiri atas lima bab, yaitu:

## **BAB 1           PENDAHULUAN**

Bab ini membahas tentang latar belakang, identifikasi masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan dari penelitian ini.

## **BAB 2           TINJAUAN PUSTAKA**

Bab ini membahas tentang teori-teori yang berkaitan dengan metrik xG dalam sepak bola, serta penerapan algoritma LightGBM dalam model prediksi, termasuk tinjauan mengenai penelitian-penelitian terdahulu yang relevan.

## **BAB 3           METODOLOGI PENELITIAN**

Bab ini menjelaskan tentang tahapan metode yang digunakan dalam penelitian, meliputi metode pengumpulan data, proses *preprocessing*, analisis data, dan implementasi menggunakan algoritma LightGBM, serta tahapan evaluasi dengan metrik AUC dan *Brier Score*.

## **BAB 4           HASIL DAN PEMBAHASAN**

Bab ini berisi hasil dari penerapan algoritma LightGBM dalam perhitungan metrik xG, serta analisis mendalam mengenai kinerja model berdasarkan evaluasi yang dilakukan. Hasil juga dibandingkan dengan model lain untuk menunjukkan efektivitas LightGBM.

## **BAB V          PENUTUP**

Bab ini berisi kesimpulan dari hasil penelitian mengenai penerapan algoritma LightGBM dalam perhitungan metrik xG, serta saran-saran yang dapat digunakan untuk penelitian selanjutnya dalam bidang analisis sepak bola dan penerapan *machine learning*.