

# **BAB I**

## **PENDAHULUAN**

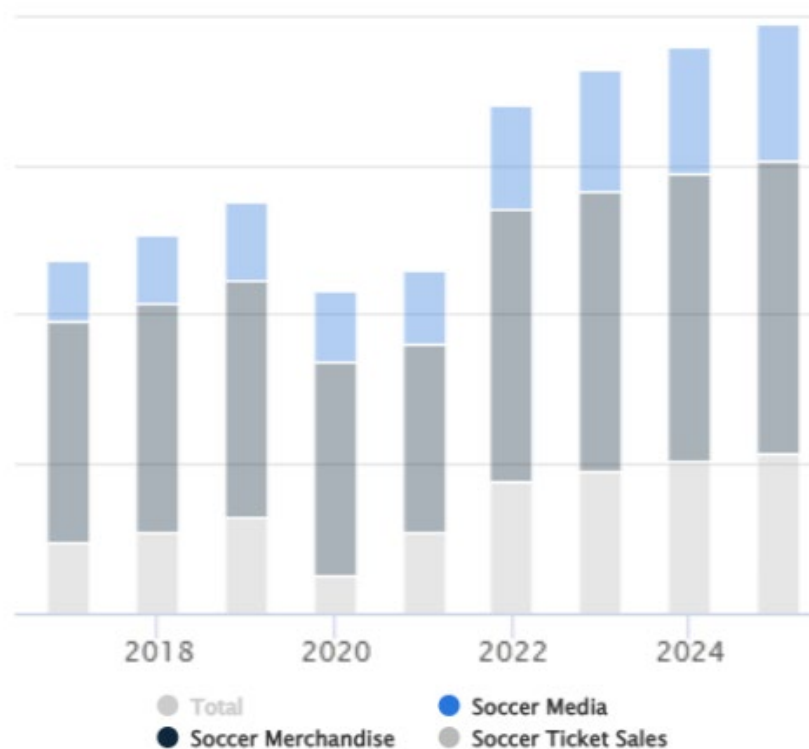
### **1.1 Latar Belakang**

Ketidakpastian adalah elemen tak terpisahkan dalam setiap ranah olahraga, dan justru inilah yang membuatnya begitu menarik bagi banyak orang. Dalam sepak bola contohnya, perpaduan antara performa dan keberuntungan sering kali menjadi penentu kemenangan atau kekalahan, menjadikan olahraga ini semakin kompleks dan menarik untuk dianalisis secara mendalam. Sepak bola di era modern saat ini telah mengalami perubahan signifikan dengan adanya integrasi *data science* dan *machine learning* dalam berbagai aspek pengelolaan olahraga ini. Saat ini, sepak bola bukan hanya sekadar kompetisi olahraga, melainkan juga bagian dari industri yang lebih luas. Dalam era ini, data tidak lagi sekadar angka, tetapi menjadi landasan penting untuk mengukur performa dan membuat analisis berbasis fakta. Teknologi data *analytics* dan metrik spesifik semakin mendukung pengambilan keputusan dalam manajemen sepak bola. Penggunaan perangkat *wearable* dan sistem informasi mutakhir memungkinkan tim untuk mengumpulkan serta menganalisis data dalam jumlah besar, yang berperan penting dalam aspek taktik pertandingan, *scouting* pemain, hingga pencegahan cedera (Chatziparaskevas et al., 2024).

Di Indonesia sendiri, pasar sepak bola diproyeksikan menghasilkan pendapatan sebesar USD 158,30 juta pada tahun 2024. Pertumbuhan tahunan sebesar 2,80% (CAGR 2024-2029) diperkirakan akan meningkatkan volume pasar

hingga mencapai USD 181,70 juta pada tahun 2029. Secara global, pendapatan terbesar akan dihasilkan oleh Inggris, dengan nilai sebesar USD 9.696,00 juta pada 2024.

Selain itu, rata-rata pendapatan per pengguna (ARPU) di pasar sepak bola Indonesia diproyeksikan mencapai USD 28,27 pada 2024, dengan jumlah pengguna yang diperkirakan mencapai 6,2 juta pada tahun 2029. Tingkat penetrasi pengguna dalam pasar ini akan mencapai 2,0% pada tahun 2024 (Statista, 2024). Gambar 1.1 menampilkan proyeksi pendapatan tahunan pasar sepak bola di Indonesia.

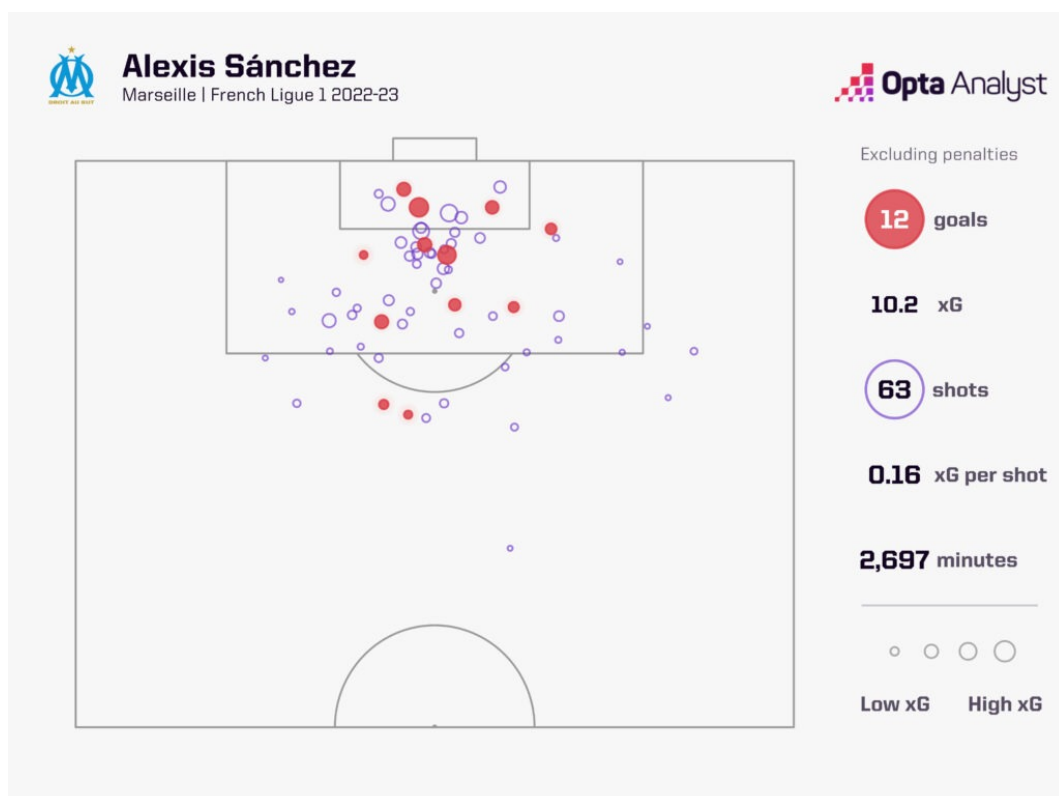


Gambar 1.1 Pendapatan Tahunan Pasar Sepak Bola di Indonesia

Dalam analisis sepak bola modern, metrik yang dikenal sebagai *Expected Goals* (xG) menjadi komponen penting dalam mengevaluasi kualitas peluang mencetak gol. Metrik ini menghitung probabilitas suatu peluang menghasilkan gol, yang pada akhirnya memberikan wawasan lebih akurat terkait hasil pertandingan.

Dengan mengakumulasi nilai xG dari setiap peluang dalam suatu pertandingan, dapat diperoleh hasil pertandingan yang diharapkan. Pemanfaatan xG terbukti memberikan wawasan berharga terkait performa tim dan pemain, terutama dalam menentukan efektivitas serangan serta pengambilan keputusan yang lebih baik di masa mendatang (Eggels, 2016).

Salah satu contoh penerapan xG dalam sepak bola adalah melalui analisis *shot-map*, yang memperlihatkan distribusi dan kualitas peluang yang dihasilkan oleh seorang pemain. Gambar 1.2 menampilkan *shot-map* Alexis Sanchez, di mana setiap titik tembakan dilengkapi dengan nilai xG yang menunjukkan peluang keberhasilan menjadi gol (Whitmore, 2023). Analisis semacam ini membantu pelatih dan manajemen tim dalam mengevaluasi kontribusi pemain serta mengidentifikasi area yang memerlukan perbaikan atau penguatan strategi.



Gambar 1.2 Visualisasi *Shot-map* xG

Perhitungan xG harus dilakukan dengan akurasi dan ketepatan tinggi untuk memberikan wawasan yang andal terkait kualitas peluang gol. Ketepatan ini sangat penting dalam mendukung keputusan strategis, seperti menentukan taktik permainan dan mengevaluasi kinerja pemain. Namun, perhitungan xG secara manual menghadapi banyak tantangan, terutama di era teknologi saat ini, karena data yang digunakan semakin kompleks dan bervolume besar. Penggunaan metode manual tidak hanya memakan waktu tetapi juga rentan terhadap kesalahan manusia. Oleh karena itu, metode *machine learning* menjadi solusi untuk mempercepat dan mempermudah perhitungan xG dengan memanfaatkan data historis, sehingga hasilnya lebih konsisten dan akurat.

Dengan menerapkan konsep ini, komputer dapat belajar dari data dan pengalaman untuk menghasilkan keputusan atau prediksi secara mandiri. Menurut Pratama et al. (2017), *machine learning* adalah bidang keilmuan yang memungkinkan komputer atau mesin untuk menjadi cerdas dengan cara belajar dari data yang diberikan. Metode ini sangat bermanfaat dalam berbagai aplikasi, termasuk dalam analisis sepak bola, karena mampu memproses data yang kompleks dan menghasilkan hasil yang lebih presisi.

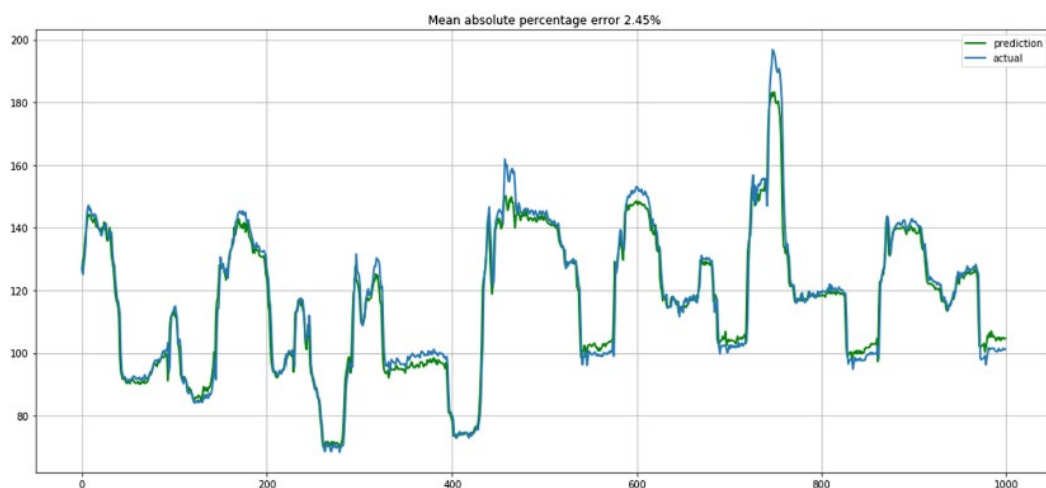
Dalam mengoptimalkan penerapan analisis xG yang telah dibahas sebelumnya, diperlukan model *machine learning* yang mampu menangani data historis dengan efisiensi tinggi. Pada tahap inilah *Light Gradient Boosting Machine* (LightGBM atau LGBM) memainkan peran penting. Secara umum, LightGBM

adalah metode *machine learning* berbasis *Gradient Boosting Decision Tree* (GBDT) yang digunakan untuk prediksi dan klasifikasi data (Ramadanti et al., 2024). Teknologi ini dirancang untuk mengatasi tantangan dalam pemrosesan data besar dan kompleks, menjadikannya unggul dalam berbagai aplikasi berbasis data.

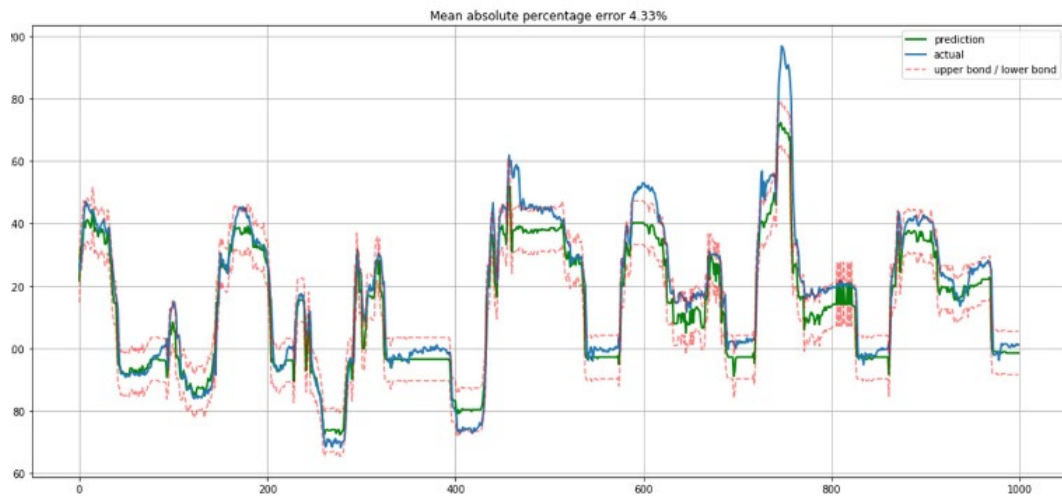
LightGBM memungkinkan pengembangan model xG secara efisien dengan mempelajari pola dari data historis. Dibandingkan dengan GBDT tradisional, LightGBM menawarkan proses pelatihan yang hingga 20 kali lebih cepat tanpa mengurangi akurasi yang dihasilkan (Hartanto et al., 2023). Kemampuan ini sangat penting dalam analisis sepak bola, terutama saat bekerja dengan data berukuran besar dan kompleks. Dengan efisiensinya, LightGBM memungkinkan evaluasi peluang gol dilakukan lebih cepat dan akurat, sehingga memberikan wawasan yang dapat diterapkan dalam waktu nyata untuk mendukung pengambilan keputusan strategis.

LightGBM juga memiliki beberapa keunggulan yang menjadikannya unggul dalam konteks pemodelan prediktif. Sebagai *framework gradient boosting* berbasis algoritma *decision tree*, LightGBM tidak hanya cepat tetapi juga mendukung distribusi data dalam skala besar dengan performa tinggi. Kemampuannya dalam klasifikasi yang unggul telah diaplikasikan secara sukses dalam diagnosis penyakit dan prediksi hasil klinis, memperkuat keandalannya dalam berbagai bidang analitik (Artzi et al., 2020). Dalam sepak bola, kemampuan klasifikasinya dapat digunakan untuk memprediksi hasil pertandingan dan mengidentifikasi peluang yang berpotensi menghasilkan gol.

Dibandingkan dengan model lainnya, seperti XGBoost, LightGBM menunjukkan kinerja prediktif yang lebih baik. Studi menunjukkan bahwa LightGBM menghasilkan nilai *Mean Absolute Percentage Error* (MAPE) sebesar 2,45%, yang lebih kecil daripada MAPE yang dihasilkan oleh XGBoost. Ini menandakan bahwa LightGBM lebih akurat dalam memprediksi variabel target, termasuk dalam konteks *output* daya termal maupun analisis sepak bola (Nemeth et al., 2019). Keunggulan dalam akurasi dan kecepatan ini membuat LightGBM menjadi pilihan utama dalam berbagai aplikasi berbasis data, termasuk pemodelan xG di sepak bola. Gambar 1.3 menunjukkan grafik yang membandingkan MAPE dari berbagai model, di mana LightGBM menonjol dengan nilai MAPE terbaik. Sementara itu, Gambar 1.4 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 1.3 Grafik Performa Model LightGBM



Gambar 1.4 Grafik Performa Model Lain

Dalam rangka mendukung analisis yang telah dilakukan sebelumnya dengan menggunakan model LightGBM untuk memprediksi nilai xG, penting untuk menggunakan dataset yang kredibel dan komprehensif. Oleh karena itu, penelitian ini memanfaatkan dataset *open-data* dari StatsBomb, yang dipublikasikan secara gratis oleh perusahaan tersebut untuk mendorong riset akademis dan analisis dalam bidang olahraga (StatsBomb, 2022). Data ini mencakup berbagai informasi rinci terkait pertandingan sepak bola, termasuk data peluang dan tembakan, yang relevan untuk pengembangan model xG. Dengan menggunakan dataset ini, penelitian dapat dilakukan secara lebih komprehensif, memungkinkan analisis mendalam berbasis data historis yang kredibel.



Gambar 1.5 Logo StatsBomb

StatsBomb merupakan perusahaan data olahraga yang didirikan oleh analis dan untuk para analis. Perusahaan ini memiliki tim yang berdedikasi tinggi dalam mengumpulkan serta menganalisis data olahraga terlengkap di dunia. Platform StatsBomb dirancang dari awal untuk memastikan pengumpulan dan analisis data lebih relevan serta fleksibel dibandingkan dengan penyedia lainnya. Kemampuan platform ini untuk merespons kebutuhan, peluang, dan tantangan baru secara cepat menjadikannya salah satu pemimpin dalam industri data olahraga (StatsBomb, 2024).

Penelitian awal mengenai xG dilakukan oleh Lucey et al. (2015), yang menggunakan dataset berisi 9.732 tembakan dan 10 detik cuplikan video sebelum setiap tembakan dalam pertandingan profesional. Dataset ini berasal dari Prozone, yang sekarang dikenal sebagai Stats Perform. Dalam penelitian tersebut, model yang dikembangkan dinamakan *Expected Goal Value* (EGV), dengan algoritma yang digunakan berupa *Conditional Random Models* berbasis model probabilistik. Penelitian ini menunjukkan bahwa pemanfaatan data spasial-temporal tidak hanya meningkatkan prediksi hasil tembakan tetapi juga memberikan wawasan lebih mendalam tentang strategi pemain selama pertandingan.

Penelitian sejenis dilakukan oleh Fairchild et al. (2018), yang memanfaatkan dataset berisi 1.115 tembakan non-penalti dari 99 pertandingan *Major League Soccer* (MLS) di Amerika Serikat. Berbeda dengan penelitian sebelumnya, dataset ini dikumpulkan secara manual. Model yang dibangun dalam penelitian ini adalah *Expected Goal Model*, dengan penggunaan algoritma *Logistic Regression* untuk memprediksi peluang gol berdasarkan posisi dan karakteristik



tembakan. Studi ini berfokus pada analisis spasial tembakan dan menyajikan dimensi fraktal yang menggambarkan pola distribusi peluang mencetak gol di MLS.

Tureen dan Olthoff (2022) memperluas kajian mengenai analisis individual dalam sepak bola melalui pengembangan model *Estimated Player Impact* (EPI). Penelitian ini menggunakan data dari 580 pertandingan *Premier League* dan 326 pertandingan *Women's Super League*, yang semuanya berasal dari penyedia data StatsBomb, sama seperti penelitian kali ini. Model EPI dibangun menggunakan algoritma *Generalised Linear Mixed Models* (GLMM), yang memungkinkan kuantifikasi dampak individu pemain terhadap berbagai aksi dalam pertandingan sepak bola. Studi ini menekankan pentingnya analisis hierarkis untuk mengidentifikasi peran spesifik pemain dalam memengaruhi hasil pertandingan.

Meskipun model xG telah diterapkan dalam berbagai penelitian, penggunaan algoritma LightGBM dalam pengembangan model ini masih terbatas. Cavus dan Biecek (2022) melakukan eksplorasi dengan menggunakan beberapa model, termasuk LightGBM, melalui Forester AutoML. Dataset yang digunakan terdiri dari 315.430 tembakan selama tujuh musim dari lima liga top Eropa, yang disediakan oleh Understat. Model yang dikembangkan dalam studi ini adalah *Explainable Expected Goals*, dengan memanfaatkan algoritma seperti XGBoost, *Random Forest*, LightGBM, dan CatBoost. Namun, hasil penelitian menunjukkan bahwa model terbaik dalam memprediksi peluang gol adalah *Random Forest*, yang mengindikasikan bahwa masih terdapat ruang untuk meningkatkan kinerja LightGBM dalam model xG di masa mendatang.

Berdasarkan penelitian sebelumnya, terdapat beberapa masalah dan tantangan yang mengindikasikan perlunya pengembangan lebih lanjut dalam penerapan model xG menggunakan algoritma LightGBM. Meskipun *Random Forest* telah menunjukkan kinerja terbaik dalam penelitian Cavus dan Biecek (2022), penggunaan LightGBM tetap menjanjikan karena memiliki keunggulan dalam efisiensi waktu dan kemampuan menangani data dalam skala besar. Selain itu, model xG sebelumnya sebagian besar menggunakan algoritma tradisional seperti *Logistic Regression* atau GLMM, yang mungkin kurang optimal dalam menganalisis data kompleks secara cepat dan akurat. LightGBM, dengan kemampuan *gradient boosting* yang canggih, dapat memberikan kombinasi ideal antara kecepatan pelatihan, skalabilitas, dan akurasi tinggi.

Dengan meningkatnya volume dan kompleksitas data sepak bola, khususnya data spasial dan temporal yang disediakan oleh StatsBomb, diperlukan model yang mampu memanfaatkan data tersebut secara maksimal. LightGBM diharapkan dapat mengatasi keterbatasan model-model sebelumnya dengan menghasilkan prediksi yang lebih akurat serta mendukung analisis yang lebih mendalam, misalnya melalui eksplorasi pola tembakan atau pengaruh pemain dalam waktu nyata. Implementasi algoritma ini dalam penelitian xG juga memungkinkan evaluasi baru dalam aspek kualitas peluang, yang tidak hanya berfokus pada jumlah tembakan, tetapi juga pada strategi dan konteks permainan. Oleh karena itu, penelitian ini mencoba membuktikan bahwa penerapan LightGBM dapat menjadi solusi yang lebih baik dalam pengembangan model xG untuk analisis sepak bola.

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. Perhitungan manual xG tidak efisien memakan waktu dan rentan terhadap kesalahan.
- b. Keterbatasan model xG tradisional yang menggunakan algoritma tradisional (misalnya, *Logistic Regression* atau GLMM) belum optimal dalam menangani data spasial-temporal yang kompleks secara cepat dan akurat.
- c. Kurangnya optimalisasi lebih lanjut pada penerapan LightGBM dalam Analisis xG.

### 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*.